1

Concepts de base et terminologie

Objectif

Ce chapitre décrit brièvement Unicode et les raisons qui ont poussé à sa création. Ensuite, afin de rendre plus tangible le reste du livre, on aborde directement sous forme succincte les aspects les plus concrets d'Unicode : l'affichage de textes codés en Unicode et la saisie de caractères Unicode. Le chapitre se termine par une introduction à des termes fréquemment utilisés au cours de ce livre et dans le domaine de l'internationalisation des logiciels.

1.1 POURQUOI UNICODE ?

Les jeux de caractères utilisés avant l'avènement d'Unicode possédaient des architectures très différentes les unes des autres. Pour plusieurs, la simple détection des octets représentant un caractère était un processus contextuel complexe. Les jeux de caractères classiques ne pouvaient au mieux prendre en charge que quelques langues. La prise en charge de plusieurs langues à la fois était difficile, voire impossible. Aucun jeu de caractères ne fournissait toutes les lettres, les signes de ponctuation et les symboles techniques utilisés pour une seule langue comme le français.

Produire des logiciels destinés à différents marchés se résumait à une alternative simple : écrire des versions différentes du logiciel pour chaque jeu de caractères ou rendre le code du logiciel nettement plus compliqué pour traiter tous les jeux de caractères nécessaires. Ces deux options impliquent plus de codage, plus de tests, plus d'entretien et plus de soutien aux utilisateurs. Elles rendent la production de logiciels

destinés à l'exportation plus chère et retardent leur lancement ce qui n'est pas sans conséquence sur les revenus.

Le marché actuel rassemble souvent des données provenant de différentes sources, il suffit de penser à l'Internet. Le caractère codé correspondant à une lettre aussi simple que « A » peut varier en fonction du jeu de caractères et rendre le repérage, le tri et d'autres opérations fort difficiles. D'autres problèmes surgissent quand on mélange des données provenant de différents jeux de caractères (comme c'est le cas lors d'un copier-coller entre deux fichiers codés différemment). Il faut alors étiqueter ou baliser les données à l'aide d'informations sur le jeu de caractères d'origine de chaque morceau et créer de la sorte un nouveau codage à états. Perpétuer les anciennes méthodes de production de logiciel (ou de document informatique) dans ce contexte ne peut qu'entraîner des coûts prohibitifs, voire conduire au désastre.

Le standard Unicode[1,2] est un mécanisme universel de codage de caractères. Il définit une manière cohérente de coder des textes multilingues et facilite l'échange de données textuelles. Obligatoire pour la plupart des nouveaux protocoles de l'Internet[3], mis en œuvre dans tous les systèmes d'exploitation et langages informatiques modernes[4], Unicode est la base de tout logiciel qui veut fonctionner aux quatre coins du monde.

Grâce à Unicode, l'industrie informatique peut assurer la pérennité des données textuelles tout en évitant la prolifération de jeux de caractères et en augmentant l'interopérabilité des données. Enfin, Unicode simplifie le développement de logiciels et en réduit les coûts. En effet, Unicode permet de coder tous les caractères utilisés par toutes les langues écrites du monde (plus d'un million de caractères sont réservés à cet effet). Tous les caractères, quelle que soit la langue dans laquelle ils sont utilisés, sont accessibles directement sans aucun mécanisme d'échappement complexe contrairement à certains codages anciens comme c'était le cas dans l'ISO 2022 (§ 2.7, *ISO/CEI 2022*). Le codage de caractère Unicode traite les caractères alphabétiques, les caractères idéographiques et les symboles de manière équivalente, avec comme conséquence qu'ils peuvent se côtoyer dans n'importe quel ordre avec une égale facilité.

Le standard Unicode attribue à chacun de ses caractères un numéro[5] et un nom. À ce titre, il ne diffère guère des autres standards ou normes de codage

1. Généralement toute mention au « standard Unicode » s'applique également à la norme internationale ISO/CEI 10646. Dans cet ouvrage, on distingue les normes qui ont un caractère officiel (elles sont nationales ou internationales) des standards qui sont établis par des organismes privés comme le consortium Unicode ou l'IETF.

2. Une traduction française complète annotée et mise à jour du standard Unicode 3.2 est disponible à l'adresse <http://hapax.qc.ca>. Cette traduction a bénéficié du concours financier du gouvernement du Québec.

3. XML, HTML, WML, Corba 3.0, LDAP, etc.

4. Exemples : Java, ECMAScript (JavaScript), MS Windows 2000 et XP, Mac OS/X.

5. Une « valeur scalaire Unicode », dans le jargon d'Unicode.

de caractères. Cependant, Unicode fournit d'autres renseignements cruciaux afin de s'assurer que le texte codé sera lisible : la casse des caractères codés, leur directionalité et leurs propriétés alphabétiques. Unicode définit également des renseignements sémantiques et comprend des tableaux de correspondance de casse ou des conversions entre Unicode et les répertoires de divers autres jeux de caractères importants.

À l'heure actuelle, les données Unicode peuvent être codées sous trois formes principales : une forme codée sur 32 bits (UTF-32), une forme sur 16 bits (UTF-16) et une forme de 8 bits (UTF-8) conçue pour faciliter son utilisation sur les systèmes ASCII préexistants. Le standard Unicode est identique à la norme internationale ISO/CEI 10646 en ce qui concerne l'affectation des caractères (leur numéro) et leurs noms[1]. Toute application qui se conforme à Unicode se conforme donc à l'ISO/CEI 10646.

1.2 ABSENCE DE CODAGE UNIVERSEL

Un bon exemple de ce qui se produit en l'absence d'un codage universel ou lorsqu'on envoie une suite de caractères dans un code et qu'on le lit à l'aide d'un autre code est ce type de message que l'on reçoit encore parfois :

```
Bien reÃ§u ton message, on se voit Ã  NoÃ«l chez FranÃ§ois ?
```

Alors que le message aurait dû s'afficher comme suit :

```
Bien reçu ton message, on se voit à Noël chez François ?
```

Un autre exemple, plus exceptionnel, est illustré par la figure 1.1. L'enveloppe contenait un livre. Ce colis avait été envoyé à un étudiant russe par un ami français qui avait écrit à la main l'adresse qu'il avait reçue par courriel. Malheureusement, son logiciel était mal configuré et il a affiché le message écrit en cyrillique dans le code KOI8-R en le traitant comme s'il s'agissait d'un message écrit en Latin-1. Et ce sont ces caractères Latin-1 qu'il a transcrits comme adresse. Heureusement, l'adresse a été déchiffrée par les employés de la poste russe et livrée à bon port ! On voit en gris (*Россия*...) le déchiffrement du KOI8-R interprété comme du Latin-1 (en noir).

1. Le standard Unicode ne précise pas de noms français, cependant la norme internationale ISO/CEI 10646 a été publiée en anglais et en français. Les noms utilisés ici sont les noms officiels de la version française de l'ISO/CEI 10646.

Figure 1.1 – La poste victime de l'absence de codage universel

Les Russes et les Japonais — probablement parce qu'ils sont fréquemment confrontés à ce problème — ont chacun un nom pour décrire ce type de problème. En russe, on l'appelle *krakaziábry* крокозябры (« gribouillis »), alors qu'en japonais il se nomme *mojibaké* 文字化け (où *moji* signifie « caractère » et *baké* « changé »). Un autre exemple de mojibaké est le mot mojibaké lui-même 文字化け codé en UTF-8, mais décodé comme s'il s'agissait d'ISO Latin-1 : « æ–‡å—åŒ–ã ‘ ».

1.3 LANGUE ET ÉCRITURE

On confond souvent les concepts de langue, d'écriture et de codage. Il faut bien distinguer ces trois notions qui sont, globalement, orthogonales :

1) On parle de la langue d'un document : il est écrit en français, en arabe, en biélorusse, en malais, etc.
2) Un système d'écriture, ou plus simplement une écriture, est un ensemble de signes d'une même nature qui sert à transcrire une ou plusieurs langues. Une même langue peut s'écrire à l'aide de plusieurs écritures : si on écrit habituellement l'afrikaans à l'aide de l'écriture latine, les Malais d'Afrique du Sud l'ont parfois écrit en écriture arabe. Le malais lui-même s'écrit habituellement aujourd'hui en écriture latine, mais il existe un courant qui cherche à raviver sa transcription en écriture arabe (appelée jawi). Le mongol, pour sa part, s'écrit en écriture mongole, en phags-pa, et en cyrillique. Il existe même des langues qui nécessitent l'utilisation de plusieurs écritures simultanément, c'est le cas du japonais qui en utilise au moins trois : les syllabaires katakana et hiragana et les idéogrammes kanji hérités du chinois.

3) Enfin, le codage est un ensemble de caractères auxquels on attribue un numéro distinct. Un même codage peut permettre de ne coder qu'une écriture et les langues qui l'utilisent comme dans le cas du Windows Latin-1 et l'ISO/CEI Latin-1 (respectivement les codages les plus fréquents sur les plateformes Windows et Unix) ou un grand nombre d'écritures et d'autant plus de langues comme Unicode.

1.3.1 Notation, écriture, transcription et translittération

Il faut distinguer quatre mots qui ne sont pas synonymes : notation, écriture, transcription et translittération.

On utilise les notations pour les langues, la chimie et la musique. L'écriture a une extension moindre, plus proche de transcription. On l'utilise soit pour représenter des concepts (écritures idéographiques[1]) soit des paroles (alphabétiques ou syllabiques). On parle de transcription pour les écritures représentant la parole. La notion la plus limitée est la translittération, opération par laquelle on passe d'un alphabet utilisé pour l'écriture d'une langue à un autre alphabet.

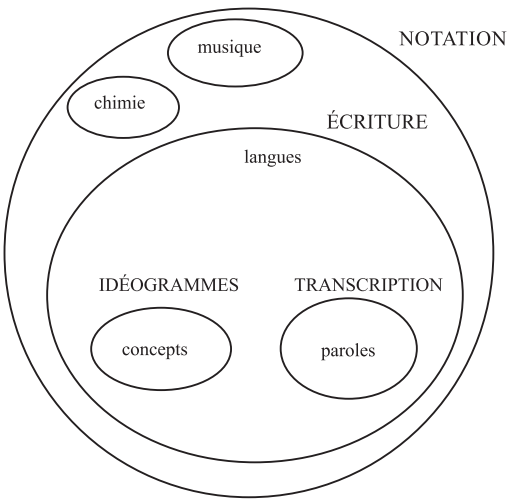

Figure 1.2 – Notation, écriture et transcription

Revenons ici sur les deux termes de transcription et de translittération. La transcription est le fait de représenter une langue en se préoccupant de faire correspondre les sons prononcés à des signes précis et non d'utiliser la graphie habituelle de cette langue pour représenter ces sons. On utilise souvent l'alphabet

1. À ce titre, les caractères chinois ne sont le plus souvent pas des idéogrammes, mais plutôt des idéophonogrammes, car ils comprennent une composante phonétique. Voir par exemple 妈 (« maman », prononcé mā,) dont l'idée est représentée par l'élément 女 (femme) et le son par l'élément 马 (« le cheval » prononcé mǎ).

phonétique international (voir § 5.1, *Latin étendu et API*) pour transcrire les langues. Le mot « chien » se transcrit en API [ʃjɛ̃].

La translittération, pour sa part, est l'opération par laquelle on passe d'un alphabet utilisé pour l'écriture d'une langue à un autre alphabet en transposant les mots lettre pour lettre indépendamment de la prononciation. C'est ainsi que le mot russe голова « tête » est translitteré en caractères latins sous la forme *golova* alors qu'il se prononce [gəlʌva]. Ce mot est souvent transcrit en français sous la forme de « galava ».

Le pinyin est le système officiel « international » de transcription du chinois depuis 1979, il fut créé à partir de l'alphabet phonétique *zhuyin zimu* de 1918 et des signes phonétiques (*zhuyin fahao*) de 1928. Il a été promulgué en 1958 par le conseil des affaires d'État chinois comme transcription officielle. Il utilise l'alphabet latin, avec un seul diacritique — le tréma —, il est fondé sur la prononciation de la langue commune *pǔtōnghuà* fréquemment appelée « mandarin » en français.

Il existe d'autres systèmes de transcription et de romanisation[1] du chinois, par exemple celui de l'École française d'Extrême-Orient. Il a été présenté pour la première fois dans un article paru en 1902 à Hanoï. L'ÉFEO est basé sur la prononciation en usage à Pékin. Bien que le pinyin soit aujourd'hui le système de transcription officielle, les transcriptions du chinois en ÉFEO sont nettement moins déroutantes pour un francophone que celles en pinyin. De nombreux toponymes et patronymes ont été popularisés dans la presse sous cette transcription ou une forme approximative — Mao Tsé (ou Tsö) Toung, Tsiang Kaï-chek, Yang-tseu-kiang, Pékin — plutôt que sous leur forme pinyin *Mao Zedong*, *Jiang Jeshi*, *Yangzijang* et *Beijing* (que le premier pékin venu pourrait prononcer Bégin ou Bédjingue alors que la prononciation en mandarin est [peitɕiŋ]).

1.3.2 Qu'est-ce qu'un caractère ?

Bien que nous utilisions tous les jours le mot caractère, ce mot tiré du grec χαρακτήρ « empreinte, marque », connaît plusieurs acceptions souvent insoupçonnées dans les domaines qui nous concernent. Selon le contexte, il peut en effet s'agir :

- Du plus petit élément d'une langue écrite pourvu d'une identité reconnue au niveau du sens ou de la forme abstraite plutôt qu'une forme particulière de cet élément. Pour les alphabets, ces éléments sont des lettres. En français, les lettres A à Z sont des caractères.
- De la forme particulière d'une lettre dans une police particulière ou un style donnée : les caractères italiques, les caractères Garamond.
- Du nom français des idéogrammes d'origine chinoise — ou mieux dit des idéophonogrammes, car ces caractères se composent généralement d'un élément phonétique et d'une clé sémantique indiquant la catégorie à laquelle appartient la notion exprimée : administration, liquide, nourriture, homme, etc.

1. Il s'agit d'une transcription en caractères latins et non en caractères romains...

- Dans le domaine informatique, des unités de base d'un codage qui peuvent correspondre à des lettres, des chiffres, des ligatures typographiques (fi), des symboles mathématiques (√) ou encore à des caractères de commande sans représentation physique distincte (pour indiquer le passage à la ligne, une espace insécable, un trait d'union optionnel, etc.).

Remarquons que la première définition mentionnée précédemment fournit des ensembles différents selon la langue considérée, même dans le cas de l'alphabet latin. Ainsi, traditionnellement en danois, considérait-on le « w » comme une variante du « v », variante qui ne faisait donc pas partie de l'alphabet. Aujourd'hui, ces deux lettres sont désormais considérées comme distinctes. C'est bien sûr la dernière acception du mot « caractère » — l'unité de codage — qui nous intéressera dans ce livre. Dans ce sens, les variantes de police d'un même caractère (exemples : 𝔞, 𝖆 et a) sont codées à l'aide d'un seul élément de code (ici « a »). Nous examinerons ce sujet en plus de détails dans le paragraphe suivant.

1.3.3 Graphème, caractère et glyphe

Le graphème est l'unité de base distinctive d'une écriture, il correspond typiquement à ce que l'utilisateur considère être un caractère. En français, par exemple, « b » et « d » dont deux graphèmes distincts, car ils permettent de distinguer les mots « bu » et « du ». À l'inverse, « a » et « *a* » ne sont pas des graphèmes distincts, car aucun mot français ne se distingue en fonction de ces deux signes (« avec » et « *avec* » sont les mêmes mots).

On dit que « a » et « *a* » sont deux glyphes du même caractère dit abstrait, abstrait car celui-ci n'a pas de forme concrète ou de style particulier.

L'idée que se fait l'utilisateur de cette unité de base peut cependant varier d'une langue à l'autre. C'est ainsi que le tchèque considère toujours le « ch » comme une seule lettre à part entière triée entre le « h » et l'« i ». Le breton considère, pour sa part, le « ch » et le « c'h » comme deux lettres à part entière.

Le tableau 1.1 illustre le rapport qui existe entre les caractères et les graphèmes. On remarque qu'un graphème ne correspond pas nécessairement à un caractère Unicode. Tout caractère Unicode n'est pas plus un graphème, puisque certains caractères Unicode ne sont pas visibles (les caractères de commande) ou ne s'utilisent jamais isolément dans une langue donnée (les diacritiques par exemple).

Tableau 1.1 – Rapport entre les caractères et graphèmes

Graphème	Caractères Unicode	Utilisation
a	a U+0061	Français, anglais, allemand
ą́	á ◌̨ U+00E1 U+0328	Lituanien
ç	ç U+00E7	Français, portugais
	c ◌̧ U+0063 U+0327	
ll	l l U+006C U+006C	Espagnol
ƛ̓	ƛ ◌̓ U+019B U+0313	Langues amérindiennes
ñg̃	n ◌̃ g U+006E U+0360 U+0067	Tagal (tagalog)
ㇷ゚	ㇷ ゚ U+30C8 U+309A	Aïnou en transcription kana

Dans le tableau 1.1, il faut lire une ligne comme :

ą́	á U+00E1	◌̨ U+0328	Lituanien

de la façon suivante : pour former le graphème lituanien « ą́ », on utilise deux caractères abstraits : le *a accent aigu* qui a comme numéro de caractère Unicode la valeur hexadécimale 0x00E1 suivi du *diacritique ogonek* qui a comme numéro de caractère Unicode la valeur hexadécimale 0x0328.

Les variantes de style d'un même caractère ou graphème sont appelées les glyphes de ce caractère. Le tableau 1.2 en fournit quelques exemples.

Tableau 1.2 – Caractères et glyphes

Glyphes (œils)	Caractères Unicode
A A A A A A A	U+0041 LETTRE MAJUSCULE LATINE A
ffi ffi ffi ffi f f i	U+0066 LETTRE MINUSCULE LATINE F + U+0066 LETTRE MINUSCULE LATINE F + U+0069 LETTRE MINUSCULE LATINE I
ف ف	U+0647 LETTRE ARABE FA'

1.3.4 Terminologie de l'apparence des caractères

L'œil ou le glyphe est la partie imprimée d'un caractère, c'est son apparence. Son pluriel en typographie est « œils ». Le fût ou la hampe de la lettre est le trait vertical que l'on trouve dans toutes les lettres dites carrées comme le « h », le « l » ou le « i ». Le contre-poinçon est la partie intérieure d'une lettre. L'empattement ou patin est le pied sur lequel se trouve à la base ou au sommet des traits verticaux (h est une lettre à empattements, h n'en a pas). Le corps ou force de corps est l'espace vertical occupé par un caractère, il est composé de l'œil et des talus de tête et de pied, les parties non imprimées verticalement de la lettre. L'approche est l'espace ou le blanc qui séparent deux lettres qui se suivent. Enfin, la chasse est l'encombrement en largeur d'un caractère donnée. La chasse change d'un caractère à l'autre, le « m » est plus large que le « n », on dit aussi que le « m » chasse plus que le « n » lequel chasse plus que le « i ».

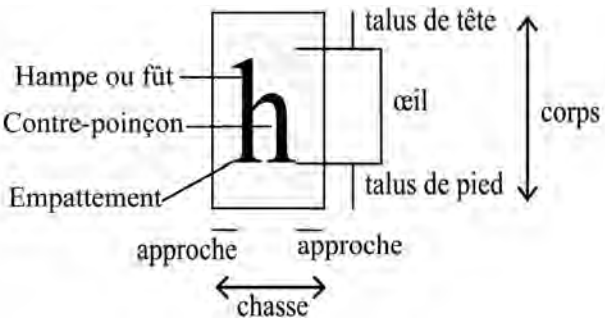

Figure 1.3 – Termes associés à une lettre à empattement

1.3.5 Classification des caractères

Complétons ce paragraphe par quelques notes terminologiques supplémentaires indispensables. On classifie habituellement les polices selon leur apparence. Il existe plusieurs classifications (Thibodeau, Vox ATypI, Gottschall), nous adopterons ici les cinq grands groupes que CSS utilise (tableau 1.3)

Tableau 1.3 – Classifications des polices

Catégorie	Font-family de CSS	Caractéristiques	Exemples de police
À empattements	`serif`	Avec petits traits qui débordent à droite et à gauche des jambages, très utilisés dans les livres comme celui-ci.	ITC Century Std, TRAJAN PRO, Times, Garamond
Sans empattements (bâton ou linéale)	`sans-serif`	Utilisé à l'écran et autres dispositifs de faible définition.	Helvetica, Gill Sans, Arial, Premi
À chasse fixe	`monospace`	Tous les caractères de largeur identique, ce style est fréquent pour les exemples en code machine.	Courier New, Prestige Elite
Cursive (scripte, manuaire)	`cursive`	Lettres écrites à la façon de l'écriture manuscrite	Mistral, French Script MT, English111 Vivace BT
Fantaisie	`fantasy`	Fourre-tout (aucune des quatre catégories précédentes)	CARNIVALE CRITTER, COTTONWOOD

Quelques mots encore sur les polices et leur terminologie : les typographes utilisent souvent le terme de « style de polices » (*typeface* en anglais) pour désigner un style général de glyphes et réservent le terme de « fonte » pour un dessin précis des œils associés aux caractères. Ainsi Garamond, une police, est-elle disponible dans les « fontes » suivantes : gras, italique, gras italique et romain. Au temps du plomb, ces fontes adoptaient des dessins légèrement différents en fonction de leur taille. Ce n'est plus le cas aujourd'hui dans les polices modernes vectorielles, sauf à des petits corps (taille) quand des instructions incorporées à ces polices permettent de corriger certains effets d'optique désagréables à petite définition (voir § 13.6.3, *Les polices de glyphes, Nuancement*).

> Nous invitons les lecteurs intéressés par cet aspect à se reporter au chapitre 11 du livre de Yannis Haralambous cité dans la bibliographie. La description des classifications de caractères typographiques latins y est très complète.

1.3.6 Caractères romains, latins, italiques et gothiques

Les caractères romains en typographie, et dans ce livre, sont les caractères droits. Penchés vers la droite, les caractères sont dits italiques. En règle générale, les textes en français sont rédigés en romain. Les caractères latins, quant à eux, sont ceux qui appartiennent au système d'écriture utilisé en anglais, en français et en italien, mais ni en russe ni en arabe.

Notons également qu'Unicode comprend des caractères de l'alphabet italique [U+10300..U+1032F], il ne s'agit pas ici du style généralement penché, mais des lettres utilisées par les langues de l'Italie antique (l'étrusque, l'osque, l'ombrien, le

falisque...). De même, l'ISO 10646 oppose distinctement le gothique (également appelée **fraktur** et **écriture brisée**) au gotique [U+10330..U+1034F] qui, selon le Petit Robert, est la langue des Goths (avec un « h » cette fois... eh, oui !), rameau oriental des langues germaniques.

1.3.7 Écriture CJC

Ce qui frappe tout d'abord l'Occidental dans l'écriture chinoise c'est le grand nombre de signes. Aujourd'hui, Unicode contient plus de 70 000 caractères chinois et plusieurs milliers d'autres caractères chinois sont en cours de normalisation ! Le caractère chinois (ou han) s'inscrit dans un carré imaginaire de surface identique pour chaque caractère. C'est pourquoi les Chinois désignent ces caractères du nom de *fàngkuàizì*, c'est-à-dire tétragramme. Certains caractères n'ont qu'un ou deux traits, d'autres en ont jusqu'à soixante-quatre. On donnera à chaque trait consécutif la taille adéquate pour qu'à la fin tous les traits tiennent dans le carré normatif.

Conçue pour écrire les divers dialectes chinois, l'écriture idéographique s'est étendue aux pays de la zone d'influence chinoise et servit à écrire pendant des siècles d'autres langues que le chinois, comme le japonais, le coréen et le vietnamien. Des locuteurs d'autres langues, comme les Yi, créèrent leur propre système idéographique en imitant le chinois.

Le chinois, langue en grande partie monosyllabique et sans flexion, convient parfaitement aux systèmes d'écriture idéographique. Les idéogrammes se prêtent moins bien à d'autres types de langues. Les Japonais résolurent ce problème en créant deux écritures syllabiques, le *hiragana* et le *katakana*. Les Coréens inventèrent un système alphabétique dans lequel les lettres sont groupées en blocs syllabiques ressemblant à des idéogrammes appelés *hangûl*.

Jusqu'au XX^e^ siècle, le vietnamien ne s'écrivait qu'à l'aide d'idéogrammes. Cette écriture, parfois qualifiée d'annamite, fut ensuite remplacée par un alphabet dérivé de l'écriture latine qu'Alexandre de Rhodes, un jésuite français, avait perfectionné et codifié dès le XVII^e^ siècle.

Le japonais emploie encore aujourd'hui abondamment des idéogrammes appelés *kanji* ; ils sont plus rares en coréen, où les idéogrammes se nomment *hanja*.

En Chine continentale, le gouvernement tente de promouvoir l'utilisation d'idéogrammes modernes et simplifiés plutôt que ceux, plus anciens et plus traditionnels, utilisés à Taïwan ou dans les communautés chinoises d'outre-mer. La simplification peut parfois être assez spectaculaire : 台 (« terrasse, belvédère, estrade ») est la version simplifiée de 臺 (de même sens, car la simplification n'est que graphique).

Traditionnellement, les écritures de l'Extrême-Orient s'écrivent du haut de la page vers le bas en colonnes alignées à partir de la droite jusqu'à la gauche de la page. Sous l'influence occidentale, on écrit également de nos jours ces écritures à l'horizontale de gauche à droite. Les glyphes peuvent fortement varier selon les différents pays et les applications. La police la plus utilisée dans un pays peut ne pas l'être dans un autre.

Les caractères han sont à chasse fixe. Chaque caractère occupe le même espace vertical et horizontal, peu importe la complexité de la forme. Cette pratique, résultat d'une longue tradition typographique chinoise, inscrit chaque caractère dans une cellule carrée. Dans ce sens, le rendu des idéogrammes est relativement simple comparé à d'autres écritures qui connaissent le réordonnancement (la dévanâgarî), les formes contextuelles (arabes), les diacritiques superposés et les ligatures (latin).

1.4 UNICODE, EN QUELQUES MOTS

1.4.1 Ce qu'Unicode est...

Unicode est une méthode pour représenter les écritures du monde à l'aide d'un répertoire de caractères dit universel (près de 100 000 caractères à l'heure actuelle). Unicode associe à chaque caractère un numéro unique, quelle que soit la plateforme, quel que soit le logiciel et quelles que soient les langues qui utilisent ce caractère. Unicode précise également des propriétés pour ces caractères afin d'en préciser la nature et le traitement. Enfin, Unicode comprend également une série de rapports techniques (certains normatifs, d'autres purement indicatifs) qui précisent des manières d'exploiter les caractères définis : comment les trier, les comparer, couper des lignes qui comprennent ces caractères.

Unicode normalise les mêmes caractères que l'ISO/CEI 10646. Les noms de caractères sont identiques en anglais. L'ISO/CEI est également publiée officiellement en français et nous utiliserons ces noms français normalisés par l'ISO dans ce livre.

1.4.2 Ce qu'Unicode n'est pas...

- **Un produit** — On pense souvent qu'Unicode est un produit. Rien n'est plus faux. Unicode (ou l'ISO/CEI 10646) n'est ni un logiciel ni une police, mais un code (codage) de caractères accompagnés de propriétés et de textes explicatifs qui décrivent la manière de traiter des données codées en Unicode.
- **Complexe** — La mise en œuvre d'Unicode seul n'est pas complexe, tout dépend de la langue que vous voulez prendre en charge. En effet, la conformité à Unicode peut s'énoncer sommairement et de manière officieuse comme suit : on peut ignorer un caractère qu'on ne comprend pas, mais ni le corrompre ni le supprimer; la mise en œuvre d'un sous-ensemble ne dépend que de vous. Ainsi, si vous ne voulez prendre en charge que les écritures de l'Europe occidentale, vous (ou votre système) n'aurez guère plus à mettre en œuvre que ce que vous avez déjà en ASCII. Pour plus de détails voir § 4.5, *Conformité*.
- **Codé sur 16 bits** — Il s'agit là uniquement d'une forme commune de sa représentation. Nous y reviendrons plus tard (§ 4.1.4, *Forme en mémoire des caractères*).
- **Une solution miracle à la production de logiciels multilingues** — Il est possible de produire de tels logiciels sans Unicode, c'est sans doute plus

compliqué alors, mais Unicode n'est pas suffisant pour produire un logiciel multilingue de façon efficace. Voir § 1.7 et le chapitre 12.

- **Un format ou un langage de mise en pages** — Unicode n'est pas un langage de description de pages capable de remplacer ou de servir de format à un traitement de texte, à un éditeur d'équations ou à un logiciel d'écriture de partitions musicales. Unicode ne code que du texte brut. Cette notion est extrêmement importante et nous l'introduirons, brièvement dans le paragraphe suivant. Nous reviendrons sur cet aspect au cours du livre.

Texte brut et texte enrichi

Unicode est conçu pour représenter du texte brut. En première approximation, on peut décrire du texte brut comme du texte sans changement de polices, sans description de colonnes, de marge, d'entête et dépourvu de toute indication de structure autre que la séparation en paragraphes. On oppose le texte brut au texte dit enrichi qui pourra comprendre des changements de police, de corps, de graisse, des notes de bas de page, etc.

Une autre définition consiste à dire qu'Unicode (et donc le texte brut) ne transmet que ce qui est strictement nécessaire à la compréhension du texte, ni plus, ni moins. Il comprendra donc des lettres, des chiffres et de la ponctuation. Notons que si l'on veut finasser, on peut dire que l'italique ou le gras peuvent modifier le sens d'un texte. Quoi qu'il en soit, Unicode ne code pas ces subtilités, car il est tout à fait possible d'écrire tous les textes français de façon parfaitement compréhensible sans italique ni gras.

Unicode ne code donc pas toutes les informations sur la mise en pages des textes qu'il permet de représenter. Unicode dit déléguer ces tâches (le choix de polices, la justification, la sélection de variantes stylistiques...) à des protocoles de haut niveau. Ces protocoles peuvent être des langages comme HTML, CSS, XSL-FO ou des formats de description de traitement de texte comme RTF ou le format propriétaire d'InDesign d'Adobe.

Texte inclus dans un protocole de haut niveau (RTF) :

```
{ \ltrch\fcs0 \f0\fs24 Petit document, je change de police. La 1}{\fs24\
super\'e8re }{\fs24 est }{\i\f35\fs24 Arial italique }{\fs24\f0, la 2}{\f0\
fs24\super e}{\f0\fs24  est }{\b\fs24\f64 Garamond gras}{\f0\fs24 .}{\par }}
```

Même texte dans autre protocole de haut niveau (HTML) :

```
<body >
<p style='font-size:12.0pt;font-family:"Times New Roman"'>Petit document, je
change de police.  La 1<sup>ère</sup> est <span style='font-family:Arial; font-
style:italic;'>Arial italique </span>, la 2<sup>e</sup> est <span style='font-
weight:bold;font-family:"Garamond Premr Pro"'>Garamond gras</span>.</p>
</body>
```

Pourquoi cette division, cette répartition des tâches entre Unicode et « les protocoles de haut niveau » ? Il y a, bien sûr, une volonté nette de simplification du travail de normalisation. Mais on peut aussi y voir une volonté de longévité : on peut

tout à fait envisager de sauvegarder un document en texte brut et de pouvoir le relire dans un siècle, pour autant qu'aucune corruption physique du texte ne l'ait altéré. On ne peut pas en dire autant des données conservées en texte enrichi au regard de la longue histoire des formats ou des langages de mise en pages, fréquemment modifiés ou abandonnés et souvent incompatibles dans les détails de leur mise en œuvre.

1.5 APPRIVOISER LES POLICES UNICODE

Unicode est à la fois un sujet très simple et très complexe. Saisir et afficher des caractères Unicode est souvent devenu relativement simple aujourd'hui. D'autres processus, comme le tri, la normalisation de chaînes Unicode ou la coupure de lignes sont bien plus délicats. Nous allons rapidement aborder ici les sujets les plus simples, afin d'apprivoiser le sujet et de mieux le saisir instinctivement. Les sujets plus complexes comme le tri ou le fonctionnement interne de processus comme l'affichage sont relégués aux derniers chapitres du livre.

1.5.1 Afficher des caractères Unicode

L'affichage de caractères Unicode se fait à l'aide de polices souvent appelées *polices Unicode*. Il faut bien comprendre ce qu'on entend par là et, pour ce faire, il nous faut introduire quelques notions ici avant d'y revenir en plus de détails dans le chapitre 13 consacré aux polices.

Une police est un ensemble de dessins associés à des numéros de caractère. En termes plus techniques, on dira que les polices sont un ensemble de glyphes ou d'œils indexés par des valeurs de caractère. Voir la figure 13.3, *Deux espaces*, dans le § 13.6.2, *Fonctionnement d'un moteur de rendu*.

Dire qu'une police est une police Unicode signifie que les numéros de caractères qui servent à repérer ces œils sont exprimés sous la forme de numéros de caractère Unicode, ceci ne signifie pas que la police en question contient un œil (un glyphe) pour chaque caractère Unicode. Avant de continuer, la bonne nouvelle : la quasi-totalité des polices aujourd'hui sont des polices Unicode, elles imprimeront correctement les caractères Unicode pour lesquels elles ont des glyphes.

Une police peut contenir un même glyphe pour plusieurs caractères. C'est le cas le plus souvent pour les polices qui prennent en charge plusieurs écritures et qui comprennent un seul œil pour le A latin, le A cyrillique et l'alpha majuscule grec (A).

Chaque police a un nom. Il s'agit souvent d'une marque de commerce. C'est pourquoi on trouve des polices quasi identiques, créées par des fondeurs différents, qui portent des noms très différents. C'est le cas des polices Helvetica (conçues dans les années 1950 en Suisse par la fonderie Haas) et Arial (la version postérieure de Microsoft) qui sont très proches l'une de l'autre.

1.5.2 S'assurer que sa police est une police Unicode

Vous avez en votre possession une police, vous voudriez savoir si elle peut être utilisée avec des textes Unicode. Comment vérifier la chose ? Il existe plusieurs moyens. Si vous possédez une machine Windows, vous pouvez télécharger un logiciel que Microsoft offre gratuitement à l'adresse suivante : <http://www.microsoft.com/typography/TrueTypeProperty21.mspx>.

Après son installation, si vous cliquez sur une police du répertoire *Polices* (par exemple via `Démarrer > Panneau de configuration > Polices`) et cliquez avec le bouton droit puis choisissez `Propriétés`, la boîte de dialogue illustrée par la figure 1.4 apparaîtra.

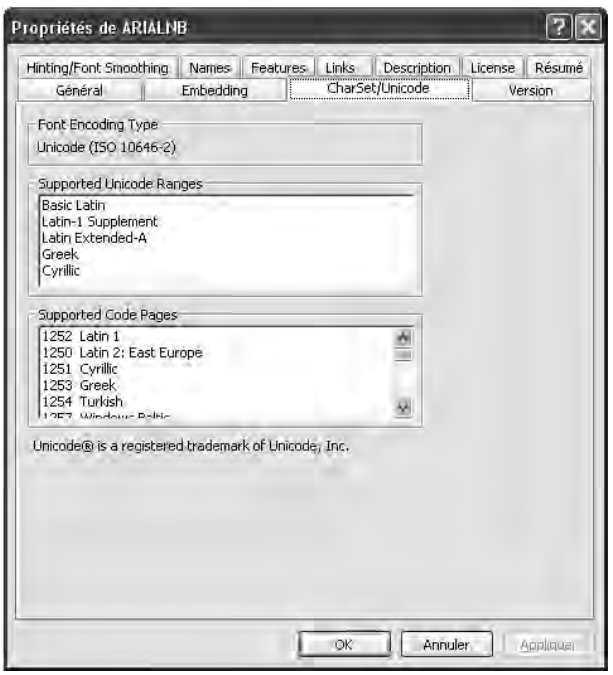

Figure 1.4 – Fenêtre de propriétés de police sous MS Windows

Malheureusement, Microsoft a décidé de ne pas traduire celle-ci complètement ! Si vous cliquez sur l'onglet « CharSet/Unicode » vous remarquerez que la police (de la famille Arial) est indexée par des valeurs Unicode (« Font Encoding Type = Unicode (ISO 10646-2 ») et qu'elle comprend non seulement des caractères latins, mais également grecs et cyrilliques.

Il existe un autre outil gratuit, BabelMap, qui permet également ce type de découverte. Il est téléchargeable ici <http://hapax.qc.ca/BabelMap_fr.html>. Une fois téléchargé et installé, lancez-le (nous l'utiliserons encore !). Cliquez sur `Outils>Analyse de police`. La fenêtre de la figure 1.5 apparaît.

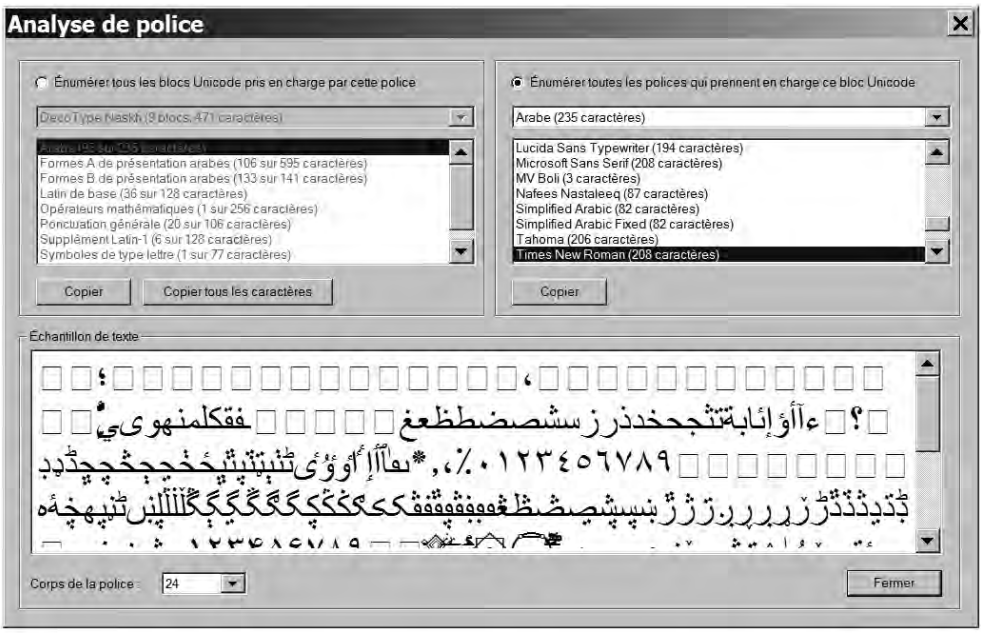

Figure 1.5 – Analyse de police sur BabelMap

Vous pourrez alors analyser la police courante et voir tous les blocs Unicode qu'elle prend en charge. Si les caractères affichés pour un bloc ne correspondent pas aux noms de caractères : vous voyez des caractères arabes à la place de caractères latins, par exemple, votre police n'est pas Unicode (c'est-à-dire qu'elle n'est pas indexée à l'aide de valeurs Unicode, voir § 13.5, *Pas d'expédients ASCII, de l'Unicode !*). Cet outil permet également, comme c'est le cas dans la capture d'écran, de vérifier la couverture d'un bloc à travers toutes les polices installées sur votre système. Ceci n'est pas possible avec l'outil de Microsoft. Dans l'exemple précédent, on voit le résultat de l'analyse du bloc arabe. On remarque que la police Times New Roman comprend 208 glyphes (pour les 235 caractères du bloc arabe), mais que la police MV Boli n'en reprend que trois. Il s'agit d'une police pour le thâna, une écriture en partie d'inspiration arabe utilisée aux Maldives pour écrire le divéhi ou maldivien, une langue indo-iranienne.

1.5.3 Où trouver des polices multi-écritures supplémentaires ?

Il est possible que votre système n'ait pas de police assez riche pour afficher les caractères dont vous avez besoin. La première chose à faire, si votre système d'exploitation est MS Windows, est de vous assurer que Windows a été configuré pour prendre en charge un maximum de langues. Si ce n'est pas le cas, il vous faudra ajouter des polices supplémentaires.

Sur Windows XP, procéder de la façon suivante :

- Sélectionner `Démarrer > Panneau de configuration > Options régionales et linguistiques`.
- Cliquer sur l'onglet `Langues`, on y voit deux cases : « écritures complexes[1] et écrites de droite à gauche » et « langues d'Extrême-Orient » (chinois japonais coréen, CJC). Cocher sur la case correspondant aux écritures qui vous intéressent. Le système installera les polices correspondantes ainsi que les modules algorithmiques nécessaires à une bonne prise en charge de ces langues. Appuyer sur OK.
- Il se peut que le système vous demande ensuite d'insérer le cédérom de Windows ou de préciser l'endroit où ces fichiers d'installation se trouvent. Suivez les instructions.

Windows Vista a intégré dans sa distribution standard ces écritures complexes et ces langues d'Extrême-Orient (CJC) en y ajoutant de nombreuses nouvelles écritures et au moins une police pour chacune. Parmi ces nouvelles écritures prises en charge par Windows, on retrouve : l'éthiopien, le khmer, le lao, le mongol, l'oriya, le ouïgour (à l'aide de l'écriture arabe), le singhalais, le tibétain, le syllabaire autochtone canadien et le yi.

MS Office installe également de nombreuses polices très riches, dont Arial Unicode MS qui couvre de très nombreuses écritures.

Il existe d'autres polices gratuites ou offertes comme « partagiciel » qui prennent en charge un très grand nombre d'écritures :

- **Code 2000** — <http://code2000.net/index.htm> — Cette police comprend des glyphes pour la quasi-totalité d'Unicode 4.0 sauf ce qui est dans la police Code 2001, techniquement on dit qu'elle comprend des glyphes pour tout le PMB (Plan multilingue de base) d'Unicode 4.0, nous reviendrons sur le terme PMB par la suite, il suffit de savoir à ce stade que presque tous les caractères Unicode sont codés dans ce plan. La qualité de la police est relativement médiocre, mais il s'agit d'une bonne police de derniers recours. Il est recommandé de la télécharger (et si possible de laisser une obole à son créateur bénévole).
- **Code 2001** — <http://code2000.net/code2001.htm> — Cette police comprend surtout des glyphes pour des écritures anciennes.
- **Cardo** — <http://scholarsfonts.net/cardofnt.html> — Police destinée aux spécialistes des lettres classiques (grec, italique, latin, alphabet phonétique international, hébreu).
-

1. À savoir l'arabe, l'arménien, le géorgien, l'hébreu, les écritures de l'Inde, le thaï et le vietnamien.

- **Charis SIL** — <http://scripts.sil.org/CharisSILfont> — Prend en charge quasiment toutes les langues utilisant les écritures latine, cyrillique, et l'alphabet phonétique international.
- **DéjàVu** — <http://dejavu.sourceforge.net/> — Prend en charge le latin, le grec, le cyrillique, les écritures africaines romanisées, l'arabe, l'hébreu, le n'ko, le tifinagh, etc.
- **Doulos SIL** — <http://scripts.sil.org/DoulosSILfont> — Prend en charge le latin, le cyrillique, l'alphabet phonétique international, les écritures africaines romanisées.
- **Gentium SIL** — <http://scripts.sil.org/Gentium> — Prend en charge le latin, l'alphabet phonétique international, le grec, les écritures africaines romanisées.

1.5.4 Absence de glyphe pour un caractère

Il arrive parfois qu'un programme ne puisse pas afficher des caractères Unicode stockés dans un document. Ce peut être dû au fait que le texte est mal décodé (mojibaké), qu'il n'existe pas de caractère Unicode pour ce numéro de caractère ou que les polices associées à ce document n'ont pas de glyphe pour un caractère Unicode légitime.

Rappelons au passage que CSS et XSL-FO permettent d'associer une liste de polices à un document ou à des passages de celui-ci. Le logiciel utilisera les polices dans l'ordre dans lequel elles sont mentionnées. Soit donc un passage HTML libellé de la sorte[1] :

```
<p style="font-family:Batang,SBL Greek,serif">L'esclave en grec : ὁ δοῦλος.</p>
```

Le navigateur affiche le texte en français en utilisant la police Batang, il affiche également les caractères non accentués grecs en Batang, car ils se trouvent dans cette police. Les caractères accentués en sont, toutefois, absents, le navigateur se tourne alors vers la seconde police SBL Greek qui elle les contient. Notez que ce changement de police ne produit pas un résultat très esthétique, il s'agit d'un pis-aller. Pour éviter ce problème, il suffit d'inverser la liste des polices :

```
<p style="font-family:SBL Greek,Batang,serif">L'esclave en grec : ὁ δοῦλος.</p>
```

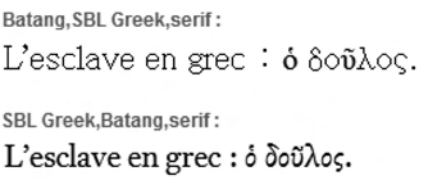

Figure 1.6 – Choisir une police complète et harmonieuse

1. Nous avons, par désir de concision, orné le <p> d'un style particulier. Une feuille de style CSS isolera habituellement ce style et l'appliquera à tous les paragraphes du document de la sorte :

```
p { font-family : "SBL Greek", Batang, serif; }
```

La police SBL Greek comprend tous les caractères grecs accentués nécessaires, les mots « ὁ δοῦλος » s'affichent donc sans changement inopportun de police. Elle comprend d'ailleurs tous les autres caractères latins, la phrase entière s'affiche dans une police conçue pour afficher du grec accentué et du latin de façon harmonieuse. Si, dans notre exemple, un caractère venait encore à manquer (il n'est ni dans SBL Greek, ni dans Batang), le navigateur se replierait alors sur n'importe quelle police à empattements puisque nous avons mentionné « serif » comme dernière police. Si aucune police à empattements ne permet d'afficher un caractère, le navigateur se tournera vers sa police de dernier recours. Si celle-ci ne contient aucun glyphe pour un caractère du texte, le navigateur affichera habituellement un glyphe pour indiquer l'impossibilité d'en trouver un : un « ? », un « □ » ou encore un « 🞏 ».

Il n'y a pas qu'en HTML que des caractères peuvent manquer. C'est également le cas pour d'autres documents électroniques comme les PDF ou les fichiers de traitement de texte, si des polices mentionnées dans le document en question ne sont pas disponibles. MS Word affiche alors typiquement un carré blanc : « □ ».

1.5.5 Incorporation des polices

Pour pallier ce problème, une technique consiste à incorporer les polices nécessaires à l'affichage d'un document dans ce même document. Ce n'est pas le comportement habituel des traitements de texte, car cette incorporation de polices peut considérablement alourdir le document. Habituellement, les traitements de texte supposent que les polices sont installées sur le système local et ils n'incluent que des références à ces polices dans le document, sans inclure les polices elles-mêmes. Il existe, toutefois, au moins un cas où l'on peut vouloir inclure la police dans le document de traitement de texte, il s'agit des cas où, pour des raisons de droit d'auteur, l'on ne désire pas transmettre (et donc donner) une police, mais l'on veut cependant que le document s'affiche correctement chez son correspondant.

Si l'incorporation automatique des polices n'est pas souvent souhaitable pour les fichiers de traitement de texte, elle est cruciale pour les documents PDF qui se targuent de pouvoir s'afficher sur toutes les plateformes de manière identique. Cette incorporation n'est cependant pas obligatoire[1]. Les fichiers PDF n'incorporent habituellement pas ce qu'il est convenu d'appeler les 14 fontes PostScript standard : *Times Roman* (italique, gras, gras italique, romain), *Courier* (italique, gras, gras italique, romain), *Helvetica* (italique, gras, gras italique, romain), *Symbol* et *Zapf Dingbats*. Tous les logiciels d'affichage de fichier PDF doivent incorporer ces polices et il est donc inutile que tous les fichiers les incorporent.

1. Ainsi, pour les applications industrielles qui envoient de très nombreux fichiers qui utilisent quasi systématiquement les mêmes polices, il est possible d'installer ces polices sur le serveur d'impression qui transforme les fichiers PDF en images tramées pour les imprimantes qui lui sont reliées et éviter de la sorte que chaque fichier PDF envoyé à ce serveur contienne à chaque fois toutes ces polices.

Notons enfin que l'incorporation de police est souvent optimisée : seuls les glyphes d'une police donnée qui sont réellement utilisés dans le document sont incorporés, on parle alors de jeu partiel incorporé. Cela allège, bien sûr, d'autant le document qui incorpore les portions de police en question. Ceci est essentiel avec les polices idéographiques, dites CJC, car ces polices contiennent un très grand nombre de glyphes.

Toutes les polices ne sont pas incorporables dans des documents. Si elles le sont, vous pourrez alors, par exemple, les inclure dans des documents **MS Word 2003** en le mentionnant parmi les options de sauvegarde (`Fichier > Enregistrer sous... > Outils > Options d'enregistrement`, puis cocher `Incorporer les polices TrueType`). Avec **MS Word 2007**, il faut cliquer sur l'icône Microsoft Office, en haut à gauche, puis cliquer sur `Options Word > Enregistrement` puis cocher `Incorporer les polices dans le fichier`.[1] Il est sage de décocher cette option une fois le document enregistré, car cette option s'appliquera à toutes les futures sauvegardes quel que soit le fichier, ce qui est en général totalement superflu. Pour déterminer si une police est incorporable, utiliser l'utilitaire de propriétés de police que nous avions déjà vu précédemment (figure 1.3).

Malheureusement, l'incorporation des polices n'a jamais vraiment décollé dans le domaine de l'HTML peut-être parce qu'aucune norme n'est apparue suffisamment tôt dans ce domaine. Microsoft a développé sa propre solution peu répandue : WEFT qui n'est pas prise en charge par Firefox. D'autres techniques d'incorporation existent, plus particulièrement sIFR[2] qui combine JavaScript, CSS et Flash ou encore Glyphgate (<http://www.glyphgate.com/info/>), un module à ajouter au serveur qui gère les pages HTML et qui, au pire, envoie des images du texte à afficher pour les polices qui ne seraient pas prises en charge correctement par votre navigateur[3]. Pour plus de détails sur ce sujet, voir le chapitre 10, *Fontes et pages web*, de l'ouvrage de Yannis Haralambous cité dans la bibliographie.

1.6 SAISIR DES CARACTÈRES UNICODE

Il existe de nombreuses manières de saisir des caractères Unicode. Celles-ci varient selon le programme, le document ou la plateforme. Plusieurs peuvent exister en parallèle. Celles-ci s'opposent souvent par deux qualités : la simplicité ou l'universalité. La méthode la plus simple est souvent le clavier avec une touche gravée (ou une petite combinaison de touches) affectée à un caractère. Étant donné le nombre limité de touches gravées, c'est évidemment de loin la méthode la moins universelle. Une

1. Dans MS Word 2003 comme dans MS Word 2007, il existe une case à cocher qui permet de « ne pas incorporer les polices système communes », il s'agit ici des polices livrées avec MS Windows et non des quatorze fontes PostScript standard.

2. Voir <http://www.journaldunet.com/developpeur/tutoriel/dht/060713-typographie-sifr.shtml>.

3. Voir le site de l'Assemblée législative du Nunavut pour un exemple de site géré par GlyphGate et capable d'afficher de l'inuktitut (esquimau) même sur un navigateur dépourvu de police appropriée : <http://www.assembly.nu.ca/inuktitut/index.html>.

méthode universelle — mais peu mnémotechnique — est de mentionner le numéro des caractères Unicode que l'on veut saisir.

On peut, grosso modo, classer les différentes manières de saisir des caractères en six grandes catégories :

1) **Clavier** — Il s'agit ici d'une combinaison de touches. Elle peut comprendre l'utilisation de touches mortes comme « Alt », « Alt Gr » ou « Ctrl ».
2) **Méthode d'entrée** — Les méthodes d'entrée sont utilisées dans les langues idéographiques. L'utilisateur précise souvent en orthographe phonétique le mot chinois (par exemple en pinyin) ou par une suite de portions du caractère (dans le cas de la méthode des « quatre coins »). Comme plusieurs caractères chinois partagent typiquement une même prononciation, l'utilisateur doit le plus souvent choisir parmi les choix proposés par la méthode pour conclure la saisie.
3) **Menu** — L'utilisateur parcourt une série de menus et choisit une option qui insère un caractère particulier. Certains éditeurs HTML (comme le Composeur de Mozilla) utilisent cette méthode qui ne permet, toutefois, que d'insérer un nombre très limité de caractères différents.
4) **Appel de caractère** — Série de caractères facilement accessibles à tous les utilisateurs (souvent réduite aux seuls caractères ASCII de base) qui sera interprétée par la suite comme un caractère d'une autre valeur. C'est ainsi que la chaîne de caractères « È » sera interprétée comme un « È » par les analyseurs HTML et XML, 00C8 étant la valeur hexadécimale Unicode du caractère « È ».
5) **Sélection à partir d'un tableau** — On appelle un module d'un programme (par le moyen d'un menu) qui affiche ensuite une grille de caractères. On peut ensuite pointer sur une case du tableau pour choisir le caractère désiré parmi les caractères voisins. Cette option permet d'insérer plus de caractères que le simple menu (l'option 3), qui ne permet que de choisir un caractère parmi une courte liste affichée dans un menu déroulant par exemple.
6) **Clavier virtuel** — Un petit clavier s'affiche à l'écran et on peut sélectionner les touches grâce au clavier ou en cliquant sur les touches affichées à l'écran. C'est un peu une forme hybride du clavier standard et du tableau à l'écran.

Le tableau 1.4 résume différentes options d'insertion de caractères Unicode.

Tableau 1.4 – Méthodes pour insérer des caractères Unicode

Catégorie	Contexte	Méthode	Remarques
Appel	CSS	`\41B`	Notation hexadécimale, se termine après 4 chiffres (0-F) ou au premier caractère qui n'est pas un chiffre hexadécimal. L'espace qui suit cet appel de caractère est « avalé ». \41B est la lettre cyrillique Л. Voir aussi, § 11.6, *Notation des caractères*.
Appel	HTML	`&eacute;`	Appel d'entité, correspond à « é ».
Appel	Java	`\u041B`	Le Л cyrillique.
Clavier	Windows	`Alt-133`	Correspond à « à » (numéro décimal en CP 850, un codage hérité de MS DOS).
Clavier	Windows	`Alt-0133`	Correspond à « … » (numéro décimal en Windows Latin 1, voir § 2.4, *Windows Latin 1*).
Menu + Tableau	Windows	`Démarrer > Programmes > Accessoires > Outils système > Table des caractères`	Cocher la case « Affichage avancé », sélectionner par numéro ou par bloc le caractère.
Clavier	Word en français	`41B Alt-C`	Le numéro hexadécimal du caractère Unicode suivi d'Alt-C est transformé en Л.
Menu + Tableau	Word 2003	`Insertion > Caractères spéciaux`	Sélectionner par numéro ou par bloc le caractère.
Menu + Tableau	Word 2007	`Insertion > Symbole > Autres symboles…`	
Clavier	Wordpad et Word en anglais	`41B Alt-X`	Correspond à Л, ne fonctionne pas dans MS Word français.
Appel	Perl	`chr(0x263C)`	Correspond au soleil blanc « ☼ ».
Appel	TeX ou LaTeX	`\biguplus`	U+2A04 ⨄ UNION N-AIRE AVEC PLUS.
Appel	XML, HTML	`&#xB6;`	Appel de caractère numérique hexadécimal (U+00B6 PIED-DE-MOUCHE).
Appel	XML, HTML	`&#182;`	Appel de caractère numérique décimal (U+00B6 PIED-DE-MOUCHE).

1.6.1 Claviers

Claviers préinstallés

Votre système d'exploitation comprend de très nombreuses définitions de clavier. Habituellement, une seule de ces définitions est activée : celle qui correspond aux touches gravées sur votre clavier. Il est cependant tout à fait possible d'ajouter d'autres définitions de clavier pour d'autres langues et que vous pourrez utiliser à loisir. Ceci signifie, implicitement, que les touches de clavier n'envoient pas directement des numéros de caractère au système d'exploitation, mais des numéros de touche que le pilote de clavier interprète en fonction de la définition de clavier active.

Dans MS Windows XP, l'ajout de clavier se fait par le même menu d'Options régionales et linguistiques (`Démarrer > Panneau de configuration > Options régionales et linguistiques`) sous l'onglet `Langues`, dans `Services de texte et langues d'entrée`, cliquer sur `Détails`. Choisir ici les langues et les dispositions de clavier qui vous conviennent.

Dans MS Windows Vista, l'ajout de clavier se fait par l'onglet Clavier et langues : (`Démarrer > Panneau de configuration`[1] `> Clavier et langues`) puis en cliquant sur le bouton `Modifier les claviers...` Choisir ici les langues et les dispositions de clavier qui vous conviennent.

Si vous installez plusieurs claviers, il est également utile de faire apparaître au bas de l'écran la barre des langues qui vous permettra de changer rapidement de clavier.

Pour ce faire, **dans Windows XP**, toujours dans fenêtre de Détails où vous avez ajouté des claviers, sous `Préférences`, cliquer sur `Barre de langue`, puis dans la boîte de dialogue `Paramètres de la barre de langue`, cocher la case `Afficher des icônes supplémentaires de la barre de langue dans la zone de notification`. Cliquer sur `OK` à trois reprises.

Pour faire apparaître cette barre des langues dans **Windows Vista**, cliquer sur l'onglet `Barre de langues`, puis cocher les options adéquates.

Pour changer de clavier, vous pourrez alors cliquer sur la barre de langue et manuellement changer de clavier ou alterner entre les claviers à l'aide de touches de raccourci (typiquement les touches *Alt* de gauche + *Maj*).

Attention — Cette dernière option explique pourquoi il est déconseillé d'installer des claviers supplémentaires sur la machine d'autrui sans l'avertir, puisque cette personne pourra avoir la désagréable surprise de se voir subitement taper en une langue exotique après avoir, par inadvertance, tapé sur les touches de changement de clavier.

1. En mode classique, passer par le niveau intermédiaire `Options régionales et linguistiques`, puis continuer avec l'onglet `Clavier et langues`.

Figure 1.7 – Choisir un autre clavier et préciser une autre langue

Habituellement à chaque langue n'est associé qu'un clavier et le simple fait de choisir une langue suffit à choisir un clavier. Toutefois, si vous souhaitez avoir accès à plusieurs dispositions de clavier pour une même langue (français AZERTY et français canadien QWERTY), un petit clavier apparaît à côté du code langue dans la barre de langue. Ce clavier vous permettra de choisir manuellement la disposition de clavier que vous préférez pour cette langue. Dans l'exemple précédent, le système définit deux dispositions de clavier possibles pour l'arabe (marocain), nous avons choisi la variante à 101 touches (la variante à 102 touches pourrait être utilisée avec un clavier externe, par exemple).

> Le même principe existe sous Mac OS X, consultez <http://hapax.qc.ca/polices-et-clavier.html> pour un exemple d'installation et de sélection d'un clavier tifinagh (l'écriture touarègue).

Claviers virtuels

Maintenant que vous avez installé un clavier étranger, si vous sélectionnez ce clavier, les lettres gravées sur les touches de votre clavier sont probablement de peu d'utilité. Idéalement, au changement de clavier, les lettres sur les touches devraient changer. On l'a tenté à l'aide de diodes électroluminescentes[1], malheureusement ce genre de clavier n'a pas réussi à s'imposer. À la place, la plupart des systèmes proposent des claviers virtuels.

Pour afficher le clavier virtuel qui correspond au clavier courant sous **Windows XP**, cliquer sur `Démarrer`, pointer sur `Tous les programmes`, sur `Accessoires`, sur `Accessibilité`, puis cliquer sur `Clavier visuel`. Sous **Windows Vista**, cliquer sur `Démarrer`, pointer sur `Tous les programmes > Accessoires > Options d'ergonomie > Clavier visuel`.

Le clavier virtuel sert non seulement d'aide-mémoire, mais il permet aussi la saisie en pointant et cliquant sur les touches représentées sur ce clavier visuel. La figure 1.8 correspond au clavier visuel pour le clavier marocain arabe.

1. Disponibles ici <http://www.artlebedev.com/everything/optimus/>.

Figure 1.8 – Clavier visuel marocain arabe

Définir son propre clavier

Le format réduit de ce livre ne permet pas de décrire la manière de composer soi-même un nouveau pilote de clavier. C'est aujourd'hui chose assez facile sur les systèmes modernes pour les claviers à la disposition relativement simples. Si vous désirez créer un pilote de clavier pour Windows, utilisez l'utilitaire MSKLC[1]. Sur Mac OS/X (versions 10.2 et ultérieures), deux utilitaires : Ukelele[2] et KeyLayoutMaker[3]. Les claviers tifinaghs mentionnés ici[4] ont été créés à l'aide de MSKLC pour Windows et Ukelele pour Mac OS X.

Méthode d'insertion directe par numéro de caractère

Dans MS Word et MS Wordpad, par exemple, si vous connaissez la valeur (hexadécimale) Unicode d'un caractère, vous pouvez insérer directement ce caractère dans votre document à l'aide du raccourci clavier Alt-X (Alt-C dans Word en français) :

- Tapez la valeur (hexadécimale) Unicode du caractère. La chaîne de valeur peut également commencer par U+.
- Appuyez sur Alt-X (Alt-C dans Word en français). L'application remplace la chaîne située à gauche du point d'insertion par le caractère spécifié.

Remarquez qu'un raccourci particulier à l'éditeur dans lequel vous désirez saisir, ou une macro définie après l'installation, un caractère à l'aide d'Alt-X (ou Alt-C) pourrait masquer cette méthode de saisie du système d'exploitation. La substitution Alt-X est fournie par un module appelé Uniscribe dont nous parlerons par la suite (voir § 13.8, *Un moteur de rendu : Uniscribe*).

Il existe également deux méthodes de saisie plus anciennes sur Windows qui utilisent des numéros de caractères décimaux. Il s'agit des méthodes Alt-*n* et Alt-0-*n*.

1. <http://www.microsoft.com/globaldev/tools/msklc.mspx>
2. <http://scripts.sil.org/ukelele>
3. <http://scripts.sil.org/keylayoutmaker>
4. <http://hapax.qc.ca/polices-et-clavier.html>

Ces méthodes fonctionnent avec la quasi-totalité des applications, contrairement à la méthode d'Alt-X qui dépend de l'utilisation par le programme du module d'écriture complexe appelé Uniscribe.

La méthode Alt-*n*, quand *n* est ≤ 255, précise le caractère dans la page DOS locale, il s'agit de CP 850 pour les systèmes francophones où Alt-133 correspond à « à ». Dans certains programmes plus récents de Windows, le *n* peut être plus grand que 255. Il s'agit alors de la valeur décimale Unicode du caractère. Alt-8470 correspond ainsi à « № » (U+2116 SYMBOLE NUMÉRO). Ceci fonctionne dans MS Word et Wordpad sur Windows XP et Vista.

La méthode Alt-0-*n* précise le caractère dans la page Windows locale (Latin-1 pour les systèmes français, Windows 1251 en Russie). Dans ce cas-ci, le *n* doit toujours être plus petit ou égal à 255. Alt-0133 correspond sur une machine Windows francophone aux points de suspension « … ».

1.6.2 Méthodes d'entrée extrême-orientales

Une méthode d'entrée est un programme utilisé pour permettre la saisie des milliers de caractères différents des langues dites CJC (chinois, japonais, coréen), à partir d'un clavier normal à 101 touches. Un éditeur de méthode d'entrée est constitué à la fois d'un moteur qui convertit les frappes de touches en caractères phonétiques et idéographiques, et d'un dictionnaire des mots idéographiques les plus usités. Au fur et à mesure que l'utilisateur frappe sur les touches, le moteur de l'éditeur de méthode d'entrée tente de reconnaître le ou les caractères et de les convertir en idéogrammes à l'aide du dictionnaire et de différents algorithmes.

Sous MS Windows, on ajoute et on active une méthode d'entrée de la même manière que les claviers (figure 1.7). Après avoir sélectionné la méthode d'entrée (un type de clavier chinois dans l'exemple suivant), il est alors possible de commencer à saisir des idéogrammes. Dans notre cas, nous avons choisi une méthode phonétique d'entrée, on tape du pinyin et la méthode affiche l'idéogramme qui y correspond le mieux. Si ce choix n'est pas adéquat, l'utilisateur peut choisir un autre idéogramme manuellement.

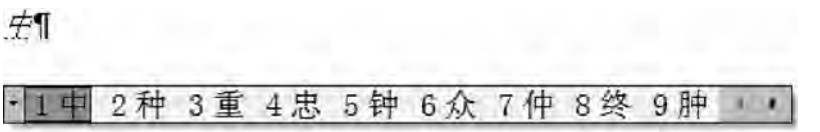

Figure 1.9 – Méthode d'entrée pour idéogrammes

Dans l'exemple de la figure 1.9, l'utilisateur vient de taper « zhong » et la méthode d'entrée a substitué à cette transcription pinyin l'idéogramme 中. L'utilisateur, à l'aide d'une touche (flèche vers l'arrière ici), a ensuite voulu choisir un autre idéogramme correspondant à « zhong ». C'est pourquoi l'éditeur de méthode d'entrée propose en dessous du point d'insertion les meilleurs autres idéogrammes qui correspondent à « zhong ».

1.6.3 Tableau de caractères

Il existe de nombreuses applications qui permettent de choisir des caractères à partir d'un tableau.

Une fois n'est pas coutume, commençons par le cas du Mac OS X 10.2 qui possède une jolie palette de caractères livrée en standard qui permet de sélectionner des caractères Unicode et de les insérer dans la fenêtre d'édition[1].

Pour Windows, il existe un tableau de caractères offert par le système d'exploitation et un autre offert par MS Office. Le tableau de caractères de MS Windows permet la recherche de caractères par leur nom (ISO 10646 en français). On peut ainsi donc facilement trouver toutes les flèches disponibles dans une police. La version d'Office, pour sa part, comprend également une liste de caractères fréquemment utilisés qui ne sont pas accessibles à l'aide des claviers habituels.

Tableau de caractères Windows

Windows offre un outil de sélection de caractères à l'aide d'une grille de caractères. Pour y avoir accès passer par `Programmes > Accessoires > Outils système > Table des caractères`. Une fois la grille affichée, cocher de préférence la case Affichage avancée. Double-cliquer sur les caractères que vous désirez voir copier dans le presse-papiers. Selon l'application cible, il est aussi possible de copier immédiatement le caractère sélectionné en le glissant vers cette application. Dans l'exemple de la figure 1.10, on a sélectionné toutes les flèches répertoriées dans la police *Lucida Sans Unicode* afin de pouvoir choisir la meilleure. Pour plus d'informations sur l'utilisation de cet outil, cliquez sur le bouton `Aide` de la fenêtre. Malheureusement, l'outil de Windows ne vous permet pas d'insérer des caractères ajoutés à Unicode après la sortie du système d'exploitation ou même quelques années auparavant. Ceci signifie par exemple qu'on ne peut sélectionner aujourd'hui sous Windows XP des caractères tifinaghs dans une police berbère ou touarègue qui code ces caractères avec les valeurs approuvées par Unicode depuis 2005.

Tableau de caractères MS Office

On accède au tableau Unicode des applications **Office 2003** en choisissant l'option du menu principal `Insertion`, puis `Caractères spéciaux`... Dans **Office 2007**, passer par le menu `Insertion > Symbole > Autres symboles`... Ensuite, dans toutes les versions d'Office, choisir l'onglet `Caractères spéciaux` pour insérer des caractères typographiques fréquents, mais absents des claviers européens, la liste est courte et donc plus pratique à consulter qu'un long tableau reprenant tous les caractères Unicode. Cette même liste énumère les raccourcis qui permettent d'entrer la plupart de ces caractères à l'aide de quelques touches. Exemple : Ctrl-- (c'est-à-dire « Ctrl » suivi du « - ») pour insérer le TRAIT D'UNION CONDITIONNEL.

1. Pour plus de détails, voir <http://www.osxfacile.com/palette.html>.

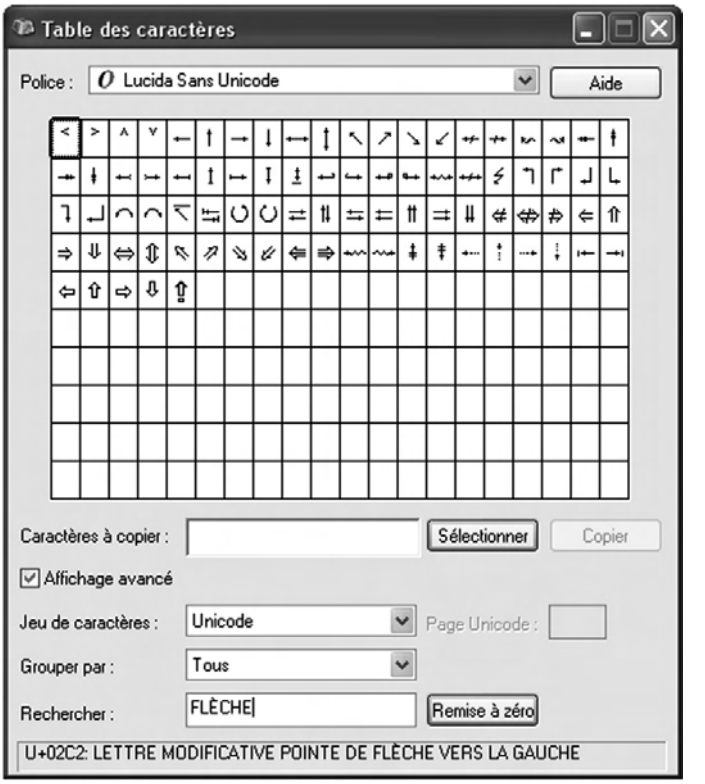

Figure 1.10 – Table des caractères de MS Windows

Pour passer à la grille des caractères Unicode, cliquer à l'endroit où vous souhaitez insérer le symbole, dans le menu `Insertion`, cliquer sur `Caractères spéciaux`, puis sur l'onglet `Symboles`. Dans la zone `Police`, cliquer sur la police souhaitée. Si vous utilisez une police qui comprend plus de 255 caractères, comme *Arial* ou *Times New Roman*, la liste `Sous-ensemble` s'affiche. Cette liste vous permet de choisir parmi de nombreux caractères, y compris des caractères grecs et russes (alphabet cyrillique), s'ils sont disponibles. Enfin, double-cliquer sur le caractère que vous souhaitez insérer. Lorsque vous sélectionnez un caractère Unicode sous l'onglet `Symboles`, le numéro du caractère s'affiche dans la zone `Code`[1] `du caractère`. Si vous connaissez déjà le numéro du caractère, vous pouvez le taper directement dans cette zone pour accéder au caractère Unicode.

Malheureusement, l'outil d'Office ne vous permet pas d'insérer des caractères ajoutés à Unicode après la sortie de votre version d'Office ou même quelques années auparavant. C'est pourquoi si votre machine est une machine Windows nous vous recommandons vivement l'option suivante : BabelMap.

1. Il s'agit d'un fâcheux anglicisme de la part de Microsoft, car en français un code est un ensemble de règles, un recueil de lois, un système de symboles, pas un élément de cet ensemble ou de ce système.

Tableau de caractères BabelMap

BabelMap est probablement l'outil le plus complet sous Windows pour la recherche de caractères Unicode. On peut bien sûr, comme pour toutes les palettes de caractères, cliquer sur une case et copier le caractère correspondant dans un tampon, mentionner un numéro de caractère et le voir s'afficher pour ensuite le copier, afficher les caractères d'une écriture particulière et en choisir un ou encore chercher un caractère par son nom comme dans l'outil de Windows. Mais, là où BabelMap semble imbattable, c'est dans la recherche poussée des caractères. On peut ainsi chercher un caractère non seulement en fonction de son nom officiel, mais aussi de ses synonymes non officiels, des commentaires qui lui sont associés dans les tableaux officiels, selon sa propriété. On peut ainsi afficher tous les signes de ponctuation ou tous les symboles monétaires pour une version particulière d'Unicode ! Les options de recherche sont innombrables[1].

Copier-coller et presse-papiers

Une autre manière de saisir des caractères est de les copier d'un document qui les contient déjà ! C'est en fait ce que vous faites après avoir sélectionné des caractères dans une application comme BabelMap. Dans Windows, on copie le texte sélectionné à l'aide de Ctrl-c dans ce qu'on appelle le presse-papiers et on colle le contenu du presse-papiers dans une fenêtre à l'aide de Ctrl-v.

Les applications qui ajoutent du contenu au presse-papiers peuvent copier ce même contenu sous plusieurs formats (RTF, image, HTML, texte brut « ANSI »[2], texte brut Unicode, etc.) pour augmenter le nombre d'applications susceptibles de copier ces données. Dans le cas de BabelMap, l'information n'est copiée que sous la forme de texte brut Unicode, mais de nombreuses autres applications copient le texte avec leur formatage (par exemple en RTF ou HTML). Quand vient le temps de recopier ce texte du presse-papiers vers votre traitement de texte, celui-ci copiera habituellement le texte formaté dans votre document.

Cela est souvent utile, mais cela peut-être également très déplaisant, car vous voilà soudain avec du texte dont le formatage ne respecte peut-être pas du tout votre gabarit et qui pourrait contenir des hyperliens dont vous n'avez que faire. Pour éviter ce désagrément, copier d'abord le texte du presse-papiers vers un éditeur de texte brut (comme TextPad ou NotePad) celui-ci ne sélectionnera pas la version formatée du presse-papiers mais la version en texte brut, puis réselectionner ce que vous venez de copier, le copier dans le presse-papiers (Ctrl-c) et enfin le coller (Ctrl-v). Cette fois-ci, il sera dépouillé de tout formatage.

---

1. Pour plus de détails, voir <http://hapax.qc.ca/BabelMap_fr.html>.
2. Ce nom déroutant et erroné est perpétué par Microsoft dans le sens de code de caractères propriétaires Windows, dans le cas des codes sur 8 bits ils sont proches des codes ISO/CEI. Contrairement à leur nom, ces codes n'ont pas été normalisés par l'ANSI, l'organisme de normalisation américain.

MS Word 2003 et **Open Office** permettent de s'en tirer autrement : il faut non plus copier le texte du presse-papiers à l'aide de Ctrl-v, mais grâce à `Edition > Collage spécial`, puis choisir `Texte Unicode sans mise en forme` (`Texte non formaté` dans Open Office). Pour **MS Word 2007**, passer par le menu `Accueil > Coller > Collage spécial` (Alt-Ctrl-v).

1.7 INTERNATIONALISATION ET LOCALISATION

L'internationalisation (mot souvent abrégé en i18n – la lettre initiale « i » du mot suivie des 18 lettres intermédiaires et le tout terminé par un « n » final) est un terme général qui désigne le processus qui consiste à préparer les logiciels afin qu'il puisse servir plus d'une culture, afficher plus d'une langue, s'utiliser sur plus d'un marché. C'est un processus technique qui ne requiert aucun talent de traducteur. Une des techniques de base de l'internationalisation consiste à extraire du code d'un programme tous les messages qu'il affichera et de les regrouper dans un fichier séparé qui pourra être traduit sans devoir modifier (ou même consulter) le code de ce programme. D'autres techniques d'internationalisation consistent à structurer le code de telle façon qu'il emploie des services (des bibliothèques logicielles) eux-mêmes internationalisés pour ce qui est du tri, du formatage des dates et montants afin que le code ne dépende plus d'une langue particulière.

La localisation (l10n) est l'adaptation d'un logiciel à destination d'une culture, ou d'une langue particulière. Cette culture particulière se nomme la « locale » en jargon informatique. Plus l'i18n est bien conçue, plus la localisation est techniquement facile à effectuer. La localisation implique principalement la traduction, mais elle ne se limite pas à cette activité, l'adaptation peut également s'intéresser aux aspects suivants :

- la modification des formats de date ou de montant ;
- la modification de la devise ;
- le tri des données qui pourrait devoir être adapté ;
- l'utilisation judicieuse de couleurs, d'icônes, de symboles adaptés à la culture cible ;
- les modifications imposées par la loi du pays visé, etc.

Nous étudierons les techniques d'internationalisation au chapitre 12.

Résumé

Dans ce chapitre nous avons d'abord vu ce qui a justifié la création d'Unicode : la multiplicité des jeux de caractères, multiplicité qui soulève de nombreuses difficultés lors de l'échange de données et la conception de logiciels internationalisés.

Le mot caractère prend de nombreux sens et il est important de les distinguer du mot « glyphe » (ou œil) qui représente une forme particulière d'un caractère et du terme « graphème » (ce que l'utilisateur perçoit comme une lettre, une entité distinctive de son écriture). Dans certains cas, comme le « c'h » breton, un graphème est composé de plusieurs caractères dans le sens d'éléments d'un jeu de caractères.

Les typographes utilisent une terminologie particulière pour décrire la dimension et l'apparence d'un caractère, les termes les plus importants sont :
– l'approche (le blanc imprimé avant ou après un caractère),
– le talus (le blanc imprimé au-dessus ou en dessous d'un caractère),
– la chasse (la largeur du rectangle dans lequel s'inscrit le caractère imprimé, ce rectangle comprend les approches droite et gauche),
– le corps (la hauteur de ce rectangle, ce rectangle comprend les talus de pied et de tête),
– les empattements (les petits traits aux extrémités d'un jambage).

On oppose le texte brut (sans balisage, sans formatage) au texte riche (balisé ou formaté).

Afin de bien fixer les idées, nous avons brièvement décrit Unicode (un code de près de 100 000 caractères, une série de propriétés affectées à ces caractères et une série d'algorithmes de référence) et surtout ce qu'Unicode n'est pas : Unicode n'est ni un produit, ni un format de mise en pages ni une solution miracle aux problèmes de traduction de logiciel ou de documents.

Par la suite, nous avons vu comment saisir des caractères sur Windows et dans plusieurs logiciels et langages de programmation et comment s'assurer qu'une police était bien une police Unicode.

Pour conclure, nous avons décrit ce qu'on nomme les caractères latins (ceux qu'on utilise en français et en anglais), les caractères chinois, coréens et japonais également appelés caractères han ou idéogrammes CJC. Enfin, nous avons introduit deux termes importants dans le domaine qui nous concerne et qu'il faut distinguer : l'internationalisation (la préparation des logiciels, des documents pour qu'ils puissent traiter plus d'une écriture, plus d'une langue) et la localisation (l'adaptation linguistique et culturelle d'un document ou d'un logiciel, elle suppose le plus souvent la traduction). Une bonne internationalisation facilite la localisation.