

Annexe G

Glossaire

À chasse fixe

On dira d'une police de caractères qu'elle est à chasse fixe, ou monochasse, quand tous les caractères ont la même chasse, comme c'est le cas habituellement sur une machine à écrire. Antonyme : à chasse variable ou à espacement proportionnel.

À chasse nulle

Caractéristique de certaines espaces, de caractères de commande ou de formatage qui ne consomment pas de place le long de la ligne de base horizontale. Cf. *Signe à chasse nulle*.
Synonyme : *sans chasse*.

Accent

Signe placé au-dessus, en dessous ou à côté d'un caractère, souvent afin d'altérer sa valeur phonétique. Cf. également *Diacritique*.

À éviter

Caractère codé fortement déconseillé. Unicode conserve ce type de caractères mais ils doivent être évités. Voir la définition D7a à la section 3.3, *Caractères et représentations codées*. À distinguer de *Désuet*.

Alphabet

Ensemble de symboles qui, dans le contexte d'une langue écrite particulière, sert à représenter les sons de cette langue. La correspondance entre les symboles et les sons peut être plus ou moins forte ; la plupart des alphabets ne présentent pas de correspondance biunivoque entre les sons distincts (phonèmes) et les symboles distincts (graphèmes).

Alphasyllabaire

Système d'écriture où les voyelles sont représentées par des signes diacritiques écrits au-dessus des consonnes ou à proximité de celles-ci.

À même le texte.

Cf. Dans le texte.

Annotation.

Association d'un contenu textuel secondaire à un passage textuel principal. Contrairement au *balisage*, on considère que la valeur de cette annotation fait partie du « contenu » du texte. Quelques exemples typiques : gloses, citations, ampliements, yomi japonais, etc.

Anounâssika

Cf. *Tchandrabindou*.

Anousvâra

Terme sanscrit employé dans les écritures dérivées de la *brâhmî* pour désigner des aspects de la nasalisation d'une voyelle sous l'influence d'une consonne nasale voisine, par un phénomène comparable à celui qu'on observe en français méridional (prononciation provençale de l'*a* dans *année*). L'*anousvâra* s'écrit en dévanâgarî à l'aide d'un point posé au-dessus du caractère qu'il modifie. C'est aussi généralement le cas dans les autres écritures dérivées de la *brâhmî*. On parle également de *bindou*.

ANSI

(1) Abréviation de l'organisme de normalisation américain, l'*American National Standards Institute*. (2) Nom collectif donné par Microsoft à toutes les pages de code de Windows. Ce nom est quelque fois utilisé spécifiquement pour qualifier la page de code 1252, un sur-ensemble d'ISO/CEI 8859-1.

Antialias

Caractère invisible qui affecte la liaison des caractères voisins lors du rendu. Voir la section 9.2, *Arabe* et « Liaison cursive » dans la section 14.2, *Commandes de disposition*.

API

(1) Alphabet phonétique international. (2) Association phonétique internationale responsable de la définition de l'Alphabet phonétique international et de sa mise à jour.

ASCII

Acronyme de *American Standard Code for Information Interchange*, un code à 7 bits, qui est la variante américaine de la norme ISO/CEI 646. Officiellement, il s'agit de la norme américaine ANSI X3.4.

Balise

Association d'attributs textuels ou structurels à un passage textuel principal. En règle générale, on ne considère pas qu'une balise fait partie du « contenu » textuel. SGML et XML sont des langages de balisage structurel, les étiquettes Unicode/ISO 10646 sont des balises linguistiques. Certaines balises indiquent également la police à utiliser pour un passage particulier.

Bas de casse

Synonyme de *minuscule*. Cf. *Casse*.

Base de données des caractères Unicode

Ensemble de fichiers qui fournit des correspondances et des propriétés normatives et informatives reliées aux caractères Unicode. Cf. 4, *Propriétés des caractères*, et le fichier *UnicodeCharacterDatabase.html* sur le disque accompagnant ce livre.

Bicaméral

Écriture qui distingue entre les *majuscules* et les *minuscules*. Terme utilisé le plus souvent dans le contexte des alphabets européens.

Bidi

Abréviation de bidirectionnel. Utilisé pour désigner un texte pouvant s'écrire de droite à gauche et de gauche à droite.

Bloc

Regroupement de caractères apparentés au sein de l'espace de codage d'Unicode. Un bloc peut comprendre des positions non attribuées qui sont alors réservées.

Bloc de caractères

Voir *Bloc*.

BNF

Abréviation de *Forme de Backus-Naur*, une métasyntaxe formelle utilisée pour décrire des syntaxes hors-contexte

Bopomofo

Transcription alphabétique du chinois principalement utilisée à Taïwan (Formose), utile pour l'enseignement et la saisie de texte. Le bopomofo permet de transcrire le mandarin ainsi que certaines langues minoritaires. Chaque symbole correspond aux sons initiaux ou aux sons finaux de la syllabe. Le nom bopomofo est dérivé du nom des quatre premiers éléments, repris ci-dessous : ㄅ ㄆ ㄇ ㄏ. Cette écriture se nomme *zhuyin zimu* ou *zhuyin fuhao* en mandarin.

Boustrophédon

Type d'écriture (grec archaïque, étrusque, etc.) dans lequel les lignes se succèdent dans l'ordre où l'on trace les sillons d'un champ, c'est-à-dire alternativement de gauche à droite et de droite à gauche. Les glyphes utilisés dans une direction sont habituellement l'image miroir de ceux utilisés dans la direction opposée. On parle alors de glyphes spéculaires.

Brâhmî

Écriture historique de l'Inde. La brâhmî a donné naissance à partir du IV^e siècle aux écritures indiennes ultérieures (télougou, kannara, tamoul, malayalam, goudjarati, etc.). Parmi les écritures indiennes dérivées de la brâhmî, la nâgarî (ou *dévanâgarî*) mérite une place à part. Avec l'apparition de l'imprimerie, elle est devenue l'écriture principale du sanscrit. Elle sert également aujourd'hui à noter le hindî.

Braille

Écriture constituée de points en relief à l'usage des aveugles ou des mal voyants. Voir la section 13.9, *Braille*.

Canonique

(1) Qui se conforme aux règles générales de codage, c'est-à-dire ni compressé, ni compacté ni dans toute autre forme indiquée par un protocole de plus haut niveau. (2) Caractéristique d'une transposition normative et d'une forme d'équivalence décrite dans 3, *Conformité*.

Capitale

Synonyme de majuscule. Cf. *Casse*.

Caractère

(1) Le plus petit élément d'une langue écrite pourvu d'une valeur sémantique au niveau du sens ou de la forme abstraite plutôt que d'une forme particulière (Cf. également *Glyphe*), bien que dans les tableaux de code il soit essentiel de se référer à une représentation visuelle particulière pour que le lecteur reconnaisse le caractère concerné. (2) Synonyme de *caractère abstrait*. Voir la définition D3 à la section 3.3, *Caractères et représentations codées*. (3) L'unité de base utilisée par le codage de caractères d'Unicode. (4) Le nom français des éléments idéophonographiques d'origine chinoise. Cf. *Idéophonogramme*.

Caractère abstrait

Unité d'information utilisée pour organiser, commander ou représenter des données textuelles. Voir la définition D3 à la section 3.3, *Caractères et représentations codées*.

Caractère attribué

Synonyme de *caractère codé*.

Caractère codé

Caractère abstrait et sa *valeur scalaire Unicode* associée (le numéro entier qui lui est associé). Le caractère abstrait seul n'a pas de valeur numérique, c'est le « codage du caractère » qui lui attribue une valeur scalaire Unicode. Cette association constitue un « caractère codé ».

Caractère combinatoire

Caractère qui se combine graphiquement avec le caractère de base précédent. On dit que le caractère combinatoire est adjoind à ce caractère de base. Voir la définition D14 à la section 3.5, *Combinaison*. Cf. également *Signe à chasse nulle*.

Caractère complémentaire

Caractère Unicode codé appartenant au plan complémentaire.

Caractère composite

Cf. *Caractère décomposable*.

Caractère de base

Caractère qui ne se combine pas graphiquement avec les caractères précédents et qui n'est ni un caractère de commande ni un caractère de formatage. Voir la définition D13 à la section 3.5, *Combinaison*.

Caractère de compatibilité

(1) Caractère codé uniquement à des fins de compatibilité avec les normes de codages de caractères préexistantes et de transcodage entre celles-ci et Unicode. (2) Un caractère qui possède une décomposition de compatibilité, Voir la définition D21 à la section 3.6, *Décomposition*.

Caractère décomposable

Caractère équivalant à une suite d'un ou plusieurs autres caractères, selon les correspondances de décomposition précisées dans la liste des noms de la section 15.1, *Liste des noms de caractères*. Également appelé *Caractère précomposé* ou *Caractère composite*. Cf. la définition D18 à la section 3.6, *Décomposition*.

Caractère de description idéographique

Caractère graphique utilisé avec d'autres caractères graphiques pour former une suite de description idéographique (SDI). Une telle suite peut être utilisée pour décrire un idéogramme ne faisant pas partie d'Unicode.

Une SDI décrit un idéogramme de façon abstraite. Elle n'est pas interprétée comme une séquence composite et n'entraîne aucune forme de présentation particulière.

Une SDI n'est pas un caractère et par conséquent ne fait pas partie du répertoire.

Caractère de remplacement

Caractère utilisé à la place d'un caractère non interprétable issu d'un autre codage. Unicode utilise U+FFFD CARACTÈRE DE REMPLACEMENT à cet effet.

Caractère du PMB

Caractère Unicode codé appartenant au plan multilingue de base.

Caractère graphique

(1) Caractère typiquement associé à une représentation visible y compris n'importe quel type d'espace. Cf. également *Glyphe*. (2) Tout caractère qui n'est pas avant tout associé à une fonction de commande ou de formatage.

Caractère neutre

Caractère que l'on peut écrire de droite à gauche ou de gauche à droite selon le contexte. Voir la section 3.12, *Comportement bidirectionnel*.

Caractère précomposé

Cf. *Caractère décomposable*.

Caractères han

Écriture idéophonographique chinoise. Elle est dérivée de 214 (ou 227 selon les dictionnaires) clés ou radicaux dont la présence dans un signe indique habituellement la catégorie de choses, d'idées, etc. représentée par le caractère. Exemples : 中, 國. Syn. Sinogramme.

Casse

(1) Trait de certains alphabets où les lettres ont deux formes distinctes. À l'origine, la casse était une boîte plate divisée en compartiments (les *cassetins*). On distingue les lettres du *haut de casse*, ou *capitales*, ou encore *majuscules*, et les lettres de *bas de casse*, ou *minuscules*. Ces variantes peuvent différer sensiblement dans leur aspect et taille. (2) Propriété normative de caractères : majuscule, minuscule, casse de titre (*Lu*, *Ll* et *Lt*). Voir la section 4.1, *Casse – normatif*.

Casse de titre

Lettre initiale majuscule, les autres lettres du mot étant en minuscules. Dans certaines langues, la lettre initiale en casse de titre peut avoir une forme de glyphe différente d'une lettre majuscule. Cf. *Casse*.

Casseau

Dans la typographie ancienne, moitié de *casse*, à grands compartiments, servant de réserve pour différents caractères. Dans le cas d'Unicode, il s'agit d'un ensemble de caractères décoratifs ou de fantaisie.

Catégorie générale

Division des caractères en classes principales telles que les lettres, la ponctuation ou les symboles et en sous-classes subséquentes pour chacune de ces classes principales. Cf. la section 4.5, *Catégorie générale – en partie normatif*.

Cédille

Un signe placé à l'origine sous la lettre c en français, en portugais et en espagnol pour indiquer que la lettre doit être prononcée comme un s (cf. *açã*o). Diminutif désuet espagnol de *ceda*, la lettre z.

Cellule

Dans la terminologie de l'ISO/CEI 10646, place dans une rangée à laquelle un caractère isolé peut être affecté.

Cellule de rendu

Zone rectangulaire sur le dispositif d'affichage au sein de laquelle un ou plusieurs glyphes sont représentés.

Chiffres

Cf. *Chiffres arabes*, *Chiffres européens*, *Chiffres de l'Inde*.

Chiffres arabes

Formes particulières des chiffres décimaux utilisés dans la plus grande partie du monde arabe (par exemple U+0660 ٠, U+0661 ١, U+0662 ٢, U+0663 ٣). Bien que les *chiffres européens* soient issus de ces formes, elles sont visuellement distinctes et sont codées séparément. On désigne parfois les chiffres arabes sous le nom de chiffres indiens, cette désignation entraîne cependant une ambiguïté inutile avec les chiffres utilisés actuellement par les langues de l'Inde. Unicode désigne les chiffres arabes sous le nom de *chiffres arabo-hindî*. Des variantes de ces chiffres sont utilisées principalement en Iran et au Pakistan, on leur donne le nom de *chiffres arabo-hindî orientaux*.

Chiffres de l'Inde

Formes particulières des chiffres décimaux utilisés par plusieurs écritures d'origine *brâhmî* (par exemple, en dévanâgarî : U+0966 ६ U+0967 ७ U+0968 ८ U+0969 ९). Les *chiffres arabes* (et, par suite, les *chiffres européens*) dérivent de ces formes.

Chiffres décimaux

Chiffres qui peuvent être utilisées dans les nombres à base dix.

Chiffres européens

Formes des chiffres décimaux d'abord utilisées en Europe et aujourd'hui à l'échelle planétaire. Ces chiffres dérivent historiquement des chiffres arabes, on les appelle donc parfois « chiffres arabes », il y a lieu de préférer la forme « chiffres européens » qui ne porte pas à confusion avec les *Chiffres arabes* d'Unicode.

Chữ hán

Nom des caractères chinois au Viêt-nam dérivés du *hanzi*.

Chữ nôm

Mot vietnamien qui signifie écriture populaire, démotique ou vulgaire. Écriture en usage au Viêt-nam avant l'adoption de l'écriture latine au XIX^e siècle. Le *chữ nôm* est basé sur l'écriture chinoise avec une prononciation et une composition particulière au Viêt-nam.

CJC

Abréviation de chinois, japonais et coréen. Une variante, le *CJCV*, représente le chinois, le japonais, le coréen et le vietnamien.

Classe combinatoire

Une valeur numérique attribuée à chaque caractère Unicode qui précise les autres caractères combinatoires avec lequel ce caractère interagit au niveau typographique.

Classe de caractères

Jeu de caractères qui partage un ensemble précis de propriétés.

Classement

Tri des unités d'information textuelle. Chaque langue possède habituellement un classement particulier. On parle également de *tri alphabétique ou lexicographique*. Le rapport technique n° 10 « Unicode Collation Algorithm » définit un ordre déterminé, complet et univoque pour tous les caractères d'Unicode. Une norme ISO correspondante (14651) définit également un tel ordre.

Clé

Cf. *Radical*.

Code de glyphe

Une valeur de code qui désigne un glyphe. Habituellement, on identifie les glyphes contenus dans une police par leur code de glyphe. Les codes de glyphe sont généralement propres à une police particulière, en d'autres termes, une police différente comprenant les mêmes glyphes utilisera probablement d'autres codes de glyphe.

Codes de commande

Les soixante-cinq caractères des intervalles U+0000..U+001F et U+007F..U+009F. Également appelés caractères de commande.

Codes de formatage

Caractères invisibles mais qui affectent la présentation des caractères voisins.

Combinaison braille

Une des 64 (pour le braille à 6 points) ou 256 (pour le braille à 8 points) combinaisons possibles de points tangibles.

Compatibilité

(1) Cohérence avec la pratique courante ou antérieure des normes de codage de caractères. (2) Caractéristique d'une transformation normative et d'une forme d'équivalence précisée dans la section 3.6, *Décomposition*.

Composition dynamique

Création, à partir d'une suite de caractères, de formes composites, telles les lettres accentuées ou les syllabes hangûl

Conformité

Adhésion à un ensemble précis de critères relatifs à l'utilisation d'une norme ou d'un standard.

Conjointe consonantique

D'ordinaire, forme de présentation ligaturée d'un groupe consonantique. Ce terme s'applique plus particulièrement aux écritures d'origine *brâhmî*. Cf. *tranche*.

Consonne mi-formée

Également appelé *demi-forme*. En dévanâgarî, et dans d'autres écritures indiennes de la famille *brâhmî*, une consonne dévoyellée peut être représentée sous cette demi-forme. Cette consonne mi-formée reprend la forme du caractère-consonne mais sans sa hampe. Ce caractère *déhampé* ou *déhasté* peut alors servir à créer des formes conjointes, il s'agit d'ailleurs de la forme traditionnellement utilisée.

Consonne morte

Une consonne des écritures *brāhmī* suivie d'un caractère *virāma*. De ce fait, la consonne perd sa voyelle implicite. Cf. la section 10.1, *Dévanāgarī*.

Conversion par pivot

Utilisation d'un troisième codage de caractères comme étape intermédiaire lors d'une conversion entre deux codages de caractères. Le standard Unicode est fréquemment utilisé comme pivot puisque son répertoire est un sur-ensemble de celui de la plupart des autres jeux de caractères codés.

Correspondance de casses

Association des formes du haut de casse, bas de casse et casse de titre d'une lettre. Cf. la section 5.18, *Correspondance de casses*.

Crénage

Certains caractères ont un œil débordant du support – on dit que l'œil saille. Il peut se trouver dans le sens vertical pour placer un accent sur une capitale ou dans le sens latéral, pour certains caractères italiques par exemple, pour éviter des défauts d'approche trop évidents. Le crénage est le processus d'ajustement de l'approche de deux lettres adjacentes. Il est ainsi possible de comparer l'intervalle entre le A et le W dans HAWAĪ (non créné) et dans HAWAĪ (créné).

Cursif

Écriture où les lettres d'un mot se rejoignent à la manière de l'écriture manuscrite usuelle.

Dans le texte.

Informations codées à même le texte à l'aide d'une syntaxe particulière qui permet de les repérer. Les informations à même le texte sont codées dans le même jeu de caractère que le reste du texte, elles parsèment le texte et l'accompagnent. Les balisages XML et HTML en sont deux exemples.

DBCS

Cf. JC2O.

Décomposition

(1) Séparation ou analyse d'un élément textuel en ces composants. Ces composants peuvent n'avoir aucune valeur fonctionnelle mais être uniquement des unités formelles, c'est-à-dire des formes abstraites. (2) Cf. la définition D19 à la section 3.6, *Décomposition*.

Décomposition canonique

Cf. la définition D23 à la section 3.6, *Décomposition*.

Décomposition de compatibilité

Cf. la définition D20 à la section 3.6, *Décomposition*.

Définition de jeu de caractères

Ensemble pour lequel on attribue à chaque caractère qui en fait partie une valeur de code numérique appelée *point de code*, *valeur scalaire* ou plus simplement *numéro de caractère*. Syn. *Page de code* et *Jeu de caractères codés*.

Demi-chasse

Un caractère à demi-chasse est, *grosso modo*, deux fois moins large qu'un caractère similaire du même jeu ayant lui une chasse ordinaire. On dit alors que le caractère occupe une demi-cellule de rendu. Certains jeux de caractères, plus particulièrement ceux à deux octets tels que le Shift-JIS, codent deux formes différentes de certains caractères. Ces caractères à demi-chasse sont parfois codés sur un seul octet dans les jeux à nombre variable d'octets alors que les caractères à *pleine chasse* sont eux codés sur deux octets ; cette conception serait due aux premières adaptations de terminaux aux langues CJC, terminaux pour lesquels il existait une égalité stricte entre le nombre d'octets dans le flux et la chasse qu'ils occupaient à l'écran. Le terme japonais qui désigne ces caractères à demi-chasse est *hankaku*. Antonyme *zenkaku*.

Désuet

Se dit d'un caractère qui ne s'utilise plus. La désuétude d'un caractère dépend du contexte, la lettre *grand yousse* est désuète en russe, mais est toujours utilisée en bulgare moderne. À distinguer de *À éviter*.

Dévoyellement

Perte d'une voyelle. C'est la fonction du *virâma* qui crée par son adjonction à un caractère *brâhmî* une consonne dévoyellée. Cf. *halant* et *virâma*.

Diacritique

(1) Signe graphique adjoint à un symbole afin de créer un nouveau symbole qui représente une valeur nouvelle ou modifiée. (2) Signe adjoint à un symbole que celui-ci en change la valeur ou non. Dans ce cas, le diacritique représente habituellement une valeur indépendante (par exemple, un accent, un ton ou une autre information linguistique). Également appelé *signe diacritique*. Cf. également *Caractère combinatoire* et *Signe à chasse nulle*.

Diacritique à chasse nulle

Signe diacritique qui est également un signe à chasse nulle.

Digramme

Groupe de deux lettres employé pour transcrire un phonème unique ou un seul élément linguistique. L'orthographe française utilise de nombreux digrammes, par exemple : *th*, *ch*, *qu*, *ph*, etc. Ces deux lettres ne constituent pas toujours un digramme (cf. le *qu* dans les mots *quand* et *quantum*). On appelle *trigramme* un groupe de trois lettres ainsi combinées. Au-delà de trois, ces groupes sont habituellement appelés des *n-grammes*.

Diphthongue

Voyelle complexe dont le timbre se modifie au cours de son émission, par ex. [au] dans l'allemand *Auge*. Dans certains systèmes d'écriture, la diphthongue est parfois transcrite par un seul signe, parfois par plusieurs (par exemple par un *digramme*).

Direction d'écriture

Direction ou orientation des caractères écrits au sein de lignes textuelles d'un système d'écriture. Il existe trois directions habituelles dans les systèmes d'écriture moderne : gauche à droite, droite à gauche et de haut en bas.

Ductilité

Propriété d'une *police* cursive à étirer ou à étrécir la ligne de base qui relie les lettres à des fins de justification.

EBCDIC

Acronyme de l'Extended Binary-Coded Decimal Interchange Code. Un ensemble de jeux de caractères codés à 8 bits utilisés par les macroordinateurs. Soixante-quatre positions de code (x00 à x3F) sont réservées aux codes de commande, l'intervalle x41 à xFE est réservé aux caractères graphiques. L'alphabet de base est constitué de deux segments disjoints, les majuscules se trouvent de 0xC1 à 0xC9, xD1 à D9, xE2 à xE9, et les minuscules de x81 à x89, x91 à x99 et xA2 à xA9.

Écriture

Ensemble de symboles utilisés pour représenter des informations textuelles d'un ou plusieurs systèmes d'écriture.

Écriture démotique

(1) Se dit d'une écriture ou d'une forme d'écriture utilisée pour écrire la langue populaire d'une communauté linguistique. (2) Écriture cursive de l'ancienne Égypte (VII^e s. av. J.-C.-V^e s. av. J.-C.) dérivée de l'écriture hiératique, elle-même dérivée des hiéroglyphes monumentaux.

Élément textuel

L'unité textuelle minimale sur laquelle opère une manipulation de texte particulière, dans le contexte d'un système d'écriture donné. En général, il existe une correspondance n-n entre les éléments textuels et les unités de stockage.

Équivalence

Dans le contexte de manipulation de texte, relation de deux éléments qui sont identiques à certains égards.

Équivalent canonique

On dit que deux suites de caractères sont des équivalents canoniques si leurs décompositions canoniques complètes respectives sont identiques. Cf. *Équivalent en terme de compatibilité*.

Équivalent en terme de compatibilité

On dit que deux suites de caractères sont équivalents en terme de comptabilité si leurs décompositions de compatibilité complètes sont identiques. L'équivalence de compatibilité est plus lâche que l'équivalence canonique, elle élimine par exemple des différences de formatage entre des caractères proches. Cf. la définition D22 à la section 3.6, *Décomposition*.

Espace

En typographie, on désigne sous le nom d'une espace le blanc placé entre les lettres ou les mots.

Espace de code

Le domaine des valeurs numériques disponibles pour le codage des caractères.

Étiquette.

Une *balise* linguistique.

Fichier binaire

Fichier qui contient des données non textuelles.

Flottant (diacritique, accent, signe)

Cf. *Signe à chasse nulle*.

Fonte

Cf. *Police*.

Formage

Opération de mise en forme des caractères, par exemple le choix des glyphes contextuels appropriés.

Format transformé

Correspondance entre une suite de caractères codés et une suite unique d'unités de stockage.

Format transformé d'Unicode (ou du JUC)

Cf. la définition D29 à la section 3.8, *Transformations*, voir également la section C.3, *Formats de transformation JUC*.

Forme conjointe

Glyphe correspondant à la combinaison de deux ou de plusieurs glyphes représentant des consonnes. La forme conjointe est un type de ligature qui apparaît dans la plupart des écritures *brâhmî*.

En Unicode, la conjointe est formée à l'aide de *virâma* qui éliminent les voyelles des caractères pour ne laisser que les consonnes, cf. *Dévoyellement*. Ces *consonnes mortes* ou dévoyellées sont habituellement jointes à une forme consonantique subséquente pour former la conjointe. Les composants de la forme conjointe peuvent se lier horizontalement ou verticalement. Dans certains cas, on ne peut distinguer les composants dans la conjointe résultante.

Forme de présentation

Ligature ou variante de glyphe codée comme un caractère à des fins de compatibilité. Cf. également *Caractère de compatibilité* (1).

Forme de stockage

Cf. *forme stockée des caractères*.

Forme statique

Cf. *Caractère décomposable*.

Forme stockée des caractères

Correspondance entre les *numéros de caractère* (ou *points de code*) d'une définition de jeu de caractères et les *unités de stockage* (octet, seize, etc.) utilisées pour coder ceux-ci.

FSS-UTF

Abréviation de *File System Safe UCS Transformation Format*, format publié par la X/Open Company Ltd destiné à l'environnement Unix. Connu aujourd'hui sous le nom d'*UTF-8*.

Glyphe

(1) *Stricto sensu*, un trait gravé en creux dont la répétition constitue un ornement, cf. l'écriture maya. (2) Une forme abstraite qui représente une ou plusieurs *images de glyphes*. (3) Synonyme d'*image de glyphe* ou d'*œil*. Lors de l'affichage des caractères Unicode, un ou plusieurs glyphes peuvent être sélectionnés pour afficher un caractère particulier. Ces glyphes sont sélectionnés par un moteur de rendu pendant les processus de composition et de disposition.

Glyphe de remplacement

Glyphe utilisé pour rendre un caractère qui ne peut être rendu correctement dans une police particulière. Il s'agit souvent des rectangles blanc □ ou noir ■. Également appelé *glyphe manquant*. Cf. la section 5.3, *Caractères inconnus ou manquants*.

Glyphe manquant

Cf. *Glyphe de remplacement*.

Graisse

Épaisseur des traits d'un caractère. On parle également de tracés : maigre, normal, demi-gras, gras, extra-gras (noir), etc.

Graphème

(1) Plus petite unité distinctive et significative de l'écriture. Pour les écritures idéographiques, il peut représenter un concept. Dans l'écriture phonographique, il représente un élément de la réalisation phonique (syllabe, consonne, voyelle). Ainsi dans l'écriture alphabétique, le graphème est communément appelé lettre. En français, *p* et *s* sont des graphèmes distincts au sein du système d'écriture français puisque les mots *pot* et *sot* sont des mots différents. Par contre, *ɔ* et *ɑ* ne sont pas des graphèmes distincts puisque aucun mot ne diffère sur la base de ces deux formes différentes de cette même lettre. (2) Ce que l'utilisateur considère être un caractère.

Gros-boutien

Architecture informatique où l'octet le plus significatif des valeurs numériques multi-octets est stocké en premier. Cf. *Petit-boutien*.

Groupe

Terme utilisé dans l'ISO/CEI 10646 pour désigner une subdivision de l'espace de codage formée de 256 x 256 x 256 cellules. Tous les caractères du JUC et d'Unicode appartiennent au groupe 0.

Groupe consonantique

Suite de caractères qui représente une ou plusieurs consonnes.

Halant

Terme hindî, provenant du sanscrit *halanta*, il signifie « terminé par une consonne ». Ce terme désigne la fonction de dévoilement du caractère *virâma*. Cf. *Virâma*.

Hangûl

Nom de l'écriture coréenne. Cf *Jamo*.

Hanja

Nom coréen des caractères dérivé du *hanzi* (ou han-tseu).

Han-tseu

Cf. *hanzi*.

Hanzi

Le nom mandarin des *caractères han*.

Harakat

Signes qui indiquent les voyelles ou les autres modifications apportées aux consonnes dans l'écriture arabe. Cf. *Points-voyelles*.

Haut de casse

Synonyme de *majuscule*, Cf. *Casse*.

Hiragana

Une des deux séries de kana (48 signes syllabiques dont deux sortis de l'usage), l'autre étant le *katakana*. Les hiragana (les kana sans angles) servent à représenter les mots indigènes au japonais, les verbes auxiliaires et tous les éléments spécifiques, notamment les terminaisons, suffixes de substantifs et d'adjectifs, et autres particules.

Exemple : きていそ.

Hors texte.

Les informations hors texte transmettent des renseignements supplémentaires sur un texte associé de telle sorte que le contenu de ce texte demeure intact (rien n'y est ajouté, retranché ou modifié). En règle générale, on utilise à cette fin une structure de données supplémentaires qui pointe vers le texte à annoter.

HTML

Acronyme anglais de Langage de balisage hypertextuel. Langage de description textuel issu de SGML qui comporte des balises de formatage textuel et du contenu textuel brut afin de décrire du texte formaté. HTML est le langage source omniprésent qui sert à représenter les pages sur la toile, pages dites « *Web* ». À partir de HTML 4.0, le jeu de caractères de référence des données HTML est désormais l'ISO/CEI 10646. Cf. également *SGML*.

i18n

Cf. *Internationalisation*.

IANA

Acronyme de l'Internet Assigned Number Authority.

Identificateur de glyphe

Similaire au code de glyphe, l'identificateur de glyphe est une étiquette utilisée pour désigner un glyphe dans une police. Une police peut employer à la fois des identificateurs de glyphe locaux et globaux. L'AFII (*Association for Font Information and Interchange*) a défini un jeu d'identificateurs de glyphe globaux ou universels.

Identificateur de séquence Unicode

Représentation d'une suite de plusieurs points de code Unicode sous la forme d'une liste séparée à l'aide de virgules et entourée de crochets angulaire, une espace peut suivre chaque virgule.

Exemples :

<U+0061, U+0301>

<U+10339, U+0308>

<U+0633, U+0645, U+062C>

Les identificateurs de séquence Unicode (ISU) s'utilisent habituellement pour représenter une suite de caractères combinatoires, de digrammes ou de ligatures. On peut toutefois les

utiliser pour représenter n'importe quelle suite de points de code (n° de caractères) Unicode.

Idéogramme

(1) On appelle *idéogramme* un caractère graphique correspondant à une idée (concept, procès, qualité) mais n'en représentant pas le son. (2) Terme technique quelque peu imprécis pour désigner les caractères *han*, encore couramment employé en parlant du chinois à la place de *idéophonogramme*. Cf. *Idéophonogramme*.

Idéophonogramme

Caractère correspondant à une notion (idéogramme) ou au son constituée par un mot (phonogramme). Terme préféré à idéogramme pour désigner les caractères *han*.

Image de glyphe

Image concrète de la représentation d'un glyphe après son tramage ou son impression sur une surface d'affichage. Cf. *Œil*.

Indicateur d'ordre des octets

Caractère U+FEFF ESPACE INSÉCABLE SANS CHASSE quand il est utilisé pour indiquer l'ordre des octets d'un texte. Cf. la section 2.7, *Valeurs de non-caractères et de caractères spéciaux*, et la section 14.6, *Spéciaux*.

Informatif

Information dans le standard Unicode qui n'est pas impérative mais qui contribue à l'utilisation et à la mise en œuvre correctes de celui-ci.

Internationalisation

Technique permettant à un programme de s'adapter correctement à différents environnements culturels sans nécessiter de ré-écriture. Elle consiste à rendre le code source du logiciel neutre en termes de culture (indépendance de la langue, du jeu de caractères, etc.) Le terme s'abrège souvent en *i18n* (*i* suivi de 18 lettres suivies de *n*).

On oppose souvent le terme à la *localisation*.

IOO

Acronyme de *Indicateur d'ordre des octets*.

IRG

Abréviation anglaise du Groupe rapporteur idéographique, un sous-groupe de l'ISO/CEI JTC1/SC2/GT2. Voir Annexe H, *Historique de l'unification han*.

ISCI

Acronyme de l'Indian Standard Code for Information Interchange.

ISO

Organisation internationale de normalisation. Il ne s'agit pas d'un acronyme, mais le sigle prend son origine dans le mot grec « iso » lequel signifie égal (cf. isocèle, isotherme). En extrapolant, on peut considérer que « iso » signifie uniforme, normalisé. L'organisation internationale a choisi ce sigle pour sa valeur internationale et linguistiquement neutre.

Jamo

Nom coréen pour un des 40 éléments de base de l'écriture hangûl. Le mot signifie littéralement consonne ou voyelle. Les jamos n'apparaissent jamais isolément mais forment la base du *syllabaire* coréen.

JC1O

Abréviation de *Jeu de caractères à un octet*.

JC2O

Abréviation de *Jeu de caractères à deux octets*.

Jeu de caractères

Ensemble d'éléments utilisés pour représenter de l'information textuelle.

Jeu de caractères à deux octets

Codage des caractères pour les langues orientales (chinois, japonais ou coréen) qui permet la représentation des graphèmes à l'aide d'un ou de deux octets au sein d'un même flux de données. Exemple : JIS X 0208-1990. Abréviation : *JC2O* ou *DBCS*. Voir également *jeu de caractères multioctets*.

Jeu de caractères à un octet

Codage sur un octet, on oppose ce terme à *JC2O* et *JCMO*.

Jeu de caractères codés

Ensemble pour lequel on attribue à chaque caractère qui en fait partie une valeur de code numérique appelée *point de code*, *valeur scalaire* ou plus simplement *numéro de caractère*. Souvent abrégé en *jeu de caractères*.

Jeu de caractères multioctets

Jeu de caractères qui code chaque caractère sur un nombre variable d'octets. Plusieurs jeux de caractères importants sont multioctets afin de permettre une stricte compatibilité avec le sous-jeu ASCII et l'ISO/CEI 2022. Abrégé en *JCMO*.

JTC1

Comité technique mixte n° 1 de l'Organisation internationale de normalisation et de la Commission électrotechnique internationale. Ce comité est responsable de la normalisation dans le domaine des technologies de l'information.

JUC

Acronyme du jeu universel de caractères défini dans la norme internationale ISO/CEI 10646.

Kachidé

Mot persan (tatouil étant le terme arabe) qui désigne l'interlettrage utilisé pour rallonger les lignes afin de les composer, par exemple, en pavé. Le kachidé est un trait qui prolonge les liaisons entre les lettres.

Exemple :

مرحبا	<i>marhaba</i> (« bonjour ») sans tatouil
مرحبا	idem avec kachidé.

Kana

Signes de l'écriture japonaise à valeur syllabique formée des *hiragana* et des *katakana*.

Kanji

Caractères *han* tels qu'utilisés en japonais. Les *kanji* sont utilisés en fonction de leur sens, pour représenter les mots d'origine chinoise ou des mots indigènes.

Katakana

Une des deux séries de *kana* qui comporte 48 signes syllabiques dont 2 sortis de l'usage. Les *katakana* (*kana* simples) servent surtout à la transcription des mots étrangers. Exemple de *katakana* : サラリマン - *sa-ra-ri-ma-n* (prononcé *salalimane*). On reconnaît le mot anglais *salary-man*.

l10n

Cf. *Localisation*.

Lettre

(1) Élément d'un alphabet. Au sens large, comprend les éléments de syllabaires et les idéophonogrammes. (2) Propriété informative des caractères utilisés dans l'écriture de mots.

Lettre modificative

(1) Catégorie Lm dans la base de données de caractère Unicode. (2) Ensemble dans le bloc de lettres modificatives. Ressemble à une lettre ou à de la ponctuation et modifie la prononciation des autres lettres (similaires aux diacritiques). Cf. la section 8.8, *Lettres modificatives*.

Liant

Caractère invisible qui affecte la ligature des caractères voisins lors du rendu. Voir la section 9.2, *Arabe* et « Liaison cursive » dans la section 14.2, *Commandes de disposition*.

Ligature

Trait reliant deux lettres. Ensemble de lettres liées qui forme un caractère unique (p.ex. æ en français). L'arobe @ et la perluète & sont des ligatures historiques. Dans certaines langues, les ligatures de certaines lettres sont obligatoires. C'est le cas de la ligature lam-alif en arabe. La mise en œuvre de ce traitement nécessite une analyse contextuelle.

Locale

n.f., cf. Profil local.

Localisation

Adaptation régionale. Adaptation d'un logiciel ou de données pour son utilisation dans une culture particulière. Un logiciel internationalisé sera plus facile à localiser qu'un logiciel que ne le serait pas. Le terme s'abrège souvent en *l10n* (*l* suivi de 10 lettres suivies de *n*).

On oppose souvent le terme à *internationalisation* (*i18n*). En effet, un logiciel peut être localisé sans être internationalisé : il aura été adapté pour une langue en particulier et non en vue de supporter toutes les langues. Les localisations successives d'un même logiciel pour différentes locales peuvent entraîner des surcoûts importants dans la gestion de projet et dans l'effort à fournir (multiplication de versions différentes). Dans le cas d'un logiciel internationalisé, il n'existe qu'une seule version du source du logiciel pour toutes les langues : il a été conçu à cet effet.

LZW

Abréviation de Lempel-Ziv-Welch, un algorithme fréquemment utilisé pour la compression de données.

Majuscule

Synonyme de haut de casse. Cf. *Casse*.

Manipulation de texte

Tout traitement informatique effectué sur du texte : il peut s'agir d'affichage pur et simple, de repérage, de coupure de lignes, de traduction, de transcodage, etc. Il convient d'opposer le traitement de texte (un logiciel d'usage restreint) à la manipulation de texte qui comprend tous les processus qui manipulent des données textuelles.

Matra

Voyelle dépendante dans les écritures *brâhmî*. C'est le nom des lettres voyelles qui suivent les lettres consonnes dans l'ordre logique. Une matra a souvent un œil complètement différent de celui représentant la même voyelle phonologique utilisée comme voyelle indépendante.

Mécanisme de sérialisation de caractères

Une *forme stockée de caractères* et sa *sérialisation en octets*. Il existe quatre mécanismes de sérialisation de caractères dans Unicode 3.0 : UTF-8, UTF-16, UTF-16BE et UTF-16LE. Depuis la publication d'Unicode 3.0, le consortium Unicode a également approuvé UTF-32BE et UTF32LE comme mécanismes de sérialisation.

Méthode d'entrée

Cf. *Méthode de saisie*.

Méthode de saisie

Toute méthode utilisée pour saisir du texte qui ne consiste pas à taper directement les caractères. C'est grâce à ces méthodes de saisie que l'on tape les textes idéophonographiques ou d'autres caractères phonétiques.

La méthode de saisie analyse les touches sur lesquelles l'utilisateur a appuyé. Dans le cas du chinois (han-tseu) et du kanji, un éditeur de méthode de saisie effectue la conversion

entre les frappes de touche et les idéogrammes (ou autres caractères), habituellement en effectuant une recherche guidée par l'utilisateur au sein d'un dictionnaire en-ligne.

Métriques de glyphe

Ensemble de propriétés qui spécifie la taille relative, la position et d'autres traits d'un glyphe.

MIME

Acronyme anglais de *Multipurpose Internet Mail Extensions*. Grâce à ce format standard, il est possible d'imbriquer ou de joindre des fichiers de type arbitraire aux messages de courrier électronique ou d'étiqueter les fichiers HTML renvoyés au fureteur.

Minuscule

Cf. *Casse*.

Monotonique

Grec moderne écrit avec l'accent de base, le *tonos*. S'oppose à *polytonique*.

Nékoudot

Signes qui indiquent les voyelles et autres modifications aux consonnes hébraïques. Cf. *Harakat*.

Nom de caractère

Nom donné à un caractère codé. Les points de code n'ont pas de nom à proprement parler.

Non affecté

Valeur de code qui est réservée pour une normalisation ultérieure ou qui ne sera jamais utilisée.

Non-caractère

Point de code réservé qui ne peut être utilisé que de façon interne et ne peut jamais être échangé avec un autre processus. Il s'agit des valeurs $U+n\text{FFFE}$ et $U+n\text{FFFF}$, où n est compris entre 0 et 10_{16} .

Normalisation

Transformation des données dans une forme standardisée, par exemple afin d'en unifier l'orthographe. Voir la section 5.7, *Normalisation*.

Normatif

Requis pour se conformer au standard Unicode. Cf. *Informatif*.

Numéro de caractère

Nombre entier positif associé à un caractère abstrait pour en faire un caractère codé. Cf. *Valeur scalaire Unicode* et *caractère codé*.

Octet de groupe

Les 8 bits les plus significatifs d'une valeur UCS-4 à 4 octets. Identifie le groupe auquel appartient un caractère.

Octet de plan

Octet qui identifie le plan auquel appartient un caractère. Également appelé octet-P.

Œil

Dessin de la lettre qui apparaît à l'impression. En typographie, l'œil reçoit l'encre, c'est l'élément imprimant. On dit également, en imitant l'américain et parfois pour éviter la polysémie du mot œil, *image de glyphe* ou simplement *glyphe*. Le pluriel d'œil au sens typographique est œils.

Ordre de rendu

Ordre dans lequel les glyphes sont affichés lors du rendu textuel.

Ordre logique

Ordre dans lequel le texte est saisi au clavier. Généralement, l'ordre logique correspond à l'ordre phonétique. Voir section 2.2, *Principes de conception d'Unicode*.

Ordre visuel

Caractères triés dans l'ordre où ils sont présentés à la lecture. À distinguer de l'*ordre logique*.

Page de code

Synonyme de *jeu de caractères codés*, désigne souvent un jeu de caractères utilisé sur ordinateur personnel. Ainsi la page de code 437 est-elle le jeu de caractères codés implicite des versions américaines du système d'exploitation DOS.

Paire de seizezets d'indirection

Représentation d'un caractère codé sous UTF-16 correspondant à un seul caractère abstrait qui consiste en une suite de deux unités de stockage où la première valeur de cette paire est le seizezet supérieur et la seconde le seizezet inférieur. Cf. la définition D27 à la section 3.7, *Seizezets d'indirection*.

Permutation symétrique

Cf. *Propriété miroir*.

Petit-boutien

Architecture informatique où l'octet le moins significatif des valeurs numériques multi-octets est stocké en premier. L'image est empruntée aux « Voyages de Gulliver » de Jonathan Swift, dans lequel les Gros-Boutiens, partisans de la théorie selon laquelle il faut casser un œuf par le gros bout, s'opposent aux Petits-Boutiens qui, eux, soutiennent qu'il faut, au contraire, casser un œuf par le petit bout. Cf. *Gros-boutien*.

Phonème

Plus petite unité du langage parlé, dont la fonction est de constituer les signifiants et de les distinguer entre eux. Les sons interchangeables dans une langue sans changer le sens d'un énoncé ne forment qu'un seul phonème. Ainsi la consonne initiale du mot français *rail* est-elle transcrite par le seul phonème /r/ (notation phonologique), mais suivant la prononciation du locuteur, elle sera notée (notation phonétique) [r] ou [R]. Ces deux r ne forment qu'un seul phonème en français alors que d'autres langues les distingueront et il s'agira donc, dans ce contexte, de deux phonèmes distincts.

Pinyin

Romanisation du chinois normalisée par la Chine, elle est fondée sur la prononciation pékinoise (mandarin du Nord).

Plan

Terme qui correspond à une subdivision du groupe, elle se compose de 256 x 256 cellules. Il s'agit donc d'un intervalle contigu de 65.536 points de code. On numérote les plans de 0 à 16 en Unicode et de 0 à 10 pour l'ISO/CEI 10646 (en hexadécimal donc). Le plan 0 correspond à U+0000..U+FFFF, le plan 1 à U+10000..U+1FFFF et le plan 16 (10₁₆) à U+100000..U+10FFFF. Voir *Plan multilingue de base* et *Plans complémentaires*.

Plan multilingue de base

Plan défini par la norme internationale ISO/CEI 10646 qui correspond aux valeurs de code 0000 à FFFF. Également appelé plan zéro.

Plans complémentaires

Dans l'ISO/CEI 10646, plans admettant des caractères qui n'ont pas été affectés au plan multilingue de base.

Pleine chasse

Caractère des jeux de caractères d'Extrême-Orient dont le glyphe occupe une cellule de rendu au complet. Dans les jeux de caractères historiques, les caractères à chasse pleine sont habituellement codés sur deux ou trois octets. Le terme japonais pour ces caractères à pleine chasse est *zenkaku*.

PMB

Abréviation de *Plan multilingue de base*.

Point de code

(1) Position dans une table de codage utilisée pour coder un caractère. On dira plus simplement *numéro de caractère*. (2) Synonyme de *valeur scalaire Unicode*.

Point de code du PMB

Un point de code situé entre U+0000 et U+FFFF.

Point de code complémentaire

Un point de code situé entre U+10000 et U+10FFFF.

Points

(1) Voyelles sans chasse et autres signes de l'hébreu écrit. (2) Une unité de mesures en typographie.

Points-voyelles

Signes diacritiques ajoutés aux signes alphabétiques pour représenter les voyelles de certaines langues sémitiques. Cf. *Points (1)*, *Voyellement*, *Harakat et Nékoudot*.

Police

Assortiment complet de glyphes pour la représentation des données de caractères. Une police possède un ensemble de propriétés (par exemple, un corps, une graisse, une posture, des empattements ou non). L'affectation de certaines valeurs à ces paramètres permet de générer un ensemble de glyphes affichables.

Polytonique

Ancien grec écrit à l'aide de plusieurs accents distincts et d'autres diacritiques. Cf. *Monotonique*.

Profil local

Ensemble de paramètres pour les objets sensibles à un profil culturel donné ou à une langue. Par exemple, le format des dates, le calendrier, le symbole monétaire en cours, la langue et le jeu de caractères local. Synonyme (anglicisme) *la locale*.

Propriété alphabétique

Propriété informative des unités primaires des alphabets ou des syllabaires. Cf. la section 4.9, *Lettres et autres propriétés utiles*.

Propriété de directionnalité

Propriété de chaque caractère graphique qui détermine son ordre horizontal tel que précisé dans section 3.12, *Comportement bidirectionnel*. Cf. la définition D9 à la section 3.4, *Propriétés simples*.

Propriété de valeur numérique

Propriété des caractères utilisés pour représenter les nombres. Cf. la définition D10b à la section 3.4, *Propriétés simples*.

Propriété idéophonographique

Propriété informative des caractères idéophonographiques. Cf. la section 4.9, *Lettres et autres propriétés utiles*.

Propriété mathématique

Propriété informative des caractères utilisés comme opérateurs de formules mathématiques.

Propriété miroir

Propriété des caractères dont l'œil est réfléchi horizontalement, et qui représente donc une image miroir dans les textes disposés de droite à gauche par rapport à leur œil habituel rencontré dans les textes de gauche à droite. Cf. la définition D10 dans la section 3.4, *Propriétés simples*. Cf. également 4.7, *Caractères miroirs* – normatif. On peut également parler de propriété spéculaire.

Propriétés de caractère

Ensemble de noms de propriété et de valeurs de propriété associé à un caractère particulier. Cf. 4, *Propriétés de caractères*.

Protocole de niveau supérieur

Tout accord sur l'interprétation des caractères Unicode qui dépasse la portée de cette norme. Cet accord ne doit pas nécessairement être annoncé de manière formelle et explicite dans les données échangées, il peut être implicite à son utilisation dans un contexte donné. Cf. définition D8 à la section 3.3, *Caractères et représentations codées*.

Radical

Un groupe de traits dans l'écriture han que l'on traite comme une unité à des fins de tri, de recherche et de classement. On parle aussi parfois de clé. Un caractère idéophonographique (ou idéographique) peut comprendre plus d'un élément que l'on pourra considérer comme radical. Cependant chaque caractère sera constitué d'un seul élément, la clé principale, qui sera utilisé lors du tri. Le radical principal fournit souvent une indication quant au sens général du caractère alors que les autres radicaux pourront indiquer la prononciation. Cf. également *caractères han*.

Rangée

Pour l'ISO/CEI 10646, subdivision d'un plan composée de 256 cellules.

Rendu

(1) Processus lié à la sélection et à la disposition de glyphes afin de représenter des données textuelles. (2) Fait de rendre visible des glyphes sur une unité de visualisation.

Rendu bidirectionnel

Le processus ou le résultat du mélange sur une même ligne de textes orientés de droite à gauche et de gauche à droite Cf. 3.12, *Comportement bidirectionnel*.

Répertoire

Cf. *Répertoire de caractères*.

Répertoire de caractères

Ensemble des caractères faisant partie d'un jeu de caractères conventionnel, parfois qualifié d'abstrait, par exemple les 26 lettres de l'alphabet français, avec leurs formes minuscules et majuscules.

Représentation codée de caractère

Une suite ordonnée d'une ou plusieurs unités de stockage associées à un caractère abstrait dans un répertoire de caractères donné. Cf. section 3.3, *Caractères et représentations codées*.

Représentation logique

Représentation en mémoire.

Réservé

Valeur de code non attribuée et réservée aux versions ultérieures de la norme.

SACN

Abréviation de *Signe à chasse nulle (ou sans chasse)*.

Seizet

Suite de 16 éléments binaires (ou bits) considérée et traitée comme une unité. Cf. *Seizets d'indirection* et *UTF-16*.

Seizet d'indirection inférieur

Dans le codage UTF-16, unité de stockage dont la valeur se situe entre 0xDC00 et 0xDFFF. Cf. la définition D26 à la section 3.7, *Seizets d'indirection*.

Seizet d'indirection supérieur

Dans le codage UTF-16, unité de stockage dont la valeur se situe entre 0xD800 et 0xDBFF. Cf. la définition D25 dans la section 3.7, *Seizets d'indirection*.

Séquence d'échappement

Une suite d'octets utilisée pour étendre le code. Le premier octet de cette suite est *échappement* (hexa 1B).

Séquence équivalente

Cf. *Équivalent canonique*.

Sérialisation en octets

L'ordre d'une suite d'octets déterminé par l'architecture informatique d'une machine.

SGML

Acronyme anglais de Langage normalisé de balisage généralisé. Norme internationale visant à décrire un document sous sa forme logique. Le balisage se réfère à la structure du texte et définit la place de chaque partie du texte dans l'ensemble. Le balisage est descriptif, il ne s'agit pas d'instructions de traitement. Grâce à cela, les types de document décrits sont indépendants de toute plate-forme. HTML et XML sont deux langages de balisage utilisés sur la toile, il s'agit d'applications de SGML.

Le balisage permet de décrire complètement un texte de fantaisie à l'aide d'un flux de données textuelles brutes. Cf. également *HTML*, *XML* et *Texte enrichi*.

Signature Unicode

Balise implicite qui signale qu'un fichier contient du texte Unicode sous une forme codée particulière. Un *indicateur d'ordre des octets* (IOO) peut servir de signature Unicode.

Signe à chasse nulle

Un caractère combinatoire dont l'emplacement lors du rendu dépend du caractère de base. Ce caractère ne chasse habituellement pas, c'est-à-dire qu'il ne consomme pas de lui-même d'espace le long de la ligne de base visuelle. Cf. la définition D15 à la section 3.5, *Combinaison*. Cf. également *Caractère combinatoire*.

Signe avec chasse

Un caractère combinatoire qui n'est pas un signe à chasse nulle. Cf. *Signe à chasse nulle*.

Signe de cantilation

Diacritique utilisé pour indiquer de quelle manière le texte doit être psalmodié ou chanté. Ces diacritiques sont utilisés en hébreu.

Signe de ton

Diacritique ou signe sans chasse qui représente un ton phonémique. On rencontre de nombreuses langues à tons en Extrême-Orient, en Afrique ou en Europe (p.ex., serbo-croate, suédois). Les tons sont le plus souvent écrits à l'aide de signes fonctionnellement indépendants associés à un symbole vocalique car ils s'appuient toujours sur une voyelle (le noyau syllabique). Toutefois, certaines écritures comme le thaï placent les tons sur les symboles consonnes; le chinois s'écrit sans signe de ton (sauf lorsqu'on l'écrit phonémiquement).

Signe de voyelle

Dans de nombreuses écritures, signe utilisé pour représenter une voyelle ou la qualité d'une voyelle.

Sous-ensemble

L'ISO/CEI 10646 définit des sous-ensembles de caractères graphiques codés utilisés lors d'un échange par des dispositifs de réception et d'émission.

Deux types de sous-ensembles peuvent être définis : les sous-ensembles limités et les sous-ensembles sélectionnés. Un sous-ensemble adopté peut comprendre l'un des deux ou une combinaison de ces deux types.

Unicode ne définit pas de sous-ensemble, les applications qui s'y conforment doivent traiter correctement tous les caractères du standard.

Sous-ensemble limité

Sous-ensemble défini par une liste de caractères (noms ou numéros).

Sous-ensemble sélectionné

Sous-ensemble défini par une liste de collections, certaines de ces collections pouvant être *ouvertes* (c'est-à-dire susceptible de contenir des positions de code réservés)

Suite de caractères

Cf. les définitions D4 (*suite de caractères abstraits*) et D7 (*suite de caractères codés*) à la section 3.3, *Caractères et représentations codées*.

Suite de caractères codés

Suite ordonnée de représentations codées de caractère. Cf. la définition D7 à la section 3.3, *Caractères et représentations codées*.

Suite de caractères combinatoires

Cf. la définition D17 à la section 3.5, *Combinaison*.

Suite de caractères composés

Cf. la définition D17 à la section 3.5, *Combinaison*.

Suite de caractères composites

Cf. *Suite de caractères combinatoires*.

Suite défectueuse de caractères combinatoires

Suite de caractères combinatoires qui ne commence pas par un caractère de base. Cf. la définition D17a à la section 3.5, *Combinaison*.

Suite illégale d'unités de stockage

Cf. la définition D31 dans la section 3.8, *Transformations*.

Suite irrégulière d'unités de stockage

Cf. la définition D32 de la section 3.8, *Transformations*.

Suite mal formée d'unités de stockage

Cf. la définition D30 dans la section 3.8, *Transformations*.

Syllabaire

Ensemble de caractères dont chacun représente une syllabe, par exemple une consonne *C* suivie d'une voyelle *V*. C'est le cas des kana japonais qui sont des syllabes fondamentales (à l'exception du « n » final) ; dans le cas du syllabaire coréen, dit *hangûl*, les syllabes sont formées de jamos et peuvent prendre les formes *CV*, *CVC* et rarement *CVCC*. Il existe aussi un signe représentant une consonne initiale muette qui permet les formes *V* et *VC*.

Le cri et l'inuktitut s'écrivent également à l'aide de leur propre syllabaire. Les écritures dérivées de la *brâhmî*, comme par exemple la dévanâgarî, sont intermédiaires entre alphabet et syllabaire, et sont parfois appelées *alphasyllabaires*.

Syllabe

(1) Élément d'un syllabaire. (2) Unité d'articulation fondamentale qui se prononce en une émission de la voix.

Système d'écriture

Ensemble de règles régissant l'utilisation d'une ou plusieurs écritures pour transcrire une langue particulière. Exemples : le système d'écriture français, le système d'écriture anglais et le système d'écriture japonais.

Taquet

Butée réglable utilisée pour arrêter les retours de chariot. On parle également de *taquer* des caractères pour les positionner à l'emplacement désiré. Opération désignée sous le nom de *taquage*.

Unicode définit une série de flèches partant ou menant à un taquet (p.ex., ←).

Tatouil

Cf. *Kachidé*.

Tchandrabindou

L'*anounâssika* ou *tchandrabindou* (lune-point) est un signe diacritique placé au-dessus des voyelles hindî nasalisées. Si la voyelle dépendante est placée au-dessus de la ligne de tête, il ne reste plus alors de place que pour le *bindou* (le point).

T_EX

Langage informatique conçu pour la composition mathématique et d'autres ouvrages techniques. Selon son inventeur D. Knuth, T_EX se prononce comme *teck*, car le X final représente un khi grec et non un iks.

Texte brut

Texte informatique qui ne comprend que des suites d'unités de stockage d'une norme donnée sans contenir d'autres informations de formatage ou de structure. On utilise fréquemment l'échange de texte brut entre ordinateurs qui ne partagent pas un même protocole de niveau supérieur. Cf. également *Texte enrichi*.

Texte de fantaisie

Cf. *Texte enrichi*.

Texte enchâssé

(1) Texte brut entouré d'information de formatage. (2) Texte recodé afin de passer à travers un canal de transmission étroit ou pour se conformer à un protocole de communication.

Texte enrichi

Également connu sous le nom de *texte de fantaisie*. Résultat de l'adjonction d'information supplémentaire au texte brut. Exemples d'information supplémentaire : la police, la couleur, du formatage, des annotations phonétiques, du texte interlinéaire, etc. Le standard Unicode, pas plus que l'ISO/CEI 10646, n'aborde la représentation des textes enrichis. On s'attend à ce que des applications mettent en œuvre des formes propriétaires de texte de fantaisie. Certaines formes publiques de textes de fantaisie existent (par exemple, ODA, HTML et SGML). Quand un texte enrichi est entièrement dépouillé sauf de son contenu essentiel, il ne reste plus que le texte brut.

Texte formaté

Cf. *Texte de fantaisie*.

Texteur

Cf. *Traitement de texte*

Traitement de texte

Logiciel qui permet de créer et de manipuler des textes, il est l'équivalent informatique de la machine à écrire. Également appelé *texteur*.

Tranche

Il y a autant de tranches dans les écritures indiennes qu'il y a de voyelles. Chaque tranche est représentée par une voyelle indépendante ou d'une ou plusieurs consonnes qui peuvent former une *conjointe*. Il s'agit *grosso modo* d'une syllabe.

Transcodage

Transformation d'un signifiant d'un système de codage à un autre, idéalement sans modification du signifié. Exemples : transcoder du texte de CP437 à ISO-LATIN-1, transcoder du son de -law à m-law, transcoder un signal vidéo de NTSC à SECAM.

Translittération

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. Exemples : la translittération des caractères cyrilliques russes en caractères latins français, des caractères arabes en norme ISO. Cependant on parle de la *transcription* du chinois en pinyin, en système E.F.E.O (École française d'Extrême-Orient) ou en Wade. Cf. *Transcription*.

Transcription

Système de représentation de la parole qui se préoccupe de rendre compte les sons réellement prononcés et non les lettres utilisées dans le système d'écriture de cette langue pour représenter ces sons. Cf. *Translittération*.

Tréma

Signe constitué de deux points juxtaposés que l'on met sur les voyelles. Unicode ne distingue pas au niveau typographique le tréma de l'*umlaut*, ce qui ne signifie évidemment pas que les fonctions du tréma et de l'*umlaut* soient les mêmes. Cf. *Umlaut*.

Tri alphabétique

Cf. *Classement*.

Triangulation

Cf. *Conversion par pivot*.

UCS-2

Forme de stockage de l'ISO/CEI 10646, Jeu universel de caractères codé sur deux octets. Cf. annexe C, *Comparaison entre ISO/CEI 10646 et Unicode*.

UCS-4

Forme de stockage de l'ISO/CEI 10646, Jeu universel de caractères codé sur quatre octets. Cf. annexe C, *Comparaison entre ISO/CEI 10646 et Unicode*.

Umlaut

Tréma allemand, voir Köpfe. Cf. également *Tréma*.

Unicaméral

Écriture sans distinction de casse. Terme utilisé le plus souvent dans le contexte d'alphabets européens.

Unification

Le processus qui consiste à identifier les caractères communs parmi des systèmes d'écriture différents.

Unification han

Processus de sélection de sous-ensembles de *caractères han* utilisés par le chinois, le japonais, le coréen et le vietnamien et, qui pour chacun des caractères jugés communs lui attribue un seul élément de code. Les critères de sélection de ces sous-ensembles sont basés sur des similarités de représentation graphique abstraite indépendants de tout signifié ou de variations dans les traits utilisés pour représenter ces caractères. Cf. Annexe H, *Unification han*,

Unité de stockage

Entier d'une largeur variable (exemples : octet ou seize) qui sert d'unité de base pour l'expression des points de code dans la mémoire d'un ordinateur. Cf. *Forme stockée des caractères*.

Usage privé

Les valeurs scalaires Unicode (points de code) allant de U+E000 à U+F8FF, de U+F0000 à U+FFFFD et de U+100000 à U+10FFFFD sont réservés à l'usage privé. Cf. la définition D12 à la section 3.4, *Propriétés simples*. Désigne les valeurs de code et les zones de la norme dont l'interprétation n'est pas précisée par la norme et dont l'utilisation peut être déterminée de gré à gré entre des utilisateurs qui coopèrent.

UTF

Abréviation anglaise de *Unicode* (ou *UCS*) *Transformation format*. Cf. *Format transformé d'Unicode* (ou du *JUC*).

UTF-2

Nom désuet d'*UTF-8*.

UTF-7

Format transformé d'Unicode (ou du *JUC*), forme de stockage sur 7 bits, spécifié dans le RFC 2152.

UTF-8

Format transformé d'Unicode (ou du *JUC*), forme de stockage sur 8 bits. UTF-8 sérialise une valeur scalaire Unicode (ou point de code) en une suite de 1 à 4 octets, tel qu'illustré

au tableau 3-1, *Distribution UTF-8 des bits*. Cf. la définition D36 à la section 3.8, *Transformations*.

UTF-16

Format transformé d'Unicode (ou du JUC), forme de stockage dont l'unité de codage a 16 bits. UTF-16 exprime une valeur scalaire Unicode sous la forme d'un ou deux seizes. S'il y a sérialisation, chaque seize devient deux octets, transmis sous le format petit-boutien ou gros-boutien. Cf. la définition D35 à la section 3.8, *Transformations*.

UTF-16BE

Format transformé d'Unicode (ou du JUC) qui sérialise une unité de stockage sous la forme d'un seize gros-boutien. Un seize initial correspondant à U+FEFF est interprété comme une ESPACE INSÉCABLE SANS CHASSE. Cf. la définition D33 à la section 3.8, *Transformations*.

UTF-16LE

Format transformé d'Unicode (ou du JUC) qui sérialise une unité de stockage sous la forme d'un seize petit-boutien. Un seize initial correspondant à U+FEFF est interprété comme une ESPACE INSÉCABLE SANS CHASSE. Cf. la définition D33 à la section 3.8, *Transformations*.

UTF-32

Format transformé d'Unicode (ou du JUC) qui représente une valeur scalaire Unicode sous la forme d'une valeur de 32 bits. Les unités de stockage d'UTF-32 appartiennent à l'intervalle 0x00000000..0x0010FFFF.

Valeur de code

La combinaison binaire minimale qui permet de représenter une unité de texte codé pour le traitement ou l'échange. Cf. la définition D5 à la section 3.3, *Caractères et représentations codées*. Cf. *Unité de stockage*.

Valeur de code attribuée

Valeur de code pour laquelle il existe une sémantique définie et interopérable.

Valeur scalaire

Cf. *Valeur scalaire Unicode*.

Valeur scalaire Unicode

Nombre N de 0 à 10FFFF₁₆, numéro associé au caractère pour en faire un caractère codé. Également appelé point de code.

Variante contextuelle

Élément textuel qui peut posséder une forme de présentation dépendant du contexte dans lequel cet élément est rendu. .

Variante de compatibilité

Caractère qui peut en remplacer un autre sans perte d'information autre que celle liée au formatage.

Virâma

Mot sanscrit signifiant arrêt. Nom d'un symbole utilisé dans les écritures *brâhmî* pour indiquer que : (1) la voyelle implicite ne s'applique pas, fonction appelée dévoyellement ; (2) le début d'une nouvelle syllabe ; (3) le début d'un nouveau mot.

Certaines de ces utilisations sont mutuellement incompatibles. En Unicode, un virâma est utilisé pour dévoyeller une syllabe, dans ce sens, *halant* est plus approprié. Il ne reste donc plus que la consonne dite, dans ce cas, morte. Celle-ci se joint normalement à la consonne de base suivante pour former une forme conjointe ou composée. Quand le virâma n'est pas résorbé en une forme conjointe, il est alors traité comme une espace ou non, selon l'écriture employée. En dévanâgarî, c'est un signe à chasse nulle, mais en tamoul il s'agit d'une espace.

Visarga

Terme sanscrit employé pour désigner une prononciation particulière de l's final, de nature telle que l'arrêt brusque du courant d'air qui accompagne l'émission de la voyelle précédente donne à l'oreille l'impression d'une aspirée sourde. Dénoté en dévanâgarî à l'aide de U+0903 SYMBOLE DÉVANÂGARÎ VISARGA ः, ce signe se transcrit généralement *ḥ*.

Voyelle dépendante

Symbole ou signe qui représente une voyelle et qui est associé ou combiné à un autre symbole, habituellement une consonne. Dans les systèmes d'écriture basés sur le *brâhmî*, l'arabe ou l'hébreu, les voyelles sont normalement représentées par des signes de voyelles dépendantes. Cf. *Matra*.

Voyelle implicite

Dans les systèmes d'écriture dérivés de l'écriture *brâhmî*, une voyelle implicite est habituellement associée aux symboles consonnes, sauf indication contraire. La valeur de cette voyelle diffère selon les langues écrites avec ces systèmes d'écriture. La voyelle implicite peut être supprimée soit par l'utilisation explicite d'une autre voyelle ou d'un *virâma* créant de la sorte une *consonne morte* (dévoyellée).

Voyelle indépendante

Dans les écritures d'origine *brâhmî*, on dessine certaines voyelles à l'aide de lettres indépendantes détachées des autres lettres. C'est souvent le cas quand un mot commence par une voyelle ou n'est constitué que d'une voyelle.

Voyellement

Signes placés au-dessus, en dessous ou à l'intérieur d'une consonne qui représentent des voyelles ou d'autres aspects de la prononciation. Une caractéristique des écritures du Moyen-Orient. Cf. *Points-voyelles*.

wchar_t

ISO C définit un type *wide character* (caractère large), habituellement codé sur 16 ou 32 bits. La norme précise que *wchar_t* est un type entier et que le jeu de caractères source du langage C y correspond par une simple extension (extension signée ou à l'aide de zéros).

XML

Abréviation anglaise de Langage de balisage extensible. Un sous-ensemble de SGML qui constitue un métalangage de balisage de texte particulier pour l'échange de données structurées. L'ISO/CEI 10646 est le jeu de caractères de référence pour le contenu XML. Cf. également *SGML* et *Texte de fantaisie*. XML est une marque déposée du W3C.

Zenkaku

Cf. *Pleine chasse*.