Chapitre 7

Ponctuation

L'apparence et l'utilisation des signes de ponctuation varient d'une écriture et d'une langue à l'autre ; Cependant ces signes de ponctuation ont une fonction commune : ils séparent les unités de texte – comme les propositions et les phrases – en éclaircissent de la sorte le sens. Les signes de ponctuation ne s'utilisent pas uniquement en prose : on les emploie notamment dans les formules scientifiques et mathématiques. Le standard Unicode nomme ces signes de ponctuation des caractères de ponctuation.

Unicode agence les caractères en groupes connexes appelés blocs. En règle générale, un bloc regroupe des caractères d'une seule écriture. Il arrive souvent qu'un bloc de caractères représente intégralement une écriture. Il existe, néanmoins, d'importantes exceptions plus particulièrement dans le domaine de la ponctuation. En effet, les caractères de ponctuation se retrouvent parfois dans des blocs spécifiques à une écriture mais surtout dans des blocs fort différents parmi lesquels le latin de base, le supplément latin-1, la ponctuation générale et l'ensemble ponctuation et symboles CJC.

On ne code les caractères de ponctuation (U+002C , VIRGULE ou U+2022 • PUCE par exemple) qu'une seule fois et non à chaque fois qu'on les retrouve dans un système d'écriture (et donc dans un bloc) particulier. Ces signes de ponctuation universels peuvent s'employer dans n'importe quelle écriture ou combinaison d'écriture. À l'inverse, les caractères de ponctuation codés dans un bloc associé à une écriture particulière (par exemple U+058A ֊ TRAIT D'UNION ARMÉNIEN ou U+060C ، VIRGULE ARABE) sont principalement destinés à être utilisés avec cette écriture. Leur fonction est singulière, ils possèdent une directionalité particulière et diffèrent de leurs homologues universels par leur apparence ou leur utilisation.

L'emploi et l'interprétation des caractères de ponctuation peut fortement dépendre du contexte. Ainsi, U+002E . POINT peut indiquer la fin d'une phrase, une abréviation, un séparateur numéral, etc.

L'aspect des caractères de ponctuation varie en fonction du style de la police choisie ainsi que des caractères adjacents. Quelquefois, dans le contexte d'une écriture particulière, on préfère un style de glyphe précis. C'est ainsi que U+002E . POINT doit adopter une forme carrée en arménien alors qu'il est habituellement[1] circulaire en latin. Pour les textes mixtes latin-arménien, il se peut qu'il faille utiliser deux polices (ou une police qui admet des variations glyphiques contextuelles) pour rendre fidèlement le caractère.

Contrairement à ce que laisse croire Unicode, les usages en matière de ponctuation, espaces, etc. ne sont pas uniformes pour les diverses langues d'écriture latine. Ceux-ci peuvent varier d'un pays à l'autre, voire, dans un même pays, d'un auteur à l'autre souvent en fonction du matériel utilisé selon qu'il est importé ou non !

Dans un contexte bidirectionnel (voir la *Section 3.12*, *Comportement bidirectionnel*), la ponctuation commune est dénuée de directionalité implicite, elle adopte celle calculée par l'algorithme bidirectionnel d'Unicode. Quand l'œil du signe de ponctuation n'est pas bilatéralement symétrique, on en utilise l'image spéculaire (ou miroir) dans un passage droite-à-gauche (voir la *Section 4.7*, *Caractères miroirs – normatif*).De nombreux caractères de ponctuation possèdent des glyphes spéciaux pour l'écriture verticale.

[1] Le point n'est pas forcément circulaire en latin (en Helvetica, par exemple).

Certains caractères appartenant aux blocs décrits dans ce chapitre ne sont pas des caractères de ponctuation graphiques, ils modifient cependant le fonctionnement des algorithmes de mise en page. Veuillez vous référer à la *Section 14.2*, *Commandes de disposition*, pour une description de ces caractères.

7.1 Ponctuation générale

Ponctuation : U+0020 – U+00BF

Normes. Le standard Unicode a adapté la norme à 7 bits ASCII (ISO 646) en en conservant la sémantique et les numéros de caractère. Le contenu et la disposition de la norme ASCII sont toutefois loin d'être optimaux dans un espace à 16 (ou 21) bits. Unicode les a cependant conservés intacts en raison de la prévalence de l'ASCII. L'ASCII (ANSI X3.4) est identique à la norme ISO/CEI 646 :1991-IRV.

Caractères graphiques ASCII . Certains caractères ASCII qui ne représentent pas de lettre sont victimes d'une polysémie préjudiciable causée par le manque de codes disponibles dans un jeu à 7 bits. Cette section examine certaines conséquences liées à ce problème (voir ci-dessous *Codage des caractères polysémiques*, *Ponctuation générale* et *Utilisation des guillemets en fonction de la langue*). En raison de son omniprésence dans les normes et logiciels actuels, on a conservé la disposition quelque peu anarchique des signes mathématiques et de ponctuation ASCII ; ils se trouvent de la sorte isolés du gros de la ponctuation, des signes et des symboles Unicode (codé à partir de U+2000).

Variantes typographiques. Les valeurs de code de l'intervalle ASCII sont aujourd'hui bien établies et utilisées par un grand nombre de mises en œuvre. Le standard Unicode ne décrit donc que sommairement l'apparence typographique des glyphes correspondants. La valeur U+0024 ($) dérivée de la valeur 24 ASCII a, par exemple, comme valeur sémantique le *dollar*, Unicode ne précise pas si celui-ci doit être représenté à l'aide d'une barre verticale ou de deux. Le numéro de caractère Unicode U+0024 fait référence au dollar en tant qu'entité sémantique et non à un aspect précis de celui-ci.

De même, pour les nombres écrits en capitales ou en bas de casse dont la position des chiffres peut varier par rapport à la ligne de base, le séparateur des milliers ou des unités peut être représenté à l'aide d'un point légèrement surélevé. Il n'est pas souhaitable d'encourager des variations dans la représentation sous-jacente du texte entre ceux qui codent des chiffres modernes et ceux qui codent des chiffres à l'ancienne, c'est pourquoi Unicode considère ce point surélevé comme une simple variante glyphique du *point* U+002E « . ». Le standard Unicode représente les autres caractères aux œils multiples à l'aide de l'œil le plus dépouillé ou le plus fréquent ; les logiciels de rendu peuvent choisir un autre œil pour afficher ce caractère.

Codage des caractères polysémiques. Certains caractères connaissent divers emplois, soit à cause d'une ambiguïté dans la norme d'origine ou par réinterprétation successive d'un ensemble restreint de codes. Ainsi, ANSI X3.4 définit le code 27 hexa comme une *apostrophe* (*guillemet apostrophe, accent aigu*) alors qu'il définit 2D hexa comme un *signe-moins-trait-d'union*. En règle générale, le standard Unicode conserve aux valeurs de code communes leur sémantique, sans rien y ajouter ou soustraire. Le standard Unicode fournit par ailleurs des valeurs de codes supplémentaires *univoques* pour la plupart des sens de ces caractères ASCII ; les entrées de ces caractères univoques dans le tableau de codes renvoient à leur caractère ASCII polysémique. Pour une liste complète des caractères d'espacement et des tirets dans le standard Unicode, veuillez vous reporter ci-dessous à *Ponctuation générale*.

Pour des raisons de fait, U+0027 est un caractère particulièrement surchargé. En ASCII, il représente un signe de ponctuation (un guillemet-apostrophe gauche ou droit, une apostrophe proprement dite, une ligne verticale ou le signe prime) ou encore un signe modificateur (un accent aigu ou un diacritique apostrophe). Il est permis de couper les mots au signe de ponctuation alors qu'on considère habituellement que les signes modificateurs font partie d'un mot (ils sont insécables).

Unicode recommande d'utiliser U+2019 pour désigner l'apostrophe, toutefois U+0027 est bien plus fréquent sur les claviers. Les logiciels de saisie modernes remplacent donc fréquemment U+0027 par le caractère adéquat. Pour ces systèmes, un U+0027 est toujours représenté dans le flux de données par un trait vertical droit et jamais par une apostrophe courbée[2] ni par un guillemet-apostrophe fermant. Pour plus d'informations, veuillez lire la sous-section *Apostrophes* ci-dessous.

Sémantiques de la ponctuation appariée. Dans le contexte des textes bidirectionnels ou verticaux, il est impératif d'interpréter sémantiquement plutôt que graphiquement les signes de ponctuation appariés tels que les parenthèses (U+0028, U+0029), les crochets (U+005B, U+005D) ou les accolades (U+007B, U+007D). Ceci signifie que *ces caractères ont une sémantique constante mais des glyphes variables selon le flux directionnel rendu par un logiciel particulier*. Ce logiciel doit s'assurer que les glyphes rendus sont les bons. Lors de l'interprétation sémantique (plutôt que graphique), tout caractère dont le nom est qualifié par un « GAUCHE » désigne un caractère *ouvrant* alors que le qualificatif « DROITE » indique un caractère *fermant*. Ainsi, U+0028 (PARENTHÈSE GAUCHE et U+0029) PARENTHÈSE DROITE définissent, respectivement, une parenthèse ouvrante et fermante dans tous les contextes, y compris ceux bidirectionnels ou verticaux. Dans un passage droite-à-gauche, U+0028 est rendu à l'aide d'une «) ». Dans un passage gauche-à-droite, on rend le même caractère sous la forme d'une « (».

Tilde. U+007E ~ TILDE s'emploie en tant que clone à chasse du tilde diacritique[3] (voir *Clones à chasse des diacritiques* à la *Section 8.1, Latin*) ou, le plus souvent, en tant que tilde centré sur la ligne; il ressemble alors à U+223C ~ OPÉRATEUR TILDE. On emploie souvent le tilde pour indiquer une valeur approximative ou, dans un dictionnaire, pour répéter le terme défini dans le corps de sa définition. Bien que le rendu d'U+007E ~ TILDE dépende du contexte, les polices modernes le centrent habituellement sur la ligne, comme dans le tableau des codes.

> Remarque : en fonction de la locale et du profil de l'utilisateur, les caractères U+002C, U+002E, U+060C, U+066B et U+066C (et d'autres sans doutes) peuvent chacun servir de séparateur de chiffres.

Ponctuation générale : U+2000 – U+206F

La ponctuation générale comprend les signes de ponctuation et les éléments ressemblant à des caractères qui permettent de modifier la composition. Certains caractères de ponctuation peuvent s'utiliser dans de nombreuses écritures. On retrouve également beaucoup de ces caractères dans le bloc latin de base (ASCII) et dans le bloc supplémentaire latin-1.

Il est courant que les normes actuelles codifient des caractères de ponctuation génériques plutôt que les caractères correspondants plus précis utilisés en imprimerie. On peut citer comme exemples les guillemets anglais simples (apostrophes) et doubles, le point, le tiret et l'espace. Unicode codifie également ces caractères génériques tout en incluant

[2] Dans de nombreuses polices (Futura, Optima…) cette apostrophe n'est pas courbée. U+2019 est une apostrophe typographique (et, dans ce cas, directionnelle), alors que U+0027 est une apostrophe dactylographique (et, dans ce cas, non directionnelle). Bien que l'apostrophe courbée et le guillemet-apostrophe fermant soient sémantiquement distincts,dans la pratique ils sont totalement homoglyphes.

[3] Il est toutefois préférable d'utiliser à cette fin U+02DC ˜ PETIT TILDE.

indépendamment les caractères univoques : les différentes formes de guillemet, le point numéral, le tiret sur cadratin, le tiret sur demi-cadratin, le moins, le trait d'union, le cadratin, le demi-cadratin, l'espace fine, l'espace sans chasse et ainsi de suite. Mais avec une vision anglo-américaine que nous corrigerons ici ou là !

Unicode place la ponctuation utilisée essentiellement avec une écriture particulière dans le bloc correspondant à cette écriture. C'est le cas du U+061B ؛ POINT-VIRGULE ARABE (dans le bloc arabe) ou de la ponctuation utilisée avec les idéogrammes (dans le bloc *Symboles et ponctuations CJC*).

Espacement

Du temps du plomb, les espacements (notamment entre les mots) étaient obtenu en insérant entre ceux-ci un ou plusieurs caractères sans relief (ne laissant donc aucune trace imprimée, aucun glyphe !). Ces caractères sont connus en typographie française sous le nom d'espaces (au féminin). Nous dirons donc « une espace » pour parler du caractère ; nous parlerons aussi, par abus de langage mais par tradition, d'« une espace » pour le glyphe (donc l'espacement ou le blanc) correspondant. Ces espaces étaient essentiellement de deux types : celles à chasse fixe (par exemple un cadratin, un demi-cadratin, etc.) et les cadrats de valeurs diverses qui permettaient de justifier les lignes en mettant ces espaces de tailles variées entre les mots. Hélas, le principe d'Unicode de ne coder que des caractères et non des glyphes est pris ici en défaut car la quasi-totalité des espaces proposées dans la rangée U+2000 sont des glyphes (de chasses diverses) du caractère (entité abstraite) espace.

Le caractère d'espacement le plus fréquemment utilisé est l'espace-mot U+0020 ESPACE. On utilise aussi souvent sa contrepartie[4] insécable, U+00A0 ESPACE INSÉCABLE. Ces deux caractères sont indetiques, mais adoptent un comportement différent lors de la coupure en fin de ligne. Dans le contexte de la composition bidirectionnelle, U+00A0 ESPACE INSÉCABLE se comporte comme un séparateur numéral. (Voir la *Section 3.12*, *Comportement bidirectionnel*, pour un examen détaillé de l'algorithme bidirectionnel d'Unicode.) On utilise habituellement U+3000 ESPACE IDÉOGRAPHIQUE pour les textes idéographiques car sa chasse correspond à celle des idéogrammes.

La principale différence entre les autres espaces est leur chasse. Les caractères U+2000…U+2006 sont les espaces typographiques habituelles définies par rapport au cadratin. U+2007 ESPACE TABULAIRE, également appelée *espace-nombre*, a une largeur égale à celle d'un chiffre dans les tableaux. U+2009 ESPACE FINE et U+200A ESPACE ULTRAFINE sont des espaces de largeur successivement plus étroite employées, d'une part, en typographie française devant les signes **; ? !** et les appels de note et, d'autre part, dans la justification. Les espaces à chasse fixe (U+2000..U+200A) sont héritées de la typographie conventionnelle (au plomb). Le crénage et la justification en typographie numérique n'utilisent pas ces caractères. Toutefois, lorsqu'on les utilise, comme c'est le cas en algèbre, leur chasse dépend généralement de la police utilisée et ils conservent cette chasse pendant la justification[5]. U+2009 ESPACE FINE constitue une exception car sa chasse est parfois ajustée[6].

---

[4] Ce caractère n'existait pas du temps de la composition manuelle ou mécanique (Linotype) puisque le compositeur savait où il fallait couper et n'aurait donc pas saisi une espace en fin de ligne ni rejeté un point virgule en début de ligne ! Ce caractère a été imposé par les systèmes permettant la saisie au kilomètre (photocomposeuses ou PAO) pour indiquer que cette espace est indispensable et ne peut pas être remplacée par une coupure en fin de ligne. Voir la section 14.2.

[5] En d'autres mots, on n'augmente pas l'espacement là où apparaissent ces espaces.

[6] Cette notion de fine est typique du manque de cohérence dans la terminologie. Pour les Anglo-américains, une fine vaut souvent 1/5 ou 1/6 de cadratin. En France elle peut aussi valoir un quart de cadratin (c'est la valeur utilisée en linotypie) et de plus en plus les logiciels de formatage considèrent que cette fine est de chasse légèrement variable !

Veuillez vous reporter à la sous-section *Coupure de mots et de lignes* à la *Section 14.2*, *Commandes de disposition* pour une description des caractères d'espacement et de leur comportement lors de la coupure de mots et de lignes.

On retrouve des caractères d'espacements dans d'autres blocs Unicode. Le tableau 7-1 reprend la liste des espaces.

Tableau 7-1. Caractères d'espacement Unicode

U+0020	ESPACE
U+00A0	ESPACE INSÉCABLE
U+2000	DEMI-CADRATIN
U+2001	CADRATIN
U+2002[7]	ESPACE DEMI-CADRATIN
U+2003	ESPACE CADRATIN
U+2004	TIERS DE CADRATIN
U+2005	QUART DE CADRATIN
U+2006	SIXIÈME DE CADRATIN
U+2007	ESPACE TABULAIRE
U+2008	ESPACE PONCTUATION
U+2009	ESPACE FINE
U+200A	ESPACE ULTRA FINE
U+200B	ESPACE SANS CHASSE
U+202F	ESPACE INSÉCABLE ÉTROITE
U+3000	ESPACE IDÉOGRAPHIQUE

On remarque donc que, dans ce tableau, il y a certains caractères définis par rapport au cadratin et d'autres (notamment U+2007..U+200A) qui sont plus subjectifs (et culturels). On retrouve ici, par ailleurs, le problème de codage des caractères polysémiques (voir ci-dessus) : à part ces espaces U+0020 et U+00A0, toutes les autres sont insécables et d'utilisation contextuelle bien définie. De nombreux systèmes savent utiliser le même mode de saisie (c'est-à-dire la barre d'espacement) à la fois pour l'espace-mot et pour les espaces fixes (par exemple avant un point virgule). Ce que l'on tape « Premier prix : Pierre ; » étant interprété comme « PremierU+0020prixU+00A0:U+0020Pierre+U2009; ».

La *Section 14.2*, *Commandes de disposition*, décrit U+200B ESPACE SANS CHASSE, qui joue un peu — pour les espacements — le rôle du zéro en arithmétique, ainsi que plusieurs simili-espaces sans chasse aux propriétés particulières.

Tirets et traits

Tout comme pour les espaces, à la notion abstraite de tiret, Unicode a ajouté les glyphes typographiques usuels et préconise plutôt une terminologie américaine.

En raison de sa prévalence dans les jeux de caractères préexistants, U+002D - TRAIT D'UNION-SIGNE MOINS est le signe le plus fréquemment utilisé pour représenter le trait d'union. Il est ambigu et il est, en règle générale, rendu avec une chasse intermédiaire. U+2010 - TRAIT D'UNION représente le trait d'union d'un mot comme « chef-d'œuvre » (signe que les typographes utilisaient aussi pour « diviser » les mots en fin de ligne, d'où son nom

[7] Les espaces sur demi-cadratin et cadratin sont rigoureusement identiques aux demi-cadratin et cadratin. Ils ne doivent leur présence dans Unicode qu'à la suite d'une erreur des normalisateurs.

« division ». Sa chasse est étroite. Lors de la composition, il est préférable d'utiliser U+2010 ‐ TRAIT D'UNION et non le caractère ambigu U+002D - TRAIT D'UNION-SIGNE MOINS. Unicode reprend le caractère U+2011 ‑ TRAIT D'UNION INSÉCABLE à des fins de compatibilité avec les normes actuelles (en l'occurrence le XCCS, jeu de caractères normalisé de Xerox, qui a souvent servi de modèle au début d'Unicode). Il possède la même valeur sémantique que U+2010 ‐ TRAIT D'UNION, mais il ne peut être coupé en fin de ligne.

U+2012 ‒ TIRET NUMÉRIQUE a été inclus pour des raisons de compatibilité : il a le même sens (ambigu) que U+002D - TRAIT D'UNION-SIGNE MOINS, mais sa chasse est celle des chiffres (s'ils sont de chasse fixe). U+2013 – TIRET DEMI-CADRATIN s'emploie dans la notation des intervalles tels[8] que 1914–1918. Il faut le distinguer de l'opérateur arithmétique U+2212 − SIGNE MOINS, même si les typographes ont pris l'habitude de confondre les deux. Pour des raisons de compatibilité générale lors de l'interprétation des formules, comme « $x = a - b$ », il faut interpréter les caractères U+002D - TRAIT D'UNION-SIGNE MOINS, U+2212 − SIGNE MOINS et U+2012 ‒ TIRET NUMÉRIQUE comme étant le *signe moins*.

On emploie U+2014 — TIRET CADRATIN pour représenter une incise — comme celle-ci — au sein d'une phrase[9]. On le tape souvent à la machine sous la forme de deux traits d'union. Jadis, en typographie mathématique, on l'employait à la place d'un *signe moins binaire*. U+2015 ― BARRE HORIZONTALE s'utilise par certains typographes, notamment francophones, pour indiquer un changement d'interlocuteur dans des dialogues.

On trouve des traits d'union et des tirets dans d'autres blocs de caractères Unicode. Le tableau 7-2 reprend une liste des tirets Unicode. Pour un examen détaillé relatif aux tirets et aux traits d'union, veuillez vous reporter au rapport technique Unicode n° 14, *Line Breaking Properties*, présent sur le disque optique ou sur le site Internet du consortium Unicode (version tenue à jour).

Tableau 7-2. Tirets Unicode

U+002D	-	TRAIT D'UNION-SIGNE MOINS
U+007E	~	TILDE
U+00AD	‐	TRAIT D'UNION VIRTUEL
U+058A	֊	TRAIT D'UNION ARMÉNIEN
U+1806	᠆	TRAIT D'UNION VIRTUEL MONGOL TODO
U+2010	‐	TRAIT D'UNION
U+2011	‑	TRAIT D'UNION INSÉCABLE
U+2012	‒	TIRET NUMÉRIQUE
U+2013	–	TIRET DEMI-CADRATIN
U+2014	—	TIRET CADRATIN
U+2015	―	BARRE HORIZONTALE
U+207B	⁻	EXPOSANT SIGNE MOINS
U+208B	₋	INDICE SIGNE MOINS
U+2212	−	SIGNE MOINS
U+301C	〜	TRAIT D'UNION EN ESSE
U+3030	〰	TRAIT D'UNION ONDULÉ

[8] En français, on emploie la division normale étroite. Sauf, évidemment, ceux qui copient le Chicago Manual of Style.

[9] En fait cet emploi d'un tiret-cadratin a tendance à se perdre aujourd'hui au profit du tiret demi-cadratin, notamment dans la presse mais pas dans les ouvrages de qualité. C'est peut-être ce qui explique que le texte anglais fasse allusion à deux traits d'union, alors que souvent ce long tiret est saisi avec trois traits d'unions !

Utilisation des guillemets en fonction de la langue

U+0022 " GUILLEMET ANGLAIS (également appelé petits guillemets ou guillemets dactylographiques) est le caractère le plus fréquemment utilisé en anglais pour indiquer des guillemets. La plupart des traitements de texte permettent de convertir automatiquement les GUILLEMETS ANGLAIS en des glyphes contextuels arrondis.

Guillemets-virgules. U+201A ‚ GUILLEMET-VIRGULE INFÉRIEUR et U+201E „ GUILLEMET-VIRGULE DOUBLE INFÉRIEUR représentent des guillemets anglais ouvrants non ambigus. Tous les autres guillemets ont une sémantique hétérogène. Ils peuvent représenter des guillemets ouvrants ou fermants selon l'usage typographique de la langue considérée.

Usage européen. Chaque langue, et parfois même chaque domaine, utilise les guillemets à sa façon. En typographie européenne, on utilise habituellement des guillemets droits (guillemets chevrons « ») pour les livres et, sauf dans certaines langues, des guillemets arrondis en bureautique. On emploie parfois des guillemets chevrons simples pour les citations reprises dans d'autres citations. La figure 7-1 illustre certaines de ces conventions. Dans cette section, on a omis dans les noms de caractères les qualificatifs simple et double quand il n'y a pas d'ambiguïté ou lorsque les deux types de guillemets sont visés.

L'anglais, l'italien, le néerlandais, le portugais et le turc utilisent un *guillemet arrondi gauche* et un *guillemet arrondi droit* pour représenter, respectivement, un guillemet ouvrant et fermant. Il est d'usage d'alterner les guillemets simples et doubles pour indiquer les citations incluses. L'emploi de guillemets simples ou doubles pour la citation de premier rang (la citation extérieure) dépend des conventions locales et stylistiques.

L'allemand, le slovaque et le tchèque utilisent les guillemets-virgules inférieurs comme guillemets ouvrants et utilisent les guillemets-virgules supérieurs comme guillemets fermants. Dans les livres allemands, lorsqu'on utilise les chevrons, ceux-ci pointent vers le texte cité. Cet usage est inverse à celui du français.

Le danois, le finnois, le norvégien et le suédois emploient des *guillemets arrondis droits* à la fois comme guillemets ouvrant et fermant. Ceci s'applique aussi bien en bureautique que dans l'édition d'ouvrages imprimés. Certains livres, toutefois, utilisent le chevron vers la droite (U+00BB » GUILLEMET DROIT) à la fois pour ouvrir et fermer une citation.

Le hongrois et le polonais suivent les conventions scandinaves, sauf qu'ils utilisent des apostrophes arrondies inférieures au début des citations. Ces langues s'abstiennent apparemment d'utiliser les guillemets-apostrophes simples afin d'éviter qu'on ne les confonde avec des virgules.

Le français, le grec, le russe et le slovène, entre autres, emploient les guillemets-chevrons, le slovène oriente cependant ces guillemets à l'allemande. Une de ces langues au moins, le français, insère une espace entre les guillemets et le texte. Dans le cas du français, on utilise[10] U+00A0 ESPACE INSÉCABLE, facilitant ainsi la bonne mise en œuvre de la coupure de ligne.

[10] En fait, ce devrait être une espace fine, mais souvent les logiciels ne gardent pas la connotation *insécable* qui devrait être associée à cette espace !

Figure 7-1. Guillemets européens

Guillemet apostrophe droit = apostrophe

‘anglais’ chef-d’œuvre

L'utilisation dépend de la langue

« français » “anglais”

„allemand“ »slovène«

”suédois” »livres suédois»

Usage extrême-oriental. Le glyphe de chaque guillemet CJC occupe la plus grande partie d’un seul quadrant de la cellule d’affichage de caractère. Le quadrant utilisé dépend de sa fonction (ouvrant ou fermant) ainsi que du sens du texte (horizontal ou vertical). Le tableau 7-3 illustre les paires de guillemets d’Asie orientale.

Tableau 7-3. Guillemets d’Asie orientale

Style	Ouvrant		Fermant	
Anglet	U+300C	「	U+300D	」
Anglet ajouré	U+300E	『	U+300F	』
Double prime	U+301D	〝	U+301F	〟

Variations glyphiques. Le dessin des guillemets « doubles primes » ressemble à une paire d’encoches inclinées vers l’avant ou l’arrière et dont la pointe se trouve en haut ou en bas. L’inclinaison des caractères ouvrant et fermant d’une paire de guillemets « double prime » s’oppose. La figure 7-2 en illustre deux variantes fréquentes. Pour rendre les choses encore moins claires, on emploie une autre forme de guillemets double-prime avec les textes latins horizontaux, en plus des guillemets arrondis simple ou doubles.

Figure 7-2. Guillemets asiatiques

Glyphes horizontaux ou verticaux

『字内』

Glyphes pour les numéros de caractères surchargés

〝字内〟 “Text”

Variantes de glyphes en fonction de la police

〟字内〃 〝字内、

Le tableau 7-4 reprend les trois paires de guillemets utilisées dans la composition des textes occidentaux horizontaux.

Tableau 7-4. Formes ouvrantes et fermantes

Style	Ouvrante		Fermante		Commentaire
Simple	U+2018	‘	U+2019	’	Affichés comme des caractères pleine chasse
Double	U+201C	“	U+201D	”	Affichés comme des caractères pleine chasse
Double prime	U+301D	〝	U+301E	〞	

Numéros de caractère surchargés. Les numéros de caractère des guillemets habituels peuvent représenter les guillemets à chasse étroite d'une police européenne ainsi que les guillemets à large chasse d'une police asiatique employés avec les autres caractères larges (les idéogrammes, par exemple). Le balisage du texte à l'aide d'étiquettes linguistiques permet habituellement de produire un résultat acceptable.

Conséquences sémantiques. La signification de U+00AB et U+00BB (les guillemets-chevrons) ainsi que celle de U+201D (guillemet-apostrophe droit) dépend du contexte. Les guillemets-virgules U+201A et U+201B représentent toujours des formes ouvrantes ; U+301E 〞 GUILLEMET DOUBLE PRIME, en revanche, est toujours, sans équivoque, un guillemet fermant.

Apostrophes

U+0027 ' APOSTROPHE est le caractère le plus fréquent pour représenter l'apostrophe. Toutefois, sa sémantique et sa direction sont ambiguës. En composition, il est préférable d'utiliser U+2019 ’ GUILLEMET-APOSTROPHE pour représenter l'apostrophe. Les texteurs fournissent habituellement une fonction qui permet de convertir automatiquement, selon le contexte, l'APOSTROPHE U+0027 en une apostrophe arrondie.

Lettre apostrophe. On recommande d'utiliser U+02BC ʼ LETTRE MODIFICATIVE APOSTROPHE lorsque l'apostrophe représente une lettre modificative (comme dans les translittérations où ce signe sert à représenter un coup de glotte). On l'appelle alors également une *apostrophe diacritique*.

Guillemet apostrophe. Il est préférable d'employer U+2019 ’ GUILLEMET-APOSTROPHE quand on désire représenter un signe de ponctuation, comme c'est le cas pour l'élision « Je l'ai. »

Une mise en œuvre ne doit pas supposer que le texte d'un utilisateur adhère aux distinctions définies entre ces caractères. Le texte peut, en effet, provenir de différentes sources et même être le résultat d'un transcodage à partir d'un autre jeu de caractères qui ne distingue pas la l'apostrophe diacritique de celles de ponctuation (y compris le guillemet apostrophe droit). Dans ce cas, on représente, en règle générale, *tous* ces caractères à l'aide de U+2019.

La sémantique de U+2019 dépend donc du contexte. Si ce signe est entouré, par exemple, des deux côtés par des lettres ou des chiffres, il s'agit alors d'un caractère de ponctuation du texte : à ce titre il ne sépare ni les mots ni les lignes.

Autres ponctuations

Point de coupure. U+2027 · POINT DE COUPURE DE MOT est un point surélevé qui peut servir, dans certains dictionnaires, à indiquer les points de coupure d'un mot : dic·tion·nai·re. Il s'agit d'un signe de ponctuation qu'il faut différencier du U+00B7 · POINT MÉDIAN polysémique.

Barre de fraction. On forme une fraction en plaçant un filet, U+2044 ⁄ BARRE DE FRACTION entre les chiffres qui la composent : 2/3, 3/9, etc. La forme typique d'une fraction formée à l'aide d'une barre de fraction se définit de la façon suivante : « Une suite d'un ou plusieurs chiffres décimaux suivie d'une barre de fraction, elle-même suivie d'un ou plusieurs chiffres décimaux. » Cette fraction doit être représentée comme une unité : ¾ ou $\frac{3}{4}$. La forme précise de présentation dépend d'informations supplémentaires de formatage.

Si le logiciel d'affichage ne parvient pas à associer à la fraction un seul glyphe, on peut alors se replier sur une représentation linéaire (par exemple 3⁄4). S'il faut séparer la fraction d'un nombre qui la précède, on peut alors insérer une espace (de chasse appropriée : normale, fine, sans chasse, etc.). Ainsi, 1 + ESPACE SANS CHASSE + 3 + BARRE DE FRACTION + 4 peut être affiché 1¾.

Tiret en chef avec chasse. Le TIRET EN CHEF (U+203E ‾) est l'équivalent suscrit du TIRET BAS (U+005F _). Ce caractère chasse, il ne faut pas le confondre avec U+0305 ◌̅ DIACRITIQUE TIRET HAUT. Comme tous les tirets souscrits et suscrits, une suite d'un de ces caractères forme une ligne horizontale continue, à la différence du U+0304 ◌̄ DIACRITIQUE MACRON dont une suite forme une ligne horizontale discontinue.

Ponctuation doublée. Unicode comprend plusieurs caractères de ponctuation doublée dont la décomposition de compatibilité est formée de deux signes de ponctuation : U+203C ‼ DOUBLE POINT D'EXCLAMATION, U+2048 ⁈ POINT D'INTERROGATION-EXCLAMATION et U+2049 ⁉ POINT D'EXCLAMATION-INTERROGATION. Ces signes de ponctuation doublée sont inclus dans Unicode afin de faciliter la mise en œuvre des textes mongols et d'Asie orientale affichés verticalement.

Puces. La puce est habituellement représentée par U+2022 • PUCE. Unicode code des formes supplémentaires de puce au sein du bloc consacré à la ponctuation générale : U+2023 ‣ PUCE TRIANGULAIRE, U+204C ⁌ PUCE NOIRE TRONQUÉE À DROITE et ainsi de suite. On utilise également souvent U+00B7 · POINT MÉDIAN comme une petite puce. La puce marque le début de paragraphes au formatage particulier, souvent des listes[11]. Il peut prendre de très nombreux aspects : images, vignettes ainsi que des formes plus conventionnelles. On emploie souvent, de façon un peu directive, U+261E ☞ INDEX BLANC POINTANT VERS LA DROITE pour faire ressortir une remarque dans un texte.

Signes de paragraphe On signale parfois, à la manière de balises d'édition, les paragraphes et les alinéas à l'aide des caractères U+00A7 § PARAGRAPHE et U+00B6 ¶ PIED DE MOUCHE. Il n'existe pas de norme précise quant à savoir quel caractère représente l'alinéa ou le paragraphe[12]. Il est assez fréquent qu'on emploie également U+204B ⁋ PIED DE MOUCHE RÉFLÉCHI pour représenter le signe de paragraphe.

Ponctuation et symboles CJC : U+3000 – U+303F

On retrouve dans ce bloc les symboles et la ponctuation principalement utilisés dans les systèmes d'écriture fondés sur les idéogrammes hans. Certains signes de ponctuation, plus particulièrement les crochets, peuvent également s'utiliser dans d'autres contextes typographiques. La plupart de ces caractères proviennent des normes d'Extrême-Orient.

Unicode a inclus U+3000 ESPACE IDÉOGRAPHIQUE à des fins de compatibilité avec les jeux de caractères historiques. Il s'agit d'une espace à chasse fixe adaptée aux polices

[11] En typographie française, ces listes sont plutôt marquées par des tirets cadratins.

[12] En typographie française, toutefois, on distingue l'alinéa (changement de ligne marquant le début d'un paragraphe), le paragraphe (dont le mot peut-être abrégé par le signe §) et les marques, comme ¶ mises dans le texte pour indiquer un alinéa.

idéographiques. Les caractères U+301C 〜 TRAIT D'UNION EN ESSE et U+3030 〰 TRAIT D'UNION ONDULÉ sont deux formes particulières du trait d'union que l'on rencontre dans les normes de jeu de caractères extrême-orientales. (Pour une liste des autres espaces et des tirets repris dans le standard Unicode, veuillez vous référer aux *Tableaux 7-1* et *7-2*.)

U+3037 〷 SYMBOLE TÉLÉGRAPHIQUE IDÉOGRAPHIQUE SÉPARATEUR DE CHANGEMENT DE LIGNE est un pictogramme indiquant un changement de ligne dans le code télégraphique chinois : il s'apparente aux pictogrammes du *Bloc de pictogrammes de commande*.

U+3005 々 MARQUE IDÉOGRAPHIQUE D'ITÉRATION remplace le deuxième élément d'une paire d'idéogrammes identiques placés côte à côte.

On écrit souvent un U+3006 〆 MARQUE IDÉOGRAPHIQUE DE FERMETURE sur les écriteaux pour indiquer qu'un magasin ou une cabine est fermé. Les Japonais prononcent ce signe *chimé*, on le rencontre le plus souvent dans le mot composé *shime-kiri.*

La chasse de U+3008 〈 CROCHET ANGULAIRE À GAUCHE et U+3009 〉 CROCHET ANGULAIRE À DROITE est ambigüe. Celle-ci est large dans un contexte asiatique, elle est cependant étroite dans les autres contextes, comme celui des mathématiques. D'autres caractères de ce bloc partagent les mêmes traits ; c'est le cas des doubles crochets angulaires, des crochets en écaille et des crochets blancs.

Le signe U+3012 〒 MARQUE POSTALE précède le code postal dans les adresses japonaises. On l'utilise également sur les formulaires pour indiquer l'endroit où il faut inscrire un code postal. Les caractères U+3020 〠 VISAGE EN MARQUE POSTALE et U+3036 〶 MARQUE POSTALE CERCLÉE sont simplement des variantes glyphiques d'U+3012 et sont inclus pour des raisons de compatibilité.

On n'utilise les caractères U+3031 〱 MARQUE VERTICALE KANA DE RÉPÉTITION et U+3032 〲 MARQUE VERTICALE KANA DE RÉPÉTITION AVEC SON VOISÉ que dans les textes japonais *écrits verticalement* ; ils servent à répéter la paire de kanas qui les précède. La variante voisée U+3032 s'emploie quand les kanas à répéter doivent être voisés. Ainsi, une proposition répétitive comme *toki-doki* peut s'écrire <U+3068 U+304D U+3032> dans un texte orienté verticalement. Ces deux caractères doivent se représenter à l'aide de glyphes de « hauteur double » qui prennent deux « cellules » idéographiques au rendu. Ceci explique l'existence dans les normes d'origine de caractères qui représentent les moitiés supérieures et inférieures de ces caractères (il s'agit des caractères U+3033, U+3034 et U+3035). Dans un texte horizontal, on utilise des caractères similaires auxquels Unicode a attribué d'autres numéros de caractère. En hiragana, les signes de répétition équivalents ont les numéros U+309D ゝ et U+309E ゞ , alors qu'en katakana il s'agit de U+30FD ヽ et U+30FE ヾ .

Idéophonogrammes inconnus ou indisponibles

U+3013 〓 SIGNE GETA sert à indiquer l'absence et la place d'un idéogramme indisponible lors du rendu. Il n'a pas d'autre utilité. Son nom provient de sa ressemblance avec les traces laissées par les socques japonais (*geta*). Deux variantes existent, l'une maigre et l'autre grasse.

Le caractère U+ 303E 〾 INDICATEUR DE VARIATION IDÉOGRAPHIQUE est un caractère graphique qu'il faut afficher. Il signale au lecteur que le caractère voulu est similaire, mais non égal, au caractère suivant. Il s'utilise à la façon du U+3013 〓 SIGNE GETA. Un signe geta remplace un caractère inconnu ou indisponible, mais il ne le désigne pas. L'INDICATEUR DE VARIATION IDÉOGRAPHIQUE constitue le premier de deux caractères qui fournissent une indication quant au glyphe ou au caractère souhaité. En bout de ligne, l'INDICATEUR DE VARIATION IDÉOGRAPHIQUE et le caractère qui le suit ont pour but de se voir remplacer par le caractère approprié après,

d'une part, que celui-ci a été repéré ou, d'autre part, qu'une police ou une méthode de saisie adéquate a été fournie.

U+303F 〿 DEMI-ESPACE IDÉOGRAPHIQUE indique de façon visible une « bourre » de cellule d'affichage utilisée lorsque des caractères idéographiques d'un jeu de caractères à deux octets ont été séparés au rendu. Unicode a repris ce caractère pour des raisons de compatibilité.

Voir également *Suites descriptives idéographiques* à la *Section 11.1*, *Han*.

Formes de compatibilité CJC : U+FE30 – U+FE4F

Certains formes de présentation codées dans ce bloc proviennent de la norme CNS 11643 de la République de Chine (Taïwan). On a inclus ces formes verticales des signes de ponctuation pour des fins de compatibilité avec les mises en œuvre existantes qui codent ces caractères explicitement quand un texte chinois est mis en page à la verticale plutôt qu'à l'horizontale. Unicode recommande plutôt d'utiliser la forme générique de ces caractères et non leurs variantes verticales. C'est au rendu qu'on sélectionne les œils correspondants à l'orientation.

Petites variantes de formes : U+FE50 – U+FE6F

La norme nationale CNS 11643 de la République de Chine (Taïwan) code également quelques variantes de la ponctuation ASCII.

Bien que les caractères de ce bloc soient interprétés comme des caractères à pleine chasse (idéographique), ils sont cependant affichés à l'aide de petits œils au sein de cellule de rendu de pleine chasse. (Veuillez vous référer à l'examen de la question à la *Section 10.3*, *Katakana*). Ces caractères sont inclus à des fins de compatibilité avec les mises en œuvre préexistantes.

Unifications. Deux variantes de petit œil issues du plan 1 de la norme CNS 11643 ont été unifiées avec d'autres caractères externes au bloc ASCII : 2136_{16} a été unifié avec U+00B7 • POINT MÉDIAN et 2261_{16} avec U+2215 ∕ BARRE OBLIQUE DE DIVISION.