Unicode et typographie : un amour impossible

Yannis Haralambous

Unicode et typographie : un amour impossible

Yannis Haralambous

Département Informatique
École Nationale Supérieure des Télécommunications de Bretagne
BP 832, 29 285 Brest, France

Yannis.Haralambous@enst-bretagne.fr
http://omega.enstb.org/yannis

RÉSUMÉ. *Dans cet article nous proposons une nouvelle approche aux concepts de* glyphe *et* caractère *et nous appliquons à l'étude du codage Unicode. Après une discussion étendue de ces deux concepts, nous nous intéressons aux autres ingrédients de base d'Unicode : les* glyphes privilégiés, *les* descriptions, *les* caractères combinants *et les* propriétés. *Dans chaque cas nous essayons en même temps de montrer l'importance de ces notions et d'indiquer leurs limitations par des contre-exemples. Cet article se veut aussi bien une « exégèse » qu'une critique (constructive) de ce codage auquel est dédié ce numéro spécial de la revue* Document Numérique.

ABSTRACT. *In this paper we give a new approach to the concepts of* glyph *and* character, *applied to the study of the Unicode encoding. After a thorough discussion of these two concepts, we describe the other fundamental ingredients of the encoding:* privileged glyphs, character descriptions, combining characters, *and* properties. *In each case we try to show the importance of the given notion and at the same time to draw its limits by using counter-examples. This paper aims to be an "exegesis" as well as a (positive) critic of the encoding to which this special number of* Document Numérique *is dedicated.*

MOTS-CLÉS : *Unicode, typographie, glyphes, caractères, langues orientales.*

KEYWORDS: *Unicode, Typography, Glyphs, Characters, Oriental languages.*

Document Numérique. Volume 6 - n° 3-4/2002, pages 105 à 137

1. Caractères, glyphes et Unicode

Depuis des années on entend parler de *caractères* et de *glyphes*. Les « polices de caractères » contiennent – malgré leur nom – des glyphes ; le codage Unicode contient des « caractères ». Qu'en est-il vraiment ? Essayons de tirer au clair les différences entre ces deux concepts, pour mieux comprendre le fonctionnement d'Unicode et les rapports entre celui-ci, la typographie et l'informatique en général.

1.1. *Les glyphes*

Tout d'abord soyons clairs sur les termes utilisés. Par *glyphe* ($\gamma\lambda\upsilon\phi\acute{\eta}$ = ciselure, gravure selon Bailly, *Dictionnaire grec-français*) nous entendons, dans le cadre du document électronique et des codages, un *signe typographique vu en tant qu'image*. Mais pourquoi cet intérêt spécifique à la typographie plutôt qu'à l'écriture manuscrite, voire même la calligraphie ?

Parce que la typographie a été une importante *modélisation* de l'écriture : par la réutilisation de types mobiles, la lettre a acquis une forme très stable et on pourrait même dire que dans un contexte assez étroit, par exemple dans le cadre d'un même ouvrage, ou celui des livres issus du même éditeur et imprimeur pendant la même période, la lettre acquiert une forme unique : ainsi, par exemple, dans le paragraphe que le lecteur a sous les yeux, toutes les lettres 'a' romaines sont représentées par des glyphes identiques. D'ailleurs, le non respect de cette règle de stabilité du glyphe, dû par exemple à des différences de densité d'encre, est considéré comme un défaut d'impression. Dans ce contexte bien étroit la modélisation est parfaite : une lettre donnée est représentée par une seule image, reproduite à l'identique des milliers de fois.

Que se passe-t-il si l'on élargit le contexte, par exemple si l'on parle de *toutes* les lettres 'a' de *tous* les livres français ? On retrouve alors la classification traditionnelle des signes typographiques en fonte, corps, style, graisse, etc. Cette classification peut être plus ou moins précise : on peut vaguement parler de « didone en petit corps », ou alors être très précis, en parlant de « Didot maigre droit 9 points de l'Imprimerie de la Librairie Nouvelle à Paris en 1856 » (fig. 2).

Qui dit classification dit description, et pour décrire les glyphes on dispose de différentes caractéristiques qui constituent autant d'*axes de variation*. Axes qui ne sont pas forcément orthogonaux et qui dépendent souvent du système d'écriture et plus généralement du contexte historique et géographique. Ainsi on aurait pu considérer que le nom de la fonte est orthogonal au corps ; et pourtant il existe des fontes qui ne sont disponibles qu'en grand corps. De même, le concept, par exemple, d'*empattement* peut être très utile à la description de caractères latins, grecs, cyrilliques, arméniens et géorgiens, mais ne s'applique pas à des écritures comme l'arabe ou la dévanagari.

Jusqu'à maintenant nous n'avons abordé que les aspects visuels du glyphe. Mais le fait qu'il s'agisse d'écriture fait que chaque glyphe a aussi une *signification*, un *contenu linguistique ou symbolique*. Et ce contenu fait, bien évidemment, aussi partie

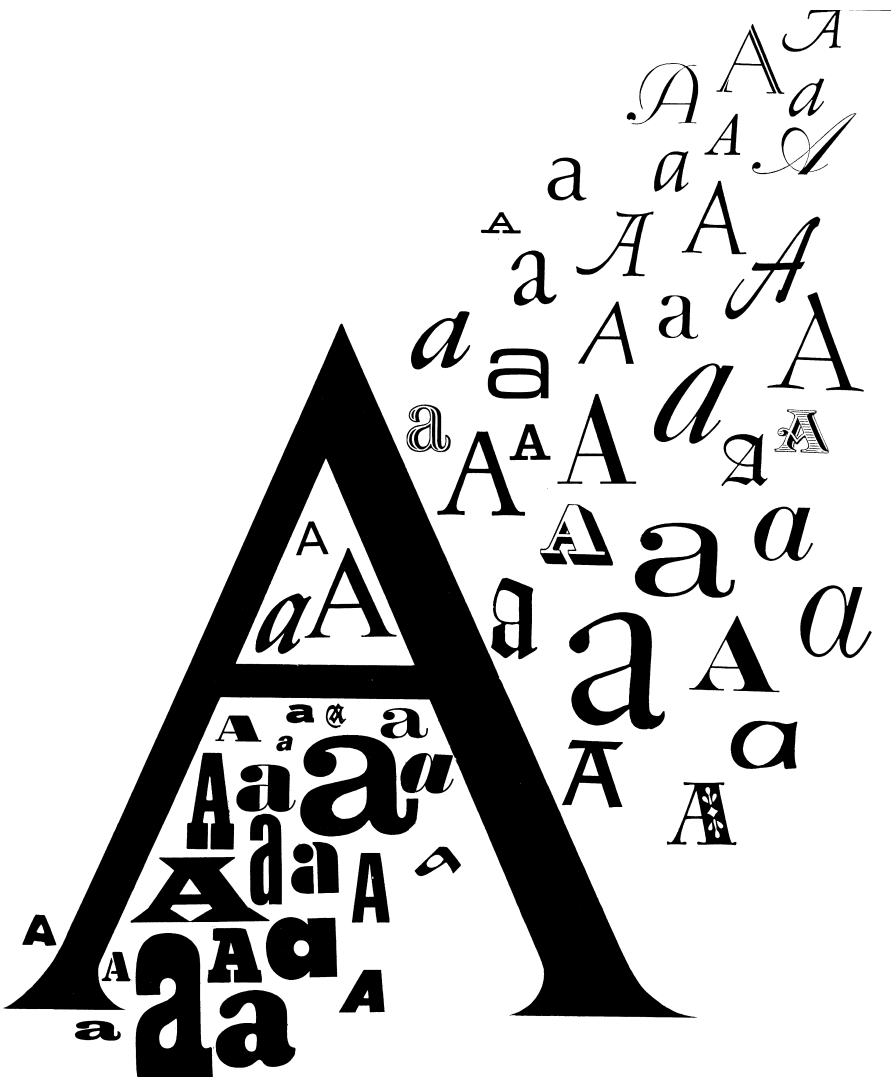


Figure 1. *Illustration de la multitude de glyphes possibles pour les mêmes caractères ‘a’ et ‘A’, tirée du livre* Design with Type *de Dair, University of Toronto Press, 1967*

Ce dimanche était le treizième de l'année 1813. Le surlendemain, Napoléon partait pour cette fatale campagne pendant laquelle il allait perdre successivement Bessières et Duroc, gagner les mémorables batailles de Lutzen et de Bautzen, se voir trahi par l'Autriche, la Saxe, la Bavière, par Bernadotte, et disputer la terrible bataille de Leipsick. La magnifique parade commandée par l'empereur devait être la dernière de celles qui excitèrent si longtemps l'admiration des Parisiens et des étrangers. La vieille garde allait exécuter pour la dernière fois les savantes manœuvres dont la pompe et la précision étonnèrent quelquefois jusqu'à ce géant lui-même, qui s'apprêtait alors à son duel avec l'Europe. Un sentiment triste amenait aux Tuileries une brillante et curieuse population. Chacun semblait deviner l'ave-

Figure 2. *Extrait d'un ouvrage français imprimé en 1856 à l'Imprimerie de la Librairie Nouvelle, à Paris*

Figure 3. *Gros Parangon (Grec 152) des* Grecs du Roi, *Robert Étienne, 1550*

Figure 4. *Logotype de la marque Coca-Cola*

de la description du glyphe. Un glyphe étant un *objet culturel*, on peut d'ailleurs aisément imaginer bon nombre d'autres manières de le décrire, que ce soit en le situant dans l'histoire de la typographie (« le alpha des Grecs du Roi »...), en utilisant des connotations culturelles (« le C du logo Coca-Cola »...), etc.

Devant l'infinité écrasante de glyphes possibles, on se met donc à les classifier en les décrivant, un procédé similaire à celui de la classification des nombres en mathé-

matiques. Dans la suite de cet article on va donc s'intéresser à la classification des glyphes et aux manières de les décrire.

1.2. *Les caractères*

Il est temps de définir le concept de *caractère*. On a vu que les glyphes peuvent être décrits et classifiés. On peut donc former des classes de glyphes suivant la description qu'on en donne : pour chaque glyphe on dira qu'il appartient à une classe donnée s'il correspond à la description de la classe. Il convient donc de bien choisir les descriptions des glyphes, pour obtenir des classes utiles. Et voici un tel choix :

Un *caractère* ($\chi\alpha\rho\alpha\kappa\tau\acute{\eta}\rho$ = signe gravé, empreinte, figure gravée, mais aussi signe distinctif, marque, caractère extérieur propre à une personne ou à une chose, selon Bailly, *Dictionnaire grec-français*) sera pour nous une *classe de glyphes dont la description est à prépondérance linguistique ou logique*[1].

Si nous décrivons, par exemple, un caractère comme « lettre latine majuscule A », cette description est bien à prépondérance linguistique puisque (a) nous nous référons au fait qu'il s'agit d'une « lettre » (donc d'un objet linguistique) et que cette lettre fait partie d'un ensemble bien connu : l'alphabet latin, et (b) puisque nous donnons une propriété linguistique supplémentaire de cette lettre, en indiquant qu'elle est majuscule. Dans cet exemple, la classe de glyphes correspondante est l'ensemble de toutes les images de signes typographiques A dans tous les corps, polices, styles et graisses possibles : un nombre énorme de glyphes d'une variété inimaginable[2].

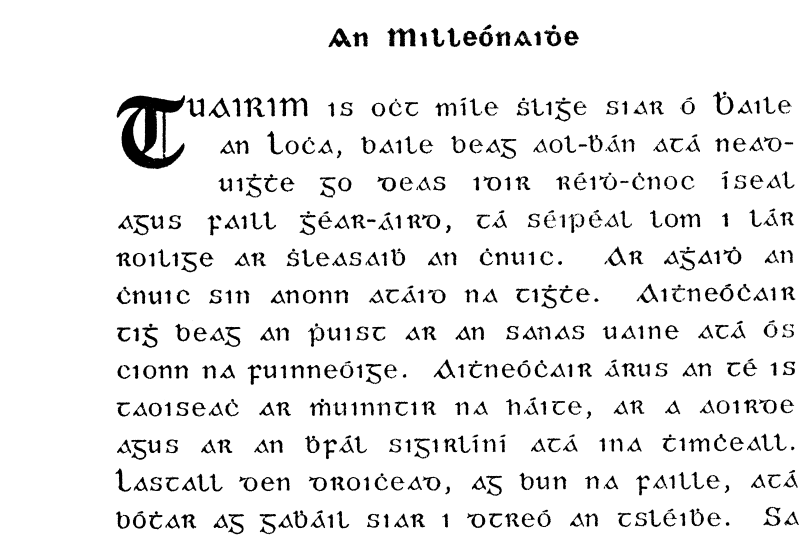
Bien sûr, pour *illustrer* cette classe de glyphes nous sommes amenés à utiliser un représentant privilégié qui sera un glyphe plutôt *neutre*. Ainsi le glyphe A est un meilleur choix que $\mathfrak{A}$ pour illustrer la « lettre latine majuscule A », exactement comme en mathématiques dans un souci d'éviter la confusion on n'utilisera pas un triangle isoscèle pour illustrer la notion de triangle en général.

En décrivant avec plus de précision le caractère, nous pouvons restreindre le nombre de glyphes dans sa classe : ainsi nous pouvons parler de *lettre latine romaine majuscule A*, ou de *lettre latine italique majuscule A* dont les ensembles de glyphes

---

1. Cette définition diffère de celle donnée dans le standard Unicode que nous trouvons moins claire et moins rigoureuse : *characters are the smallest components of written language that have semantic value* (= les caractères sont les plus petites composantes de la langue écrite qui ont une valeur sémantique). Cette définition, comme nous le verrons par la suite, ne tient pas compte des descriptions de classes de glyphes et des glyphes privilégiés dans les tables Unicode, ni de leurs valeurs normatives. C'est pour cela que nous avons préféré de redéfinir la notion de *caractère*.

2. Cas extrême : dans les polices dites « expérimentales » on trouve même des glyphes qui sont à la limite du reconnaissable ; ces glyphes ne représentent la lettre A que parce que leurs créateurs l'ont déclaré comme étant le cas...


An Milleónaide

Tuairim is ocht míle slige siar ó Baile an Loca, baile beag aol-bán atá neas-uigte go deas ioir réid-cnoc íseal agus fraill géar-áird, tá séipéal lom i lár roilige ar sleasaid an cnuic. Ar agaid an cnuic sin anonn atáid na tigte. Aitneócair tig beag an puist ar an sanas uaine atá ós cionn na fuinneóige. Aitneócair árus an té is taoiseach ar muinntir na háite, ar a aoirde agus ar an bpál sigirlíní atá ina timcheall. Lastall den droichead, ag bun na fraille, atá bóthar ag gabáil siar i dtreó an tsléibe. Sa

Figure 5. *Extrait d'un ouvrage irlandais en écriture gaëlique imprimé en 1940 à Baile átha Cliath, c'est-à-dire Dublin*

sont des sous-ensembles stricts[3] de l'ensemble des glyphes du caractère *lettre latine majuscule A.*

Mais – nous venons de le voir – pour constituer un caractère, une classe de glyphes doit être *à prépondérance linguistique ou logique*. On peut à juste titre se poser la question si ces nouvelles descriptions plus spécialisées le sont toujours ; la réponse est très probablement non.

Notons que souvent la réponse à cette question n'est pas aussi claire : quand on dit, par exemple, *lettre latine gaëlique majuscule A*[4], est-ce une description linguistique ? Sachant que l'écriture gaëlique (voir fig. 5) n'est utilisée (en typographie) que pour la composition en langue irlandaise, et que jusqu'à il y a encore 50 ans la réciproque était également vraie (l'irlandais n'était composé qu'en caractères gaëliques), on aurait pu assimiler la classe des « lettres gaëliques A » à celle des glyphes de la « lettre A en langue irlandaise écrite », ce qui est une description incontestablement linguistique et constitue, par définition, un caractère. Et pourtant, cet avis n'est pas partagé par la

3. Le fait qu'ils soient des sous-ensembles stricts n'implique pas que l'on sache bien les défi - nir : quels sont les signes typographiques « italiques » ? Comment les distingue-t-on des « romains » ? À quel moment un signe devient-il italique ou cesse-t-il d'être italique ? Autant de questions sans réponse, puisque l'« italicité » est basée sur des critères culturels, historiques, esthétiques et donc hautement subjectifs, voire même intuitifs.

4. Comme le gothique, le gaëlique est un alphabet qui se superpose à quelques détails près à l'alphabet latin mais dont les caractères sont de forme relativement distante de nos lettres latines modernes.

communauté informatique actuelle (et de ce fait, comme on le verra, il n'y a pas de table gaëlique dans Unicode, qui se veut un codage de caractères).

Avant de passer à l'informatique et aux codages, résumons les principaux points de notre discussion, jusqu'ici plutôt abstraite :

– un *glyphe* est l'image d'un signe typographique ;

– un *caractère* est une description à prépondérance linguistique ou logique d'une classe de glyphes ;

– dans le but d'illustrer un caractère on utilise un glyphe particulier, que l'on appelle *glyphe privilégié* ;

– pour limiter le nombre de glyphes correspondant à un caractère on peut raffiner la description du caractère, mais il faut le faire de manière à ce qu'elle reste autant que possible à prépondérance linguistique ou logique[5].

1.3. *L'informatique*

Passons maintenant à l'informatique. Nous savons que dans la mémoire de l'ordinateur toutes les informations ne sont au fond qu'une interminable suite de choix binaires (0 ou 1, « vrai » ou « faux », etc.). Pour obtenir des objets plus maniables, on a d'abord voulu grouper ces choix binaires et travailler avec des n-uplets. En les groupant par groupes de huit, on obtient des *octets* c'est-à-dire des nombres entre 0 et 255 ; en les groupant par groupes de 32 choix, on obtient des nombres entre 0 et $2^{32} - 1$, c'est-à-dire entre 0 et quelques 4 milliards.

Les nombres sont bien utiles pour les calculs, mais, ne vivant pas dans le monde imaginé par Bradbury et Truffaut dans *Fahrenheit 451*, nous souhaitons aussi traiter du *texte*. On a donc intérêt à faire correspondre aux nombres de la mémoire de l'ordinateur des objets informatiques qui nous permettront de traiter du texte. De quel type seront ces objets ? Eh bien, cela dépend du traitement que nous souhaitons effectuer.

Si ce traitement est une impression sur papier ou l'affichage sur écran du texte, alors ces objets informatiques peuvent très bien être tout simplement des glyphes : en les plaçant judicieusement sur la page (ou sur l'écran), nous obtenons le texte souhaité, imprimé ou affiché.

Si, par contre, il s'agit d'un traitement du contenu de notre texte, un traitement « linguistique » (par exemple une recherche de chaîne de texte, une vérification orthographique ou une indexation), alors la machine doit en fait accéder au contenu linguistique ou symbolique des glyphes, plutôt qu'aux glyphes eux-mêmes (qui, pour l'ordinateur, en l'absence d'un humain pour les interpréter, ne sont que des vulgaires images). Et puisqu'il s'agit de traitement linguistique, on a intérêt à fusionner tous

5. Soyons un peu plus précis : par « linguistique ou logique », nous entendons une description utilisant des termes tirés des définitions et descriptions de graphèmes dans des grammaires de notations, que ce soit des grammaires de langues ou des grammaires d'autres systèmes de notation, par exemple la notation musicale ou la notation des mathématiques.

les glyphes qui ont les mêmes propriétés linguistiques et de faire correspondre les nombres de l'ordinateur à ces classes de glyphes. Le lecteur l'aura déviné, on retombe sur la définition de *caractère*, que nous avons vue précédemment.

Il est donc pertinent d'utiliser les *caractères* pour stocker le texte dans la mémoire de l'ordinateur, tout en gardant les *glyphes* pour certains traitements plus « mécaniques », comme les impressions papier et l'affichage écran. Et c'est exactement ce qui se fait dans un système informatique moderne :

– le texte est représenté en mémoire par des *caractères* ; une correspondance entre les nombres de l'ordinateur et les caractères est appelé un *codage (de caractères)* ;

– les *glyphes* sont stockés dans des objets informatiques appelés *polices*[6] ou *fontes* ; une correspondance entre les nombres de l'ordinateur et des glyphes est appelée une *table de fonte*[7].

1.4. *Unicode*

Historiquement, tant qu'on utilisait des codages à un octet (256 positions), on ne pouvait couvrir qu'un petit nombre d'alphabets à la fois : il y avait des codages d'alphabet latin pour langues occidentales, d'alphabet latin pour langues d'Europe centrale, d'alphabet latin et arabe, et ainsi de suite. L'inconvénient de ce système était, entre autres, que l'on était très rapidement amené à utiliser plusieurs codages dans le même document et cela demandait la possibilité de conserver dans le document une trace des codages utilisés ainsi que des passages d'un codage à l'autre. De même, les logiciels utilisés devaient être compatibles avec tous les codages possibles et imaginables, faire des conversions à la volée, permettre un affichage correct à tout moment, et tout cela sans importuner l'utilisateur... inutile de dire que tout cela dépassait les possibilités des logiciels courants.

Unicode a de quoi résoudre ces problèmes, du moins en partie. Étant un codage à 20 bits (1 048 576 positions), ce codage a suffisamment de place pour contenir un très grand nombre de caractères ; de plus, un de ses principes fondamentaux est le fait que

6. En fait, polices de *caractères* : petite anomalie terminologique, liée à cette multitude de signifi cations que l'on donne au terme *caractère* : en typographie, un caractère n'est autre qu'une fonte. Comble de l'ironie : en grec moderne, le mot $\chi\alpha\rho\alpha\kappa\tau\eta\rho\iota\sigma\mu\acute{o}\varsigma$, c'est-à-dire « caractérisation », est, en jargon typographique, l'équivalent du terme français « enrichissement », c'est-à-dire l'adjonction de propriétés stylistiques (gras, italiques, etc.). Conceptuellement, l'action d'enrichissement est aux antipodes du concept de *caractère* donné en informatique.

7. La situation se complique davantage puisque souvent les glyphes dans les polices ont des noms. Étant donné que ces noms ont tendance à être de nature linguistique ou logique, ils pourraient *a priori* être assimilés à des caractères. Il n'en est rien puisque les polices, par la nature même de la typographie, sont obligées de contenir plusieurs glyphes ne correspondant pas à des classes « linguistiques ou logiques » (comme par exemple les ligatures 'ffi', 'fff', qui ne sont pas perçues comme des *caractères*). De même, une table de fonte peut contenir plus d'un glyphe appartenant au même *caractère*, les noms de ces glyphes ne peuvent donc pas correspondre à plusieurs caractères. Nous appelerons ces objets hybrides et souvent ambigüs, des *noms de glyphes*.

tout texte codé dans un codage « reconnu » (ISO ou commercial suffisamment répandu) peut être converti en Unicode, sans perte d'information. Unicode est donc un « super-codage » qui contient tous les précédents et les rend donc obsolètes. Le logiciel n'a plus à s'occuper des choix et changements de codage puisqu'il n'y en a qu'un seul ; reste le problème de l'affichage, qui est plutôt de l'essor du système d'exploitation, et qui est relativement bien résolu sous Linux ou Windows 2000. Unicode remplace donc bel et bien les codages qui le précédent, et nous débarasse des problèmes qui s'ensuivent.

Cela a tout l'air d'une fabuleuse réussite, surtout qu'en jouant, en quelque sorte, le rôle des Nations Unies, Unicode donne la parole aux petites minorités et contient désormais une floppée de systèmes d'écriture de plus en plus exotiques utilisés par des minorités ethniques, qui n'auront dorénavant aucun mal à accéder à l'informatique (à condition qu'elles disposent effectivement d'ordinateurs...).

Détrompons-nous ! Tout ce formidable édifice a été bâti sur l'assomption de la section précédente, c'est-à-dire qu'il est pertinent de traiter le texte de deux manières différentes : en utilisant des *caractères* pour tous les traitements linguistiques et des *glyphes* pour tous les traitements visuels.

Nos yeux voient (« lisent ») des glyphes, nos mains écrivent des glyphes, lorsqu'on imprime un texte c'est toujours avec des glyphes. Mais lorsqu'on veut traiter ce même texte dans la mémoire de l'ordinateur, on se permet la simplification énorme de passer des glyphes aux caractères. Et pas à n'importe quels caractères – puisque après tout, un caractère étant la description d'une classe de glyphes, celle-ci pourrait être arbitrairement raffinée selon les besoins – mais uniquement les caractères contenus dans Unicode. Et qui nous garantit alors qu'une fois ressorti de l'ordinateur, et donc reconverti en glyphes, notre texte sera toujours le même ?

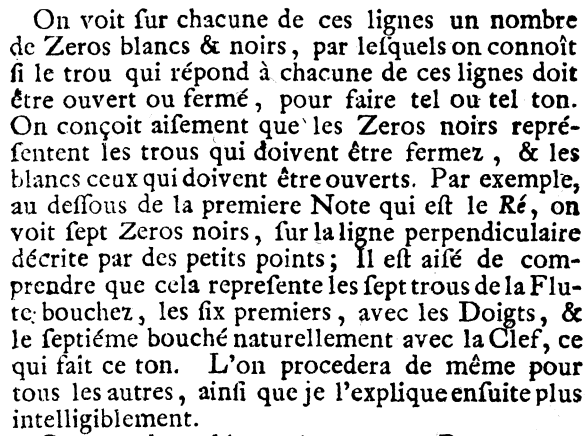
1.5. *Contre-exemples*

Quelques exemples classiques qui montrent que le dualisme caractères/glyphes n'est pas toujours la solution optimale :

1.5.1. *L'allemand gothique*

Le philosophe Kant et l'écrivain Goethe étaient des inconditionnels de l'écriture gothique. Ils n'auraient en aucun cas permis que leurs textes soient composés en romain. D'ailleurs, en Allemagne, à leur époque, cette écriture était appelée « écriture allemande », ce qui laisse supposer que bien de gens la considéraient comme un invariant de la langue allemande.

Si Unicode avait été défini dans les années 1830, il est clair que l'écriture gothique aurait eu sa propre table de caractères. Aujourd'hui on considère le gothique comme un choix purement esthétique et on classe les glyphes gothiques dans les mêmes caractères que les glyphes romains. Et les textes de Kant et de Goethe datent bien de cette époque-là ; les coder en Unicode, en oubliant délibérement la distinction sociale,



On voit ſur chacune de ces lignes un nombre de Zeros blancs & noirs, par leſquels on connoît ſi le trou qui répond à chacune de ces lignes doit être ouvert ou fermé, pour faire tel ou tel ton. On conçoit aiſement que les Zeros noirs repréſentent les trous qui doivent être fermez, & les blancs ceux qui doivent être ouverts. Par exemple, au deſſous de la premiere Note qui eſt le *Ré*, on voit ſept Zeros noirs, ſur la ligne perpendiculaire décrite par des petits points; Il eſt aiſé de comprendre que cela repreſente les ſept trous de la Flute bouchez, les ſix premiers, avec les Doigts, & le ſeptiéme bouché naturellement avec la Clef, ce qui fait ce ton. L'on procedera de même pour tous les autres, ainſi que je l'explique enſuite plus intelligiblement.

Figure 6. *Extrait d'un ouvrage français imprimé en 1728 à Amsterdam. Ici – jusqu'à preuve du contraire – le choix entre 's rond' et 's long' peut être fait par l'ordinateur : on peut donc coder ce texte avec un seul caractère pour la lettre 's' et différer le problème de choix de glyphe pour la lettre 's' à l'étape de rendu*

„Gewiß", ſchrie der Senator, deſſen Geſicht und faltiger Nacken rot angelaufen waren. „Was ſollte ich ſonſt wohl benutzen? Aber Gleichmann & Buſſe ſind aus purer Geldgier auf den Gedanken gekommen, die Schiffsmaſchinen hier am Grasbrook zu bauen. Und ich habe einige hundert Arbeiter zur halben Paſſage von Hull nach Hamburg gebracht."
„Scotch", ſagte er zu dem herbeigeeilten Kellner, „but make it a big one. — It's to my future bride."
„Aber Herr Overholt!" ſagte der Kellner, „Scotch? Am frühen Nachmittag?"
„Sie haben ein gutes Herz, Lulubelle", ſagte Oskar.
„Ich war ſelbſt ſo ein Kind", ſagte Lulubelle, „ich weiß, was ein Schilling für ſie bedeutet."

Figure 7. *Extrait d'un ouvrage allemand imprimé en 1939 à Leipzig. Ici le choix entre 's rond' et 's long' est un choix linguistique (exemple caractéristique de l'arbitraire de ce choix : alors que* Schiffsmaſchinen *et* Glasbrook *sont clairement des mots composés – ce qui justifie le 's' rond – le prénom* Oskar *n'en est pas un et pourtant contient un 's' rond) et ne peut être fait par le moteur de rendu : l'utilisation de caractères Unicode différents pour ces deux lettres s'impose, ce qui entraîne une prédestination du texte codé à une composition en écriture gothique ou romaine. Notez le passage du gothique au romain pour les mots anglais*

historique et linguistique de l'époque entre écritures « allemande » et « antique[8] », n'est-ce pas trahir la volonté et l'esprit des textes de ces auteurs ? N'est-pas un manque de respect que de dépouiller ces textes d'une propriété essentielle aux yeux de leurs auteurs ?

Le lecteur répondra peut-être que le passage par Unicode n'y est pour rien, et qu'un choix de police gothique ferait l'affaire. En fait il n'en est rien : primo, parce que le choix du gothique n'est jamais global : certains mots – ceux qui sont d'origine latine – seront toujours composés en romain. Le passage du gothique au romain doit donc logiquement être signalé d'une manière ou d'une autre dans le document. Secundo, parce que l'allemand écrit en gothique utilise deux glyphes pour la lettre 's' : un 's' court, semblable au 's' latin, et un 's' long, semblable à un 'f' mais sans barre horizontale. Unicode dispose bien d'un caractère pour le 's' long, mais on s'aperçoit alors qu'un texte destiné à être composé en gothique sera codé de manière différente (c'est-à-dire, en utilisant des caractères Unicode différents) que le même texte destiné à être composé en romain. Donc, un texte de Kant ou de Goethe codé en Unicode et stocké sur support informatique, sera *prédestiné* à être composé dans l'une ou l'autre des deux écritures, ce qui ne serait jamais arrivé si le gothique se limitait au choix des glyphes[9]. On voit déjà que la séparation artificielle entre Unicode et typographie présente une faille importante.

1.5.2. *Les ligatures de la dévanagari*

En dévanagari, écriture indienne utilisée pour des langues telles que le hindi et le marathi, mais aussi pour le sanskrit, vénérable ancêtre des langues indo-européennes, on utilise des ligatures entre les consonnes sans voyelle, que la voyelle soit explicite ou intrinsèque. Le nombre de ligatures utilisées dépend de la langue et même du niveau de langue (littéraire, technique, parlée, etc.).

Un choix de ligature peut être partiellement ou intégralement sémantique (ou lié à des critères sémantiques), il serait donc utile de considérer ces ligatures comme des caractères. Il n'en est rien : dans tous les codages actuels, les caractères sont des lettres « atomiques », et la ligature est perçue comme un phénomène d'interaction de glyphes. En passant donc par Unicode, on perd une information qui, aux yeux du débutant en ces langues, peut paraître dérisoire, mais qui se revèle être importante lorsqu'on souhaite faire de la typographie de qualité.

Les puristes d'Unicode répliqueront que le problème des ligatures a été résolu une fois pour toutes avec l'introduction du caractère ZWNJ (`ZERO WIDTH NON JOINER`) dont le rôle est d'empêcher la formation des ligatures. Mais, dans le cas de la dévanagari, son utilisation rend la saisie du texte peu commode : d'après les recommanda-

8. En allemand, le romain est appelé « Antiqua », c'est-à-dire : antique.

9. En effet, prenons une lettre 's long' d'un tel texte : si le choix de l'écriture gothique se limitait aux glyphes, alors le glyphe de cette lettre, que le texte soit écrit en gothique ou en romain, devrait appartenir au même caractère ; mais il n'en est rien puisqu'il appartient à deux caractères différents : le caractère « lettre latine s » (lorsque le texte est écrit en romain) et le caractère « lettre latine s long » (lorsque le texte est écrit en gothique).

Figure 8. *Trois manières différentes d'écrire* vva *en dévanagari : en ligature sanskrite, en (pseudo-)ligature hindi, sans ligature mais avec un signe* virama *pour indiquer qu'il n'y a pas de voyelle intrinsèque entre les consonnes. Sous Unicode on coderait : (a)* 0935 0935 *(où* 0935 *est le caractère* DEVANAGARI LETTER VA*), (b)* 0935 200C 0935 *(où* 200C *est le caractère ZWNJ), (c)* 0935 094D 0935 *(où* 094D *est le caractère* DEVANAGARI SIGN VIRAMA*)*

tions d'Unicode, en l'absence de ZWNJ, toute ligature possible – et dont un glyphe est disponible – sera effectivement formée ; il incombe donc à l'auteur du texte de prévoir *toutes* les ligatures possibles et imaginables et d'empêcher leur formation en ajoutant des caractères invisibles dans le texte...

Cette méthode résoud peut-être certains problèmes pratiques mais démontre en même temps le mal fondé de l'approche : ZWNJ est un caractère Unicode, mais ne correspond pas vraiment au profil typique de caractère Unicode, ni à la définition de caractère que nous avons donnée, c'est-à-dire la *description d'une classe de glyphes*, puisqu'il est dépourvu de glyphe. Ce problème vient du statut bâtard des ligatures, et cela ne se limite pas à la dévanagari mais concerne toutes les ligatures : si les glyphes 'f' et 'i' appartiennent clairement aux caractères *lettre latine f* et *lettre latine i*, le glyphe 'fi' n'appartient à aucun caractère. Il est considéré en quelque sorte comme un objet « dynamique » : il apparaît lorsque l'on place les glyphes 'f' et 'i' côte à côte. Le plus souvent ce phénomène ne concerne que les glyphes : par exemple, du point de vue linguistique les mots 'fille' (avec ligature) et 'fille' (sans ligature) sont strictement les mêmes, du moins en français.

Le « caractère » ZWNJ a donc été inventé pour agir de cette façon au niveau des glyphes : une décision assez controversée puisque les caractères ne sont normalement pas censés agir à ce niveau. Pour donner une analogie (pseudo-)philosophique, peut-on imaginer une idée platonicienne descendre de son piédestal et venir dans le monde réel pour indiquer aux objets qu'elle représente comment se comporter ?

Ce caractère bâtard n'est pas trop gênant dans le cas des langues occidentales[10] Mais pour la dévanagari il devient absurde : il y a foule de ligatures et l'utilisateur

10. ZWNJ a une certaine utilité pratique en allemand, où l'écriture Aufhge (avec ligature) est fausse, et l'écriture Auflage (sans ligature), qui utilise donc ce caractère, correcte. Mais on peut toujours argumenter que le logiciel qui fait le *rendu*, c'est-à-dire le passage des caractères aux glyphes, devrait tout aussi bien être capable d'introduire cette information *a posteriori*, de la même manière que les auteurs allemands n'indiqueront jamais quoi que ce soit de tel dans leurs textes et que ce sont bien les imprimeurs qui ont la responsabilité d'appliquer ou non la ligature.

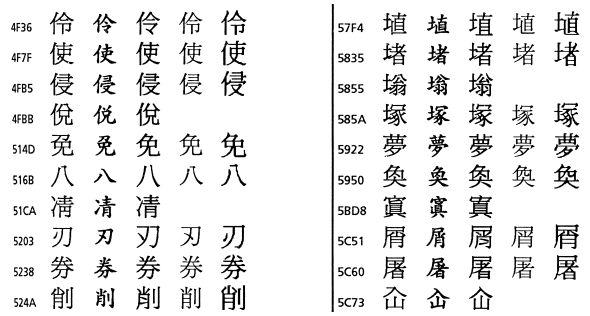


Figure 9. *Quelques exemples de variantes d'idéogrammes qui ont été unifiées : certaines parmis ces variantes sont considérées comme des idéogrammes différents dans un pays à écriture idéographique mais dans un autre. Cette liste d'exemples est tirée du livre d'Unicode version 1*

ne peut pas se rappeler de toutes les ligatures possibles pour essayer de les éviter en plaçant des ZWNJ.

1.5.3. *Les idéogrammes CJCV*

Un autre exemple classique : les idéogrammes chinois, japonais, coréens et viêtnamiens. *A priori* ces idéogrammes ne devraient pas poser trop de problèmes puisqu'ils sont basés sur une grammaire notationnelle ; en effet, il existe des éléments graphiques de base appelés *radicaux* et des méthodes de combinaison des radicaux pour former les idéogrammes. Mais en réalité, les pires choses arrivent : deux glyphes d'idéogrammes légèrement différents peuvent être considérés comme des caractères différents en Chine, mais pas au Japon, et inversement. L'appartenance des glyphes à un caractère serait donc, non pas une fonction du codage, comme prévu par Unicode, mais une fonction de la langue[11]. En outre, deux raisons aggravent encore la situation :

1) n'ayant trouvé aucun moyen de décrire les idéogrammes avec des mots de la langue anglaise, Unicode se base exclusivement sur leur glyphe privilégié. Et ce glyphe ne peut être chinois, japonais et coréen simultanément ;

2) de nouveaux idéogrammes sont inventés ou découverts dans des vieux textes tous les jours. N'ayant pu implémenter aucune grammaire notationnelle générative pour décrire des idéogrammes quelconques à partir des radicaux (qui, eux, ne changent pas), on est obligé d'attendre leur insertion dans Unicode pour pouvoir les utiliser.

On se rend compte qu'autour des idéogrammes règne un flou artistique que l'imprimerie n'a pu éviter que par la possibilité de création de nouveaux signes typographiques à tout moment. Cette méthode pourrait être adoptée au contexte informa-

11. Sans parler des problèmes que posent aux imprimeurs japonais et chinois les idéogrammes de mots chinois dans les textes japonais et inversement...

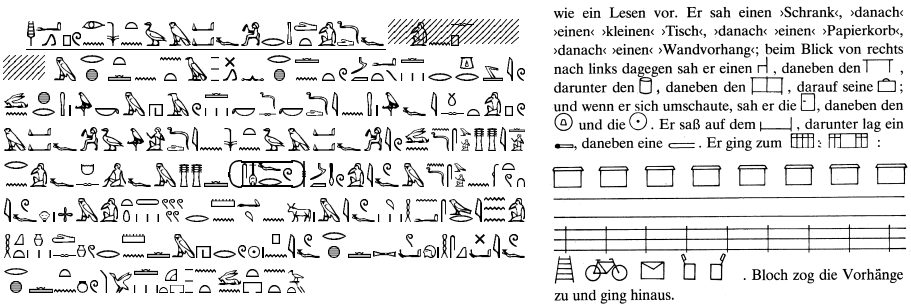


Figure 10. *À gauche : un texte écrit en hiéroglyphes égyptiens. À droite : des « hiéroglyphes modernes » utilisés par Peter Handke pour montrer le désarroi du protagoniste (meurtrier et légèrement déboussolé) devant les mots, les choses et leurs rapports...*

tique[12] mais ce serait un échec total pour Unicode : en effet, ce qui s'avère faux, c'est l'assomption implicite d'Unicode que tout système d'écriture peut être représenté informatiquement par un ensemble fini et bien défini de caractères. Les idéogrammes sont la preuve vivante du contraire, puisque, d'une part, ils ne cessent de croître en nombre et d'autre part ils ne peuvent être décrits convenablement par des mots de la langue anglaise comme tous les autres caractères Unicode.

Des problèmes similaires se posent pour les hiéroglyphes : les hiéroglyphes égyptiens ont été classifiés par les égyptologues et pourront donc facilement être intégrés dans Unicode (si on fait abstraction des problèmes de directionnalité que nous discuterons plus loin). Mais qu'en est-il des « hiéroglyphes modernes », comme ceux utilisés par Peter Handke dans l'*Angoisse du gardien de but avant le penalty* (Suhrkamp, Francfort s/ Main 1970), qui font clairement illusion aux hiéroglyphes égyptiens ? Va-t-on unifier les hiéroglyphes égyptiens et les hiéroglyphes modernes ? Comment gérer l'afflux de nouveaux signes ?

2. Les mécanismes d'Unicode

Après cette introduction à Unicode, nous allons examiner ses mécanismes fondamentaux de plus près : il s'agit des notions de *glyphe privilégié*, de *description*, de *caractère combinant* et de *propriétés*.

12. Il existe déjà des logiciels permettant de produire des nouveaux idéogrammes à partir d'une combinaison des radicaux et des modifi cations manuelles. Leur qualité est encore moyenne, mais cela risque de changer rapidement, tant le marché est prometteur.

2.1. *Les glyphes privilégiés*

Un *glyphe privilégié* est un glyphe choisi pour représenter dans le livre d'Unicode (sur papier, ou en ligne) le caractère auquel il appartient. Ce glyphe joue un rôle très important puisqu'il participe officieusement à la description du caractère. Pour la même raison il pose aussi souvent des problèmes. On peut dire qu'il contient trop d'information, et en même temps pas assez. Trop d'information puisque, inévitablement, certains éléments, parties ou propriétés du glyphe privilégié ne seront pas valables pour toutes les glyphes du caractère, et cela peut induire en erreur. Pas assez d'information puisqu'il ne peut pas être représentatif de toutes les variantes[13] des glyphes de sa classe. Ainsi, le lecteur du livre a du mal à saisir l'étendue de la classe des glyphes au seul vu du glyphe privilégié.

Il y a également un danger qui se cache dans les glyphes privilégiés : celui de l'aliénation de la tradition typographique, surtout pour des écritures non latines. En effet, les concepteurs de fontes non latines, n'étant souvent pas locuteurs d'une langue utilisant ce système d'écriture ne peuvent pas toujours se permettre d'investir du temps et de l'énergie dans des recherches sur la tradition typographique des écritures de leurs fontes. Unicode leur propose un formidable « recueil de glyphes » sur lequel ils peuvent se baser pour dessiner ou adapter leurs polices de caractères. Cela est d'autant plus vrai qu'Unicode contient souvent des caractères très rares et inexistant dans les polices du marché, ce qui entraîne un manque de répères pour les dessinateurs. Et cela peut avoir des conséquences dramatiques dans le pays où l'écriture est utilisée, surtout lorsque ce pays est sans défense devant l'invasion de polices de caractères créées à l'étranger. Un exemple classique est celui de l'accent vertical en grec : dans le livre d'Unicode version 2, l'accent aigu du grec est représenté par des glyphes (très peu esthétiques, soit-il dit en passant) ayant des accents verticaux (voir fig. 11). L'accent vertical n'a jamais existé en grec, du moins avant l'arrivée des premières polices américaines le contenant... et subitement on a donc vu des livres composés avec l'accent vertical. Le système « monotonique » est déjà un crime contre la langue grecque, voilà que cet unique accent a maintenant changé de forme, et tout cela à cause d'Unicode. Heureusement, de nombreux utilisateurs (dont l'auteur) se sont plaints auprès du consortium Unicode, et dans la version 3, la forme de l'accent a été corrigée. On a frisé la catastrophe éco-typographique...

En fait, le problème vient de la mauvais utilisation des tables Unicode : les glyphes privilégiés ne prétendent aucunement être normatifs (et encore moins, esthétiques) et c'est uniquement pour des raisons de facilité que les fonderies de polices de caractères se basent sur eux pour adapter leurs polices à Unicode. Il n'est pas du ressort d'Unicode de définir des règles *typographiques* concernant les caractères et les fonderies (américaines pour la plupart) n'ont aucune autre ressource à leur disposition. Ainsi, par exemple, au fur et à mesure que des polices sont adaptées à Unicode, on a de plus en plus de variantes graphiques du 'L' pointé catalan, les unes plus maladroites que les autres, alors qu'une règle existe : « le point doit être placé à mi-hauteur des lettres capitales, à mi-distance entre les troncs verti-

13. Dans la version 1 d'Unicode on trouvait encore des glyphes privilégiés alternatifs pour indiquer les variantes possibles. Exemple : le caractère `DOLLAR` était représenté par deux glyphes, l'un avec une barre verticale et l'autre avec deux barres. Aujourd'hui cette pratique a été abolie.

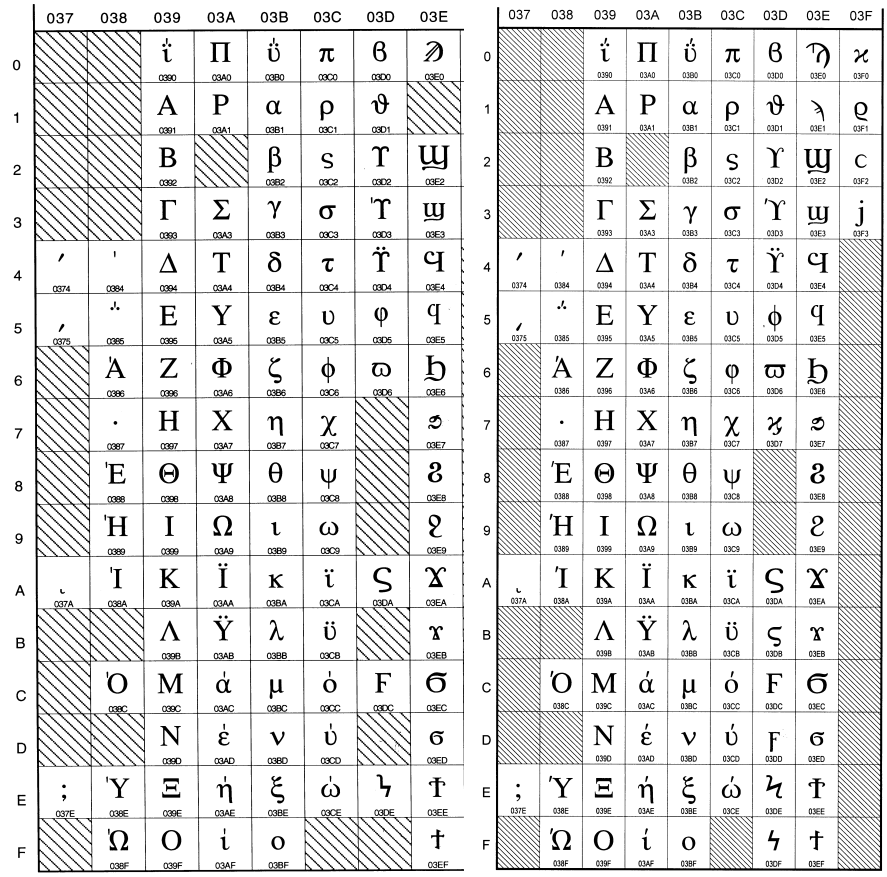


Figure 11. *À gauche : table des glyphes privilégiés du grec estropié (autrement dit : monotonique) tirée du livre d'Unicode version 2. Notez les accents verticaux. À droite : la même table dans le livre d'Unicode version 3. Les accents ont été rectifiés*

caux des lettres, et il ne doit pas entraîner de crénage entre les lettres »[14]. Le caractère Unicode correspondant étant `LATIN CAPITAL LETTER L WITH MIDDLE DOT` et non pas, par exemple, `LATIN CAPITAL LETTER L WITH DOT A MID-CAPITAL LETTER HEIGHT HORIZONTALLY EQUIDISTANT FROM BOTH STEMS`, il est tout à fait normal que la tradition typographique ne tienne qu'à un fil. Et d'ailleurs le glyphe privilégié de cette lettre ne correspond pas du tout à la règle énoncée précédemment, ce qui est un mauvais présage.

En même temps, cette « globalisation » de la typographie, comme toute globalisation, appauvrit les typographies nationales, en éliminant les différences entre elles. Les trémas (en fait des *Umlaut*) des fontes allemandes sont traditionnellement placés

14. Gabriel Valiente Feruglio, *Modern Catalan Typographical Conventions*, TUGboat, t. 16 (1995), n° 3, p. 329–338.

plus bas que les trémas des fontes françaises. Comment cette différence culturelle[15] pourrait-elle survivre si tout le monde se base sur les mêmes glyphes privilégiés ?

Cela démontre les ambiguïtés des rapports entre Unicode et la typographie : si Unicode n'utilise qu'un seul glyphe privilégié par caractère, c'est qu'il suppose que l'utilisateur connait les variantes possibles, c'est-à-dire quelles sont les variantes effectivement utilisées dans la tradition typographique. Ainsi, pour pousuivre l'exemple donné en note 13, si Unicode version 3 représente le DOLLAR par un signe à une seule barre, comment savoir si le signe à deux barres verticales fait partie des glyphes du même caractère ? Et est-ce qu'un signe à trois barres est possible ? Ici les réponses sont faciles à donner puisqu'il s'agit d'un signe monosémique et que sa description est basée sur sa signification : il n'a qu'une seule signification, celle du symbole de la monnaie « dollar », et la description du caractère est justement DOLLAR. On peut donc piocher dans la tradition typographique pour trouver toutes les variantes historiques de ce symbole, et même en inventer des nouvelles[16] : si dans mon contexte je définis clairement ⊗ comme étant le symbole du dollar, je peux, en théorie, utiliser le caractère Unicode DOLLAR pour coder ce glyphe dans un texte, puisqu'il correspond à cette signification.

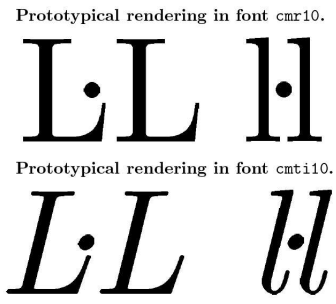


Figure 12. *Normalisation du L pointé catalan proposée par Gabriel Valiente Feruglio en 1995*

Les problèmes commencent quand les descriptions ne sont plus purement linguistiques ou logiques ; nous allons en parler dans la section suivante.

Un dernier problème résultant des glyphes privilégiés : le risque de confusion pour les utilisateurs non avertis, surtout en micro-informatique. En effet, jusqu'à l'arrivée d'Unicode, le grand public n'avait aucun contact avec les codages de caractères. D'une part, les tables des codages ISO devaient être achetées auprès de l'AFNOR, chose qu'aucun simple utilisateur de micro-informatique ne ferait, et d'autre part les codages de caractères commerciaux (DOS, Windows, MacOS, etc.) n'étaient décrits que dans certains ouvrages ultra-spécialisés. Le grand public n'entrait donc en contact qu'avec des codages de fonte, à travers des outils du type « Insertion de caractères spéciaux » (Windows) ou « PopChar Pro » (Macintosh) (voir fig. 13 et 14). Bien évidemment, à travers ces outils, en choisissant une police suffisamment « régulière » on obtenait une

15. Qui, comme toute différence historique entre deux cultures peut toujours se justifi er *a posteriori* par des (pseudo-)critères : en français le tréma agit sur le mot, puisqu'il sépare deux lettres entre elles, en allemand il agit sur la seule lettre, puisqu'il change sa prononciation...

16. Notons que le dollar à deux barres verticales représente également le symbole de Reis, et que dans certains spécimens de fonderie (comme celui de Deberny et Peignot) on trouve également des dollars à deux barres *horizontales* (information communiquée à l'auteur par Jacques André).

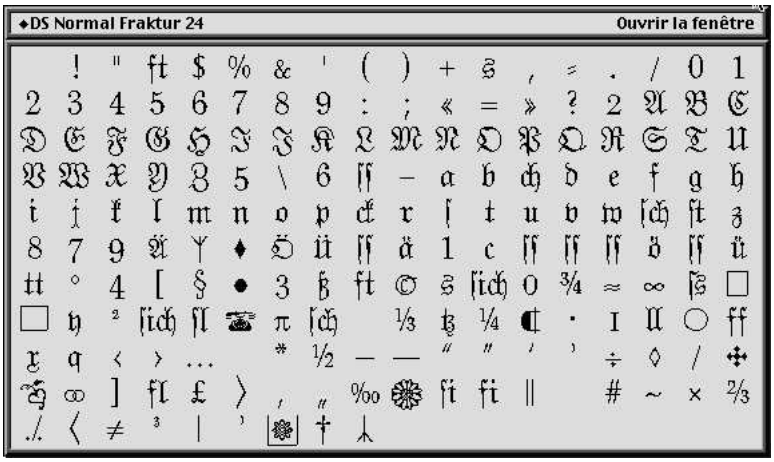


Figure 13. *L'utilitaire « PopChar Pro » sur Macintosh. La police dont la table de glyphes est affichée ici est une police gothique très réussie de la fonderie Delbanco (R.F.A.). On peut voir nombre de ligatures (qui ne correspondent donc à aucun caractère Unicode) et son codage est tout sauf standard, le comble de l'exotisme étant qu'à la place des caractère 'c', 'q', 'x', 'y', qui occupent pourtant les mêmes positions dans tous les codages dérivant de l'*ASCII*, on trouve les ligature gothique 'ch', 'ck', 'sch' et 'st'...*

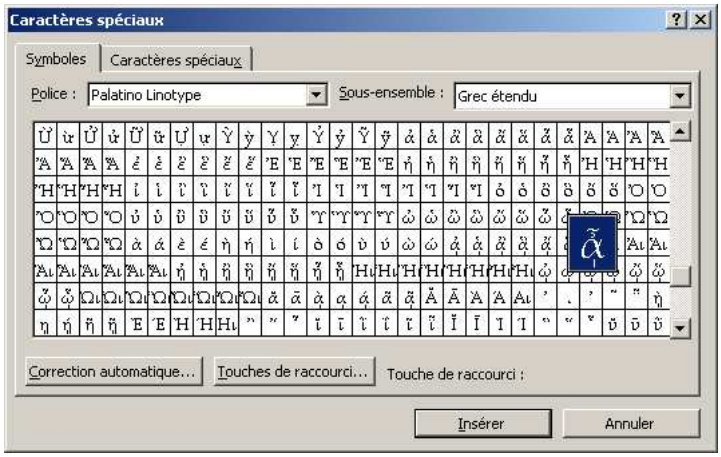


Figure 14. *La fonction « Insertion de caractères spéciaux » de Microsoft Word 2000. Les glyphes affichés proviennent de la table grecque d'Unicode, pourtant rien n'indique le code Unicode ni la description du caractère auquel est censé appartenir le glyphe sélectionné*

table de glyphes privilégiés, ce qui pourrait sembler une bonne approximation des caractères. Mais les glyphes privilégiés ne suffisent pas pour définir un codage, puisque les caractères sont, par définition, des *descriptions* et que ces outils ne proposent rien de tel[17]. Les outils les plus récents essaient de pallier ce problème en affichant également les positions Unicode des caractères correspondant aux glyphes : faute de donner la description du caractère auquel appartient le glyphe, on donne ainsi une référence à sa description dans le codage Unicode.

2.2. *Les descriptions*

Contrairement aux glyphes, les descriptions de caractères sont *normatives*, et donc bien plus importantes. En effet, beaucoup d'encre a été versée sur les problèmes de celles-ci. À commencer par l'utilisation de mots anglais[18]ou se servant de la prononciation de l'anglais dans les descriptions : pour comprendre quel est le caractère ARABIC LETTER MEEM il faut le prononcer à l'anglaise : les deux 'e' se prononcent 'î' et on a donc 'mîm' (et non pas 'mehem' comme on aurait tendance à prononcer MEEM en français). Mais ces pratiques ne sont pas uniformes : souvent des mots dans la langue la plus représentative du système d'écriture en question sont utilisés, et ceci avec les moyens du bord. Ainsi, on a dans Unicode des caractères du genre GREEK SMALL LETTER UPSILON WITH PSILI AND OXIA (description contenant 8 mots dont 3 grecs et 5 anglais), mais aussi ARABIC LETTER DAL WITH RING où il n'y a qu'un mot non anglais, le nom de la lettre.

---

17. Ce qui est parfaitement normal, puisque leur but était l'accès aux tables de glyphes des fontes et non pas l'accès à ces objets abstraits que sont les caractères. Ce qui mérite d'être remarqué ici est le fait que dans le système de classification pseudo-visuel des glyphes, ce qui jouait le rôle de classificateur était le *clavier virtuel*, c'est-à-dire la correspondance entre les touches du clavier réel et les glyphes de la fonte. Pour les créateurs et utilisateurs de polices non latines, par exemple, ce qui importait était d'avoir un « clavier adapté à la police ». Peu importe où étaient placés les glyphes dans la table de la police, du moment que l'utilisateur y accédait par les mêmes touches. Cette approche est catastrophique pour le stockage et la transmission des informations : le même glyphe est associé à autant de positions différentes, et donc à des nombres dans la mémoire de l'ordinateur, qu'il y a des polices utilisées, et ceci dans le même document... Pour pallier à ce genre de problèmes, Adobe a introduit le concept de *nom de glyphe normalisé*.

18. Des traductions de ces descriptions dans d'autres langues existent : notons, en particulier, la traduction française d'Unicode, par Patrick Andries (disponible sur `http://hapax.iquebec.com`). Mais toute traduction comporte un risque de déviation du sens original, surtout lorsqu'il s'agit de décrire des caractères dans des systèmes d'écriture totalement différents. Sans parler des problèmes bien connus de toute traduction de terminologie scientifique ou technique. À titre d'exemple, prenons le caractère SOLIDUS, autrement dit : la barre oblique. En grec, dans un contexte bureautique, ce caractère s'appelle depuis toujours κάθετος, c'est-à-dire « verticale ». Et pourtant il existe un autre caractère (VERTICAL LINE), qui est représenté par un véritable trait vertical, et qui est inconnu en Grèce, puisqu'absent des claviers de machine à écrire grecs. Comment traduire SOLIDUS sans inventer des nouveaux termes et sans tomber dans les paradoxes logiques du type « verticale oblique » ?

Ces problèmes sont des bagatelles comparés à ceux posés par les caractères *à description graphique ou typographique* : un vrai paradoxe puisqu'en contradiction avec la définition même du caractère. Voici quelques uns de ces problèmes :

1) problème de précision de description graphique : lorsqu'on a les descriptions BULLET et BLACK CIRCLE, quel est le sens exact de chacune, autrement dit : quand dira-t-on qu'un glyphe appartient au premier ou au second caractère ? À quel moment une puce devient-elle un cercle (en fait, un disque) ? On peut trouver des règles intuitives (du genre : « une puce aura le diamètre de la lettre 'o' minuscule et sera centrée verticalement par rapport à l'axe des symboles mathématiques, alors que le disque noir aura le diamètre du 'O' majuscule et sera placé sur la ligne de base »), mais ces règles ne font pas partie du standard, voire ne sont souvent même pas mentionnées dans le livre d'Unicode, ce qui confère un flou artistique supplémentaire à ces caractères.

2) problème de mélange graphique-typographique : un chef-d'œuvre de telle description est SINGLE HIGH-REVERSED-9 QUOTATION MARK, qui n'est autre que le simple guillemet ouvrant américain ‘ (que l'on peut appeler aussi « apostrophe inversée »). Dans cette description on a une composante typographique : « quotation mark », c'est-à-dire « guillemet » qui bien sûr est 100 % anglocentrique puisqu'il s'agit bel et bien d'un guillemet américain. Mais on a aussi une composante graphique : « high-reversed-9 » qui peut être analysée en « un signe similaire à un 9 inversé, placé en position d'exposant ». À part l'observation enfantine sur le fait qu'un 9 inversé est tout simplement un 6, on peut se poser des questions sur l'utilité de cette approche. À notre connaissance, le signe typographique dont il est question ici n'est utilisé que comme guillemet ouvrant dans certaines langues, dont l'anglais britannique et américain. N'eût-il pas été plus pertinent de l'appeler SINGLE OPENING ENGLISH QUOTATION MARK, puisque tel est son rôle ?

Autre exemple du même type : l'esprit doux grec est appelé tantôt PSILI, qui est la translitération de son nom grec $\psi\iota\lambda\acute{\eta}$, tantôt COMBINING COMMA ABOVE qui en est une description graphique très maladroite : d'une part, on voit mal ce qu'une virgule vient faire au-dessus d'une lettre, alors que l'on a depuis quelques siècles l'habitude d'appeler une virgule en position haute une apostrophe ; d'autre part, et cela est le comble de l'ironie, dans la typographie grecque les esprits sont plus ronds que les virgules/apostrophes, et donc cette description est non seulement maladroite mais fausse, puisque l'esprit n'a *pas* la forme de la virgule.

3) problème de mélange de descriptions linguistiques et (typo)graphiques dans le même contexte. Prenons un contexte très normalisé et fonctionnel : l'alphabet phonétique international. Alors que l'on aurait pu s'attendre à des descriptions plus rationnelles, on assiste au même mélange linguistique/(typo)graphique : le caractère LATIN LETTER PHARYNGEAL VOICED FRICATIVE (voyelle fricative pharingale) coexiste avec le caractère LATIN SMALL LETTER TURNED R WITH LONG LEG (« lettre r inversée à pied long » [traduction de l'auteur] ou « r prolongé culbuté » [traduction officielle ISO par Patrick Andries])... Heureusement dans la partie non normative on trouve des explications, parfois assez pittoresques : ainsi on nous explique que le caractère LATIN SMALL LETTER TURNED T représente le son « tsk tsk ».

Autre exemple : la lettre arabe ḥah avec trois points au-dessous, utilisée en farsi, ourdou, pachto, sindhi, etc. est appelée ARABIC LETTER TCHECH (et donc par son

nom persan ou ourdou, ce qui en fait une description linguistique) alors que la même lettre avec trois points au-dessus, utilisée surtout en sindhi, est appelée ARABIC LETTRE HAH WITH THREE DOTS ABOVE (description graphique).

2.3. *Les caractères combinants*

Contrairement à ses prédécesseurs (ISO 8859 et codages « commerciaux ») Unicode ne se contente pas de définir des caractères dont les glyphes forment des unités graphiques atomiques sans interaction, mais introduit un nouveau concept : celui de *caractère combinant*. Il s'agit de caractères dont les glyphes – comme leur nom l'indique – se combinent avec les glyphes des caractères qui les précédent (y compris d'autres caractères combinants) pour produire des glyphes composés.

Le cas typique de caractère combinant est celui d'un accent ou d'un signe diacritique, par exemple : la cédille. Unicode proclame que, moyennant une *normalisation*, la suite de caractères LATIN SMALL LETTER C et COMBINING CEDILLA est *canoniquement équivalente* au caractère LATIN SMALL LETTER C WITH CEDILLA. Cette équivalence canonique d'un caractère composé avec la suite de ses composantes est une information normative qui accompagne la description du caractère (dans le livre d'Unicode, cette information se trouve sous la description du caractère et est précédée d'un signe d'équivalence[19] ≡).

Dans le rapport technique 15, annexe normative d'Unicode, l'on décrit de manière opératoire deux normalisations :

1) la décomposition normalisée (NFD), où tout caractère possédant une décomposition par équivalence canonique est remplacé par celle-ci. Par exemple, le c cédille devient une suite de deux caractères, le c et la cédille, dans cet ordre. Cette normalisation est réitérée jusqu'à ce qu'il n'y ait plus de caractère décomposable ;

2) la composition normalisée (NFC), où après un passage par NFD, toute chaîne consistant en un caractère non combinant (un *starter* dans le jargon Unicode) suivi de caractères combinants et qui peut être interprêtée comme étant canoniquement équivalente avec un caractère Unicode, est remplacée par celui-ci.

Malheureusement il n'y a aucune stratégie claire sur le choix des caractères composés contenus dans Unicode. Le standard se trouve divisé entre deux tendances contradictoires : d'un côté les puristes qui prétendent que lorsqu'un glyphe peut être obtenu par la composition des glyphes de deux ou plusieurs caractères, alors nul besoin d'introduire un nouveau caractère pour celui-ci. De l'autre côté, ceux qui considérent les lettres avec signes diacritiques comme des lettres « avec les mêmes droits (linguistiques) que les autres » et préconisent, donc, l'introduction massive de ceux-ci dans le standard, sous forme de caractères précomposés. On assiste souvent à des

19. À ne pas confondre avec les informations précédées d'une flèche →, appelées *références croisées* et qui ne sont que de simples rappels d'autres caractères dont les glyphes ressemblent vaguement à celui du caractère courant. Par exemple, les éditeurs d'Unicode se sentent moralement obligés de nous rappeler, par le biais d'une référence croisée, que la ligature œ ressemble au digraphe scandinave æ... Bien heureusement ces informations n'ont aucune valeur normative.

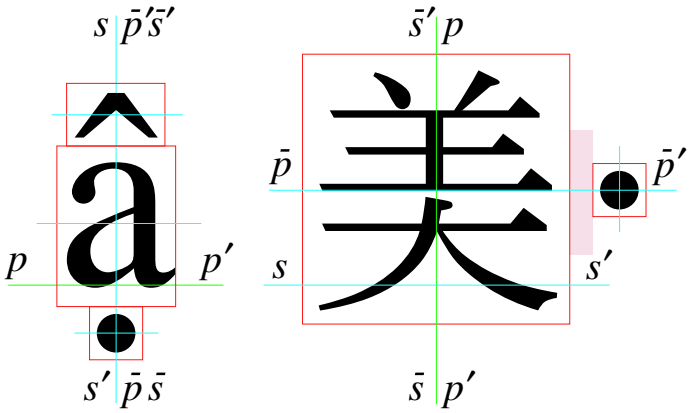


Figure 15. *Axes d'accentuation pour une lettre latine et un idéogramme*

batailles entre ces deux camps (notamment aux conférences Unicode), surtout depuis que le consortium Web a clairement pris position en faveur des caractères précomposés.

Qu'en est-il vraiment ? On assiste ici, de nouveau, à une relation illégitime entre Unicode et la typographie, qui implique notamment les fabriquants de logiciels de traitement de texte. En effet, un logiciel « compatible Unicode » devrait être compatible avec la combinatoire des caractères dans toute son étendue, autrement dit le fait de générer de nouveaux glyphes à partir des glyphes existant dans les fontes, et ceci pour *toute* combinaison possible et imaginable entre glyphes de caractères « inertes » et glyphes de caractères combinants. Ceci est très difficile, pour deux raisons :

1) pour des raisons inhérentes aux systèmes d'écriture et à la tradition typographique. Ainsi, par exemple, l'espéranto utilise un h avec accent circonflexe. Cet accent doit être centré sur le tronc ascendant de la lettre. Cela suppose que le système en question connaît les différents axes de centrage des diacritiques. Ces axes ne sont pas toujours verticaux, même si la police en question est droite : par exemple, en arabe, l'axe de placement des voyelles et autres signes diacritiques est souvent incliné : son inclinaison simule le mouvement de la main qui place ces signes après avoir tracé la lettre ;

2) pour des raisons inhérentes à la fonte utilisée. Si le précédent argument était plutôt général et pourrait s'appliquer aux caractères, ici l'on parle clairement de glyphes. Dans telle ou telle fonte, les accents ne se placent pas de la même manière. Idéalement, c'est le concepteur de la fonte qui décide du positionnement de chaque signe diacritique, et bien sûr il ne le fait que pour les combinaisons les plus courantes.

En guise de parenthèse, mentionnons le seul système actuellement disponible qui soit compatible avec la combinatoire des caractères Unicode et a l'infrastructure nécessaire pour être « typographiquement correct » : il s'agit du système Omega, dé-

veloppé par John Plaice (UNSW, Sydney) et l'auteur. Ce système se base sur des structures informatiques très souples pour les fontes (cf. figure 15 pour une description de quelques axes d'accentuation utiles pour l'alphabet latin et les idéogrammes japonais), et le concept de *micro-moteur typographique* : il s'agit de modules interchangeables, activés automatiquement dans un contexte donné, qui tirent profit des informations contenues dans les fontes pour composer les glyphes de manière arbitrairement détaillée.

Ainsi, pour un système d'écriture donné et une fonte donnée, le développeur peut d'une part insérer un certain nombre d'informations dans la fonte et, d'autre part, écrire un micro-moteur *ad hoc* pour générer des nouveaux glyphes à partir des glyphes disponibles et des informations contenues dans la fonte.

Mis à part Omega, la complexité inhérente à ces opérations dépasse de loin des possibilités des logiciels conventionnels, et cela ne peut laisser Unicode indifférent (puisque, après tout, le consortium Unicode est financé par des sociétés telles que Microsoft, Apple, Adobe, etc. toutes dûment mentionnées et remerciées dans les premières pages du livre Unicode).

La situation actuelle est celle du compromis entre le tout-décomposé et le tout-composé : on retrouve dans le standard certains caractères composés, pour des raisons historiques[20] ou pratiques. Dans certains cas le standard contient certaines combinaisons et pas d'autres, comme par exemple dans le cas du grec : on y trouve des voyelles epsilon avec accent aigu et grave, mais pas avec accent circonflexe, probablement parce que l'orthographe moderne du grec ne reconnaît pas cette combinaison. Pourtant elle est bien utile pour les textes anciens (en particulier, les transcriptions de textes épigraphiques) et on la rencontre dans des livres imprimés, certes moins souvent que les autres combinaisons.

Le cas du epsilon avec accent circonflexe est intéressant parce que la disposition de la table grecque comporte des « trous » (c'est-à-dire des positions non affectées) exactement aux endroits qui correspondent aux combinaisons de ce type. Le plus intéressant est le fait que ces positions sont *réservées*[21]. Ici le manque d'expertise est cruellement apparent : les combinaisons « hérétiques », puisque non conformes aux règles actuelles, ont été bannies parce que personne parmi les décideurs – pourtant supposés compétents – ne se doutait du fait que les mêmes combinaisons ne sont plus

20. C'est l'excuse habituelle du consortium Unicode : à chaque fois que l'on rencontre une incohérence voire une erreur énorme, dans le standard, c'est toujours pour des raisons « historiques ». Pourtant, dans les manuels scolaires d'histoire on ne cesse de nous répéter qu'un des buts de l'Histoire est de nous faire éviter de commettre les mêmes erreurs que dans le passé...

21. Ce terme signifie que le consortium Unicode se réserve la possibilité d'introduire un caractère à cet endroit-là, et souvent, dans le standard, on indique lequel, par une « référence croisée ». On se refuse donc le plaisir d'introduire et d'utiliser ce caractère, mais on admet que « s'il était inclu (ou s'il le sera dans le futur), alors ce serait à cette position-là ». Ce genre de situation nous pousse à conclure que bientôt une nouvelle discipline fera surface : l'*exégèse* d'Unicode, inspirée de l'exégèse de la bible, et destinée à interpréter le dit et le non-dit unicodiens...

erronées dans un autre contexte. Cette histoire rappelle tristement l'absence de la ligature « œ » du codage ISO 8859-1, due à l'assomption (erronée) que cette ligature n'est qu'un artéfact typographique, utilisé systématiquement dans la typographie française lorsque les lettres o et e se suivent.

Mais les caractères combinants n'ont pas que de mauvais côtés. Grâce à eux, on peut, pour la première fois, coder convenablement l'écriture phonétique. En effet, l'Alphabet Phonétique International prévoit un certain nombre de signes diacritiques, qui peuvent être arbitrairement combinés. Et les phonéticiens du monde entier ont défini une foule d'autres systèmes de notation, suppléments de l'API lui-même. De même, les caractères combinants sont très utiles en notation mathématique : la tendance des mathématiciens à inventer des nouvelles notations ou des nouvelles combinaisons d'anciennes notations fait que seul un système généraliste comme les caractères combinants aurait la moindre chance de couvrir raisonnablement leurs besoins.

Quoi qu'il en soit, les caractères combinants sont une innovation en matière de codage et leur implémentation est un challenge important pour l'industrie logicielle.

2.4. *Les propriétés*

Nous avons défini un codage comme étant une association entre nombres et caractères. En réalité, le standard Unicode contient également un certain nombre d'informations additionnelles, dont certaines sont normatives et d'autres purement indicatives (commes les glyphes privilégiés) : celles-ci sont appelées *propriétés*. On a déjà mentionné la décomposition canonique et les références croisées. Il y a d'autres types de propriétés, dont voici les plus importantes :

2.4.1. *L'inégalité explicite*

Ce terme savant est utilisé pour exprimer les homographes, c'est-à-dire des caractères qui partagent (entièrement ou partiellement) les mêmes glyphes. Ainsi le point-virgule latin `SEMICOLON` et le point d'interrogation grec `GREEK QUESTION MARK` sont entièrement homographes. Par contre, le T latin `LATIN CAPITAL LETTER T` et le T cyrillique `CYRILLIC CAPITAL LETTER TE` ne sont que partiellement homographes : en écriture cursive, voire calligraphique, le T cyrillique prend une toute autre forme (celle d'un Pi majuscule avec trois troncs verticaux). L'inégalité explicite est indiquée par une flèche.

2.4.2. *La decsription alternative*

C'est le moyen d'Unicode de « sauver les meubles » lorsque la decsription d'un caractère est partielle ou maladroite. Ainsi, la description de l'esprit rude grec `COMBINING REVERSED COMMA ABOVE` que nous avons déjà mentionné comme summum de la maladresse, est dotée d'une description alternative : « *Greek dasia, rough breathing mark* », qui s'adresse donc aussi bien aux grecophones qu'aux antiquisants occidentaux en donnant une description grammaticale précise du caractère, plutôt qu'une absurde construction pseudo-typographique. Comme preuve évidente de l'existence de l'humour unicodien, notons que la description alternative du « *smilie* »

WHITE SMILING FACE est « *have a nice day !* », phrase qui fait partie des rituels communicatifs américains, et en particulier new-yorkais. La description alternative n'est pas normative et est indiquée dans le livre par un signe d'égalité =.

2.4.3. *La note informative*

Autrement dit, un commentaire quelconque. Souvent utilisée pour indiquer les langues dans lesquelles on rencontre un caractère[22]. La note informative sert aussi de point d'entrée pour diverses escapades typographiques. Ainsi, par exemple, la note informative de l'apostrophe RIGHT SINGLE QUOTATION MARK est : « this is the preferred character to use for apostrophe », autrement dit : « nous savons tous qu'il existe, pour des raisons historiques, un caractère qui s'appelle "apostrophe", mais ce caractère-là étant plutôt laid, veuillez utiliser plutôt celui-ci dans vos textes ». Cette phrase, à première vue anodine, cache un grand malaise provoqué par les « signes de ponctuation informatiques », c'est-à-dire les signes droits ' et " que l'on retrouve dans certaines fontes, et qu'Unicode a voulu différencier de l'apostrophe et du guillemet fermant américain[23]. Ici Unicode devient « conseiller typographique », par la nécessité de justification de la présence de certains caractères, qui, sans leur background typographique, ne seraient jamais entrés dans le codage.

2.4.4. *La décomposition de compatibilité*

À ne pas confondre avec la décomposition canonique, dont le rôle est de produire un glyphe en composant les glyphes de deux caractères, la décomposition de compatibilité propose un substitut possible, c'est-à-dire une suite de caractères dont les glyphes peuvent être composés pour produire plus ou moins le même résultat, mais cette opération étant soit sans fondement linguistique, soit entre caractères non combinants, et donc dont les glyphes ne sont pas censés se combiner. Exemples : la décomposition de compatibilité du l pointé catalan LATIN SMALL LETTER L WITH MIDDLE DOT, est un l suivi d'un point centré. Le caractère « point centré » est surtout utilisé en mathématiques pour le produit scalaire, et on peut très bien imaginer une expression mathématique $l \cdot m$ où une variable mathématique l est suivie d'un point centré dénotant un produit scalaire. Il serait absurde de normaliser cette chaîne de caractères en remplaçant les deux premiers par le caractère de l pointé catalan, et cela montre la fragilité de la *décomposition de compatibilité*.

Souvent il ne s'agit même pas de « décomposition », mais d'un seul caractère, qui peut, pour une raison ou une autre, remplacer le caractère courant. Ainsi, on propose le s rond comme substitut du s long. Cette substitution est très dangereuse parce qu'elle confond, encore une fois, linguistique et typographie : effectivement, un texte

22. Ainsi on découvre que notre chère ligature nationale « œ » a également existé en islandais ancien...

23. En France on se réfère à ces signes soit par la caractérisation « informatique » – appellation absurde puisque jamais l'informatique n'a revendiqué la nécéssité d'une apostrophe et d'un guillemet qui lui soient propres – soit par leurs noms anglais : la quote (prononcée « cote ») et la double quote (prononcée « double cote »). Il est triste de constater que l'on rencontre ces signes de plus en plus souvent dans les livres et autres imprimés...

Figure 16. *Ce maître d'œuvre a inventé la version* titlecase *du O dans l'e : pourtant, contrairement au « Dz » croate, cette ligature française n'a que deux casses : c'est « Œ » ou « œ », mais jamais « Œ ».* (Photo prise à Brest, le 29 août 2002)

de Montaigne peut être composé avec deux types de s ou avec un seul, et l'on peut même remplacer le s long par un s rond, sans altérer le (contenu du) texte : on peut qualifier les deux s comme un artifice de style typographique, une modélisation par l'imprimerie de l'écriture manuscrite, sans fondement grammatical. La situation est tout à fait différente dans le cas de l'allemand écrit en gothique : ici, le choix entre s long et rond joue un rôle grammatical : les mots Wachſtube (en gros Wach-stube = chambre de veille) et Wachstube (Wachs-tube = tube de cire) sont totalement différents, et la substitution proposée par Unicode entraîne, si on compose en écriture gothique, une corruption flagrante du texte [cf. également note 9 et fig. 6].

Les décompositions de compatibilité semblent être le reflet d'un problème logiciel important qui est celui des mécanismes de substitution qui entrent en jeu lorsque le glyphe demandé par l'utilisateur n'est pas disponible. Le danger vient de la confusion entre caractères et glyphes : alors que l'on peut considérer que le remplacement d'un glyphe par un autre (par exemple, dans une autre fonte) est plus ou moins anodin, un texte codé en Unicode l'est normalement d'après les choix conscients de son auteur, et ne devrait pas être altéré par le logiciel, surtout si c'est pour des raisons aussi banales que la non disponibilité d'un glyphe. Prudence, donc, et méfiance vis-à-vis des décompositions de compatibilité, qui, malheureusement, sont considérées normatives.

2.4.5. *La casse*

Déjà par son nom, cette propriété normative a de quoi évoquer les liens étroits entre Unicode et la typographie. En fait il n'en est rien : on considère ici la différence entre lettres *majuscules* et *minuscules* d'un point de vue purement grammatical. C'est uniquement pour des raisons pratiques que le mot « casse » a été utilisé pour cette distinction grammaticale : en anglais, dans le contexte informatique et plus particulièrement celui du traitement de texte, les minuscules sont le plus souvent appelées *lowercase* (bas-de-casse) et les majuscules *uppercase* (mot à mot : « haut-de-casse », mais en français : capitale). Cette appellation traduit tout simplement le fait historique que deux casses étaient suffisantes pour composer des textes en alphabet latin – on est bien loin des cinq ou six casses nécessaires pour l'arabe.

Les choses se compliquent un peu lorsqu'on passe aux digraphes. En effet, ce qui distingue les digraphes des ligatures c'est que les deux lettres qui les composent peuvent être de casse différente. Cela n'arrive jamais dans les ligatures : la majuscule de « œ » est toujours « Œ »[24], et celle de « æ » est toujours « Æ ». Il n'en va pas de même pour les digraphes, comme « dz », « nj », etc. Ceux-ci on été introduit dans Unicode pour permettre une traduction automatique entre l'alphabet latin et l'alphabet cyrillique pour des langues balkaniques comme le serbe et le croate[25]. En alphabet cyrillique ces digraphes deviennent alors des véritables ligatures. Le problème est alors que quand on écrit en majuscules, « dz » devient, suivant le cas, « DZ » ou « Dz » (heureusement, la quatrième combinaison « dZ » n'arrive jamais). Au lieu de laisser le choix entre DZ et Dz au logiciel, Unicode a introduit une nouvelle « casse », la « casse de titrage » (*titlecase*), qui est donc Dz[26].

Heureusement ce cas est exceptionnel et se limite aux digraphes « balkaniques[27] ». Dans d'autres cas de digraphes illustres, l'attitude d'Unicode a été différente : le digraphe « l·l » a été brisé en un l pointé et un l ordinaire, le digraphe « oy » paléocyrillique est bel et bien représenté par un caractère Unicode, mais uniquement en « casse de titrage », et non pas avec deux lettres majuscules. Les digraphes Ourdou (consonne + *ha* d'aspiration) sont brisés et un caractère spécial (ARABIC LETTER HEH DOACHASHMEE est introduit pour ce *ha*, à ne pas confondre avec le *ha* ordinaire de l'arabe (ARABIC LETTER HEH).

Mais revenons à la casse. Unicode inclut dans le cédérom accompagnant le livre (et le site Web du consortium) une table de correspondance entre majuscules et minuscules. Encore une initiative utile pour les fabricants de logiciels, mais dangereuse, sinon néfaste pour les utilisateurs, puisque ces correspondances dépendent du contexte linguistique. Ainsi en turc la majuscule de la lettre « i » est « İ », plutôt que « I », qui, elle, est la majuscule de « ı ». De même, en grec, les majuscules gardent, ou ne gardent pas les accents et esprits suivant le contexte : en règle générale, lorsqu'un mot est écrit entièrement en majuscules, on ne met pas les accents et esprits, *mais* – et dans la réalité il y a toujours un *mais* – exceptionnellement, dans le but de désambiguïser, on les mettra tout de même. Un fabricant de logiciel qui applique donc bêtement les consignes

24. Encore que l'auteur a souvent vu dans des inscriptions ou enseignes la construction farandolesque « Oe »... (voir fi g. 16).

25. Attitude qui rappelle étrangement l'initiative de Tito de créer de toutes pièces une langue artifi cielle, appelée « serbo-croate », censée avoir deux écritures. Unicode n'est peut-être pas aussi innocent que cela après tout...

26. En français on aurait plutôt tendance à choisir DZ comme « casse de titre », puisqu'un titre est le plus souvent écrit entièrement en majuscules. Mais n'oublions pas que Unicode est anglocentrique, et que dans la typographie anglosaxonne – décidément, encore la typographie ! – et, à la connaissance de l'auteur, *uniquement* dans la typographie anglosaxonne, dans les titres on capitalise la première lettre de tous les mots qui ne sont pas auxiliaires. Ainsi « Unicode et typographie : un amour impossible » devient : « Unicode and Typography: a Forbidden Love ».

27. Des années de guerre sanglante entre Serbes et Croates nous empêchent le mot d'esprit de les appeler « serbo-croates ».

de capitalisation d'Unicode nuira à ses utilisateurs en leur imposant un choix unique. Heureusement ces tables de correspondance ne sont qu'informatives.

Une petite parenthèse concernant la « casse » : dans la communauté TEX on s'est aperçu depuis longtemps que dans un modèle informatique de la typographie deux casses ne suffisent pas : tout d'abord on distingue les majuscules/minuscules *obligatoires* et *facultatives*. Ainsi, lorsque j'écris PARIS, le P est une majuscule obligatoire lorsque je me réfère à la ville de Paris, et une majuscules facultative lorsque ce mot est le pluriel du mot « pari » ; les autres lettres sont des majuscules facultatives. Ainsi, un passage en minuscules produira « Paris » ou « paris[28] », suivant le cas. Ceci démontre d'autant plus que le passage d'une casse à l'autre dépend fortement du contexte. On parle également d'« inversion de casse », une invention allemande : lorsqu'un Allemand politiquement correct souhaite se référer aux étudiants de sexe indifféramment féminin ou masculin, il écrira « StudentInnen » ; le « I » majuscule à l'intérieur de ce mot devient minuscule lorsque le mot passe en majuscules : « STUDENTiNNEN ». Nous ignorons si ce jeu typographique trouvera sa place dans la tradition typographique allemande[29], mais il mérite certainement une nouvelle définition de casse, ne serait-ce que pour que les logiciels l'appliquent correctement.

Notons, pour finir, que le phénomène de casse ne concerne que les alphabets latin (y compris gothique et gaëlique), grec, cyrillique et arménien. Une bourde énorme des précédentes versions d'Unicode a été de considérer l'alphabet liturgique géorgien comme les lettres majuscules de cette langue qui, en réalité, n'en dispose pas. En géorgien on utilise des fontes spéciales pour les titres (la différence étant que les lettres géorgiennes ordinaires ont des ascendants et descendants, alors que dans les titres elles reposent toutes sur la ligne de base et ont toutes la même hauteur). Le choix d'Unicode de ne pas coder séparement ces lettres (par exemple en les qualifiant de « casse de titrage ») est, à notre avis, judicieux ; mais les avis des experts géorgiens sont plutôt partagés là-dessus, la question reste donc ouverte.

2.4.6. *La directionnalité*

Une des parties les plus techniques du standard Unicode est le chapitre sur le *comportement bidirectionnel*. En effet, Unicode a voulu résoudre complètement les problèmes logiciels d'affichage et d'interaction avec l'utilisateur dans le contexte d'écritures à directions différentes. Nous n'allons pas décrire le fameux « algorithme bidirectionnel » (appelé *bidi* par les initiés), mais dire simplement quelques mots sur les propriétés requises pour les caractères. Pour savoir comment traiter les caractères, Unicode les classe suivant leur direction. Comme les mêmes caractères sont souvent utilisés dans des contextes différents qui influent sur leur directionnalité, Unicode les divise en trois catégories : types *forts*, *faibles* et *neutres*.

---

28. Une majuscule facultative peut également produire une majuscule lors du passage en minuscules, si le mot est en début de phrase.

29. L'auteur a déjà vu des tracts écologistes français contenant le mot « étudiantEs », de toute évidence inspiré des tracts homologues allemands.

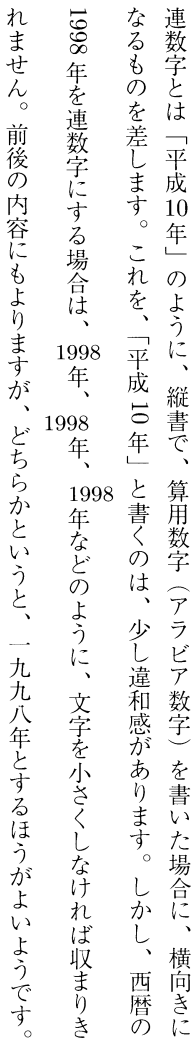


Figure 17. *Dans ce paragraphe de texte japonais vertical nous voyons quatre manières d'insérer une chaîne de caractères occidentaux (en l'occurrence, un nombre en chiffres arabes) dans du texte vertical : en l'écrivant incliné de haut en bas, ou horizontalement fer à gauche, centré ou fer à droite. Ces différents choix peuvent être en partie sémantiques ou culturels*

Les types forts sont ceux qui ont une direction unique invariable : l'alphabet latin est un type fort de direction gauche-à-droite, l'alphabet arabe un type fort de direction droite-à-gauche. Les chiffres arabes, par contre, sont un type faible puisque leur comportement est moins clair : en hébreu moderne, qui utilise des chiffres « arabes », comme dans les langues européennes, et les nombres sont écrits « de gauche à droite » alors que le texte est écrit de droite à gauche.

Évitons de tomber dans le piège de l'occidentocentrisme. En Europe, on a pris l'habitude d'écrire les nombres en mettant *d'abord* – et « d'abord » signifie sur le plan visuel : « de gauche à droite » – les unités plus grandes et ensuite les plus petites : 1 962 signifie « un millier plus neuf centaines plus six dizaines plus deux unités ». On

aurait pu prononcer et écrire ce nombre à l'envers : 269 1, seulement cette convention est profondément enracinée dans nos habitudes de notation occidentale[30]. En arabe et en hébreu moderne on a choisi d'écrire d'abord les petites unités et ensuite les grandes – et ici « d'abord » signifie, bien sûr « de droite à gauche ». Chaque écriture a le droit de choisir le système de notation des nombres et les deux choix décrits dans ce paragraphe sont aussi tout aussi valables l'un que l'autre.

L'appellation « type faible » traduit la nature profondément occidentale de l'informatique : Unicode se base sur la supposition que les nombres sont, de toute façon, notés de la plus grande unité à la plus petite, dans le texte. La différence de comportement des chiffres se réduit à une question d'affichage. Et, alors que les alphabets, eux, sont formels sur leur manière d'être affichés (gauche-à-droite ou droite-à-gauche), les nombres défient la direction du contexte, venant du système d'écriture ambiant, et sont toujours affichés de gauche-à-droite. Il ne viendrait pas à l'idée du consortium Unicode qu'en arabe et hébreu moderne les nombres sont peut-être tout simplement notés en commençant par les petites unités et qu'alors la direction d'affichage ambiant est parfaitement respectée[31]. Pour un Occidental il est plus facile d'inventer une « exception » (et donc une sorte de dysfonctionnement, une absurdité) dans les systèmes d'écriture des autres, que d'accepter un modèle différent de représentation des nombres...

À part les types forts et faibles, on a aussi les types neutres : espaces, retours-chariot, et autres caractères sans glyphe. La distinction, ici, apparaît à première vue philosophique, puisqu'il s'agit de caractères « invisibles », mais en fait est importante au niveau de l'affichage des *suites* de mots.

Pour bien expliquer la signification des trois types de directionnalité il faudrait décrire en détail l'algorithme bidirectionnel, et cela dépasse la portée de cet article.

---

30. Malgré le fait que dans certaines langues, comme l'allemand ou le russe on prononcera les dizaines après les unités : par exemple, on dira « zwei und sechsig » (mot à mot : « deux et soixante ») pour « soixante-deux ».

31. Citons également Henri Hudrisier (communication personnelle) qui écrit à ce sujet : « Les nombres sont mis en page et énoncés selon deux syntaxes qui correspondent aux deux directionalités :

Une appréhension globale de la valeur d'un nombre. On énonce d'abord et on écrit en premier dans le sens de l'écriture le chiffre correspondant au rang décimal le plus grand et on termine dans l'ordre par les unités puis les chiffres après la virgule. C'est par exemple la logique d'énoncé du français "quatre cent trente sept".

Une prise en compte "comptable" du nombre. On écrit d'abord en premier dans le sens inverse de l'écriture les chiffres après la virgule, puis les unités, puis dans l'ordre les chiffres correspondant aux rangs décimaux croissant en terminant par le plus grand. C'est l'énoncé "ein und zwanzig" ».

« On comprend bien cette logique puisqu'on doit justifi er les nombres à gauche sur leur virgule, par exemple quand on rédige une facture qui doit préserver l'alignement en colonnes de nombres comptables.

« Dès lors que l'on distingue l'énoncé, l'appréhension mentale et le sens d'écriture du nombre, la question devient plus claire. Notons que les deux syntaxes (donc les deux sens d'écriture) pouvaient coexister en sanskrit. »

Figure 18. *Une page mongole d'un ouvrage imprimé en Chine en 1982. Tournons cette page de 90 degrés dans le sens des aiguilles d'une montre : alors les notes de « bas » de page deviennent des « notes de haut de page » et le texte anglais « A tibetan christomathy ... texts and vocabularies by ... » doit être lu de bas en haut ; tournons cette page dans l'autre sens : les notes se trouvent maintenant en bas de page, mais le texte anglais est à l'envers...*

Notons simplement quelques faits plus généraux au sujet des directions d'écriture.

Premièrement, il semble étonnant qu'Unicode consacre autant d'énergie à la description des comportements de l'arabe, syriaque et hébreu vis-à-vis des autres systèmes d'écriture, et ne parle que très peu des écritures verticales et des interactions entre horizontal et vertical. Les langues à idéogrammes chinois sont traditionnellement écrites verticalement et lorsqu'on y inclut des textes dans d'autres systèmes d'écriture tantôt on les compose verticalement, tantôt horizontalement en les tournant de 90 degrés, tantôt on pratique un mélange des deux. Quelques mots français dans un paragraphe japonais s'écriront donc avec des lettres tournées de 90 degrés dans le sens des aiguilles d'une montre, et seront ainsi lus « de haut en bas ». Un problème intéressant se pose quand on mélange le japonais et l'arabe. On alors deux solutions : soit de tourner les lettres arabes de 90 degrés dans le sens inverse des aiguilles d'une montre, et cela a l'avantage de la lecture de l'arabe de haut en bas mais devient absurde lorsque dans le même paragraphe on a aussi, par exemple, du français, puisque l'arabe paraît renversé relativement au français ; soit de tourner les lettres arabes dans le sens des aiguilles d'une montre, ce qui signifie que l'on utilise, dans un contexte de haut en bas, l'algorithme bidirectionnel pour composer le japonais et l'arabe, et que, fatalement, l'arabe devra se lire de bas en haut.

La situation est encore pire quand on mélange, par exemple, du mongole et de l'anglais, puisque le mongole, contrairement au japonais, s'écrit verticalement mais en plaçant les lignes de gauche à droite. En incluant de l'anglais tourné de 90 degrés, comme on l'aurait fait pour le japonais, et en tournant la page de manière à ce que l'anglais devienne horizontal, on est obligé de lire les lignes anglaises de bas en haut (voir fig. 18)...

Unicode évite judicieusement ce genre de considérations, qui sont reprises par des normes de présentation de texte, telles que XSL-FO.

Mais il y a un problème qui se pose à Unicode, et qui n'a pas encore été résolu : le problème des écritures à directions multiples. Le berbère écrit en Tifinagh peut l'être indifféramment de gauche à droite ou de droite à gauche (en inversant, le cas échéant, les glyphes). Il en va de même des hiéroglyphes égyptiens, qui peuvent changer de direction d'écriture au plein milieu d'une inscription : comme les glyphes de ces hiéroglyphes représentent souvent des animaux ou des humains vus de profil, on suit le regard de ceux-ci pour savoir la direction d'écriture courante. Et dans certains cas, le grec ancien est écrit *boustrophédon* c'est-à-dire alternativement de droite à gauche et de gauche à droite (comme le bœuf (*bous*) qui tourne (*strophê*) pour labourer le pré). Comme ces trois écritures (le berbère tifinagh, les hiéroglyphes égyptiens et le grec épigraphique) ne font pas encore officiellement partie du codage Unicode, le problème de leur directionnalité ne s'est pas encore posé explicitement. Mais cela ne tardera pas puisque des projets d'insertion des tables correspondantes sont, depuis un moment, à l'étude.

3. Conclusion

Dans cet article nous avons voulu donner, en même temps, une approche « théorique » à Unicode et un aperçu des rapports entre Unicode et la typographie, rapports si fréquents et si ambigus, si étroits et en même temps si décriés, que l'on peut à juste titre parler d'une « histoire d'amour impossible ». D'autant plus que pour expliquer ou décrire leurs rapports on est souvent amené à se référer à la tradition typographique ou à l'histoire de l'informatique, et à des anglocentrismes ou occidentocentrismes officiellement niés par ce standard qui se veut « politiquement correct ». Il s'en suit que Unicode est un édifice tout aussi formidable que fragile, qui nécessite – et nécessitera pour les années à venir – une attention constante. Que l'on considère Unicode comme un outil néfaste de globalisation, ou comme la seule chance salutaire des écritures minoritaires à une présence internationale, on est tous d'accord qu'Unicode est un ingrédient culturel important du XXI^{e} siècle. De ce fait, tout effort d'amélioration, ne serait-ce qu'une critique constructive, est amplement justifié et devrait être encouragé. Nous espérons que cet article va œuvrer dans ce sens.

Remerciements

L'auteur souhaite remercier Jacques André, Patrick Andries et Henri Hudrisier pour leurs remarques et suggestions.