Chapitre 11

Écritures de l'Extrême-Orient

L'écriture idéographique[1] développée en Chine deux mille ans avant J.-C. est à l'origine de tous les systèmes d'écriture utilisés aujourd'hui en Extrême-Orient. Conçue pour écrire les divers dialectes chinois, l'écriture idéographique s'étendit aux pays de la zone d'influence chinoise et servit à écrire pendant des siècles d'autres langues que le chinois, comme le japonais, le coréen et le vietnamien. Des locuteurs d'autres langues, comme les Yi, créèrent leur propre système idéographique en imitant le chinois.

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

Jusqu'au XX^{e} siècle, le vietnamien ne s'écrivait qu'à l'aide d'idéogrammes. Cette écriture, parfois qualifiée d'annamite, fut ensuite remplacée par un alphabet dérivé de l'écriture latine qu'avait codifié dès le XVII^{e} siècle par Alexandre de Rhodes, un jésuite français. Le japonais emploit encore aujourd'hui abondamment des idéogrammes appelés *kanji* ; ils sont plus rares en coréen, où les idéogrammes se nomment *hanja*. En Chine continentale, le gouvernement tente de promouvoir l'utilisation d'idéogrammes modernes et simplifiés plutôt que ceux, plus anciens et plus traditionnels, utilisés à Taïwan ou dans les communautés chinoises d'outre-mer.

L'annexe *H, L'histoire de l'unification han,* décrit comment les diverses traditions typographiques de la Chine continentale, de Taïwan, du Japon, de la Corée et du Viêt-nam furent amalgamées afin de fournir, dans la norme Unicode, une série d'idéogrammes systèmes d'écritures à toutes les langues de ces régions.

Unicode inclut la série complète des syllabes hangûl coréennes ainsi que les lettres individuelles (*jamos*) servant également à écrire le coréen. La *section 3.11, Comportement de jamos jointifs,* décrit comment utiliser les jamos jointifs et comment convertir d'une méthode de représentation du coréen à l'autre.

Toutes les écritures de l'Extrême-Orient inscrivent les caractères à l'intérieur d'un carré de grandeur uniforme. Les diacritiques s'utilisent peu, bien que des annotations phonétiques ne soient pas rares. Traditionnellement, les écritures de l'Extrême-Orient s'écrivent du haut de la page vers le bas en colonnes alignées à partir de la droite jusqu'à la gauche de la page. Sous l'influence occidentale, on écrit également de nos jours ces écritures à l'horizontale de gauche à droite.

Plusieurs jeux de caractères moins récents comprennent des caractères destinés à simplifier la mise en œuvre des écritures extrême-orientales, comme c'est le cas des variantes de forme de ponctuation pour les textes écrits verticalement, des formes à demi-chasse (elles n'occupent que la moitié d'un carré) ou des formes à pleine chasse (qui permettent aux lettres latines d'occuper un carré complet). Unicode inclut ces caractères pour des raisons de compatibilité par rapport à ces anciennes normes.

[1] Le terme idéophonogramme décrit de façon plus précise qu'idéogramme la fonction de ces caractères qui possèdent souvent un élément phonétique. Pour des raisons de concision, on a cependant conservé le terme habituel, et quelque peu trompeur, d'idéogramme.

11.1 Han

Idéogrammes CJC unifiés

Les blocs idéographiques unifiés[2] contiennent une série de caractères idéographiques han unifiés utilisés pour écrire le chinois, le japonais et le coréen[3]. Le terme *han*, dérivé de la dynastie chinoise Han, se réfère généralement à la culture chinoise traditionnelle. Les caractères idéographiques han constituent une écriture cohérente qui s'écrit traditionnellement verticalement, les lignes verticales rangées de droite à gauche. Pour un usage moderne, tout particulièrement pour des textes rendus à l'ordinateur ou pour des travaux techniques, l'écriture han s'écrit horizontalement de gauche à droite et se mêle à d'autres écritures, comme le latin. Utilisés en japonais ou en coréen, des caractères propres à ces langues (*hiragana* et *katakana* pour le japonais ; syllabes *hangûl* pour le coréen) se mêlent aux caractères han.

Les caractères idéographiques han constituent un ensemble imposant de plusieurs dizaines de milliers de caractères. Ces caractères sont utilisés depuis fort longtemps en Extrême-Orient. Grâce aux efforts millénaires d'érudits chinois qui relevèrent toutes les différentes formes des caractères han (y compris variantes fausses ou créés pour l'occasion), nous disposons aujourd'hui d'un formidable recueil de caractères idéographiques han.

L'ampleur du répertoire idéographique han et sa difficile normalisation expliquent la longue description consacrée ci-dessous aux caractères han et sa division en sous-sections. Trois des premières sous-sections (normes CJC, blocs et correspondances entre les normes) décrivent respectivement les normes utilisées comme sources, la manière dont Unicode divise les idéogrammes en blocs et quelques problèmes liés à la correspondance entre ces caractères Unicode et des caractères provenant d'autres jeux de caractères. Ensuite, un examen détaillé des caractéristiques est consacré au problème de l'unification du codage de caractères utilisés par différentes langues. Vient enfin un énoncé formel sur les principes qui sous-tendent le codage des caractères unifiés han ainsi que leur ordre dans Unicode. Pour un examen détaillé de l'histoire de l'unification han et de son contexte, consultez également l'*Annexe H, Histoire de l'unification han.*

Normes CJC

Le répertoire de 27 484 caractères[4] han du PMB provient de nombreuses normes de caractères. Ces normes sont regroupées en six sources, comme l'indique le *Tableau 11-1*. Le Groupe à rapporteur sur les idéophonogrammes (GRI), un sous-groupe du JTC1/SC2/GT2 de l'ISO/CEI, effectua le gros œuvre lié à l'unification et au classement des caractères provenant de ces différentes sources.

Les sources G, T, J, K et V représentent les caractères ayant été soumis au GRI par ses membres. La source G contient des caractères provenant du territoire continental chinois, de la zone administrative spéciale de Hongkong et de Singapour. Les quatre autres proviennent respectivement de Taïwan, du Japon, de la Corée et du Viêt-nam. La source U représente les

[2] Les tableaux des idéogrammes unifiés (chapitre 15) se trouvent, sous la forme d'un fichier PDF, sur le disque optique. Les clés chinoises, par contre, font partie de ce livre.

[3] Même si on désigne les langues utilisant actuellement les caractères idéographiques han par le sigle « CJC » (en anglais « CJK »), il faut garder à l'esprit que l'ancienne écriture vietnamienne, dite « annamite », s'inspire également de ces idéogrammes. Le terme « CJCV » serait donc sans doute plus approprié d'un point de vue historique. Les idéogrammes han ne s'utilisent plus de nos jours au Viêt-nam que dans des documents historiques, religieux et pédagogiques.

[4] Le plan complémentaire idéographique (plan 2) du standard Unicode (introduits avec Unicode 3.1) ajoutent 42.711 idéophonogrammes complémentaires ainsi que 542 autres idéogrammes de compatibilité CJC.

caractères n'ayant pas été soumis au GRI par ses membres, mais retenus, néanmoins, par le consortium Unicode.

Tableau 11-1. Sources des han unifiés

Source G	G0	GB2312:80
	G1	GB12345:90 avec 92 caractères « idou » coréens et 58 de Hong-Kong
	G3	GB7589:87 formes non simplifiées
	G5	GB7590:87 formes non simplifiées
	G7	Liste des hanzi (*han-tseu*) polyvalents pour le chinois moderne et liste générale des hanzi simplifiés
	GS	Caractères de Singapour
	G8	GB8565:88
	GE	GB16500:95
Source T	T1	CNS 11643:1992 1er plan
	T2	CNS 11643:1992 2e plan
	T3	CNS 11643:1992 3e plan et quelques caractères supplémentaires
	T4	CNS 11643:1992 4e plan
	T5	CNS 11643:1992 5e plan
	T6	CNS 11643:1992 6e plan
	T7	CNS 11643:1992 7e plan
	TF	CNS 11643:1992 15e plan
Source J	J0	JIS X 0208:1990
	J1	JIS X 0212:1990
	JA	Idéogrammes contemporains de l'Union des constructeurs informatiques japonais, 1993
Source K	K0	KS C 5601:1987 (idéogrammes distincts)
	K1	KS C 5657:1991
	K2	PKS C 5700-1:1994
	K3	PKS C 5700-2:1994
Source V	V0	TCVN 5773:1993
	V1	TCVN 6056:1995
Source U		KS C 5601:1987 (idéogrammes-doublons)
		ANSI Z39.64:1989 (EACC)
		Big-5 (Formose)
		CCCII, niveau 1
		GB 12052:89 (coréen)
		JEF (Fujitsu)
		Code télégraphique de la République populaire de Chine
		Code télégraphique formosan (CCDC)
		Xerox chinois
		Caractères han autorisés dans les patronymes japonais
		Recueil IBM d'idéogrammes japonais et coréens

Dans certains cas, l'ensemble du répertoire idéographique des jeux de caractères d'origine *ne fut pas repris* par la source correspondante. Trois raisons motivèrent cette décision.

1. Lorsque les répertoires de jeux de caractères au sein d'une même source se recouvrent fortement, les caractères en double ne sont inclus qu'une seule fois. C'est la méthode utilisée, par exemple, dans les normes GB2312:80 et GB12345:90 qui possèdent de nombreux idéogrammes en commun. Les caractères GB12345:90 présents dans GB2312:80 ne font pas partie de la source G.

2. Lorsque des jeux de caractères utilisent des règles d'unification fondamentalement différentes de celles utilisées par le GRI, de nombreuses variantes de caractères présentes dans ce jeu de caractères ne font pas partie de la source. Cette situation se produit pour les normes CNS 11643-1992, EACC et CCCII. Il s'agit du seul cas où une compatibilité bijective (aussi nommée « aller-retour ») entre Unicode et le répertoire des idéogrammes han des jeux de caractères correspondants n'est pas garantie.

3. KS C 5601-1987 contient plusieurs doubles d'idéogrammes inclus en raison de leur prononciation multiple en coréen. Ces doublons d'idéogrammes ne font pas partie de la source K, mais plutôt de la source U dans le but d'offrir une compatibilité aller-retour avec KS C 5601-1987.

Blocs

Unicode comprend cinq blocs idéographiques :

1. idéophonogrammes unifiés CJC (4E00 – 9FA5) : les idéogrammes habituels ;
2. supplément A aux idéophonogrammes unifiés CJC (3400 – 4DB5) : des idéogrammes rares ;
3. idéogrammes de compatibilité CJC (F900 – FAFF) : des doublons, des variantes unifiables et des caractères propres aux entreprises ;
4. supplément B aux idéophonogrammes unifiés CJC (U+20000 – U+2A6D6) : autres idéogrammes rares :
5. complément aux idéophonogrammes de compatibilité CJC (U+2F800 – U+2FA1D) : caractères de compatibilité pour les formes traditionnelles du chinois.

Les caractères des blocs *Idéophonogrammes unifiés CJC* et *Supplément A aux idéophonogrammes unifiés CJC* ont été établis par le GRI, ils proviennent entièrement des sources G, T, J, K et V.

Le bloc *Idéophonogrammes unifiés CJC* regroupe les caractères utilisés couramment et présentés au GRI avant 1992. Les caractères du Supplément A aux idéophonogrammes unifiés CJC sont plus rares. Ils furent soumis au GRI entre 1992 et 1998 et ne peuvent être unifiés aux caractères du bloc *Idéophonogrammes unifiés CJC.*

La règle de séparation des sources fut appliquée aux *Idéophonogrammes unifiés CJC* contrairement au *Supplément A aux idéophonogrammes unifiés CJC*. Il s'agit d'ailleurs de la seule différence dans le travail d'unification effectué par le GRI pour ces deux blocs. Cette règle stipule que des idéophonogrammes distincts issus d'une même source ne doivent pas être unifiés. Pour plus d'informations, consultez la sous-section *Principes*, plus loin dans cette section.

Les caractères propres à la source U se retrouvent dans le bloc *Idéogrammes de compatibilité CJC*. Il en existe douze (U+FA0E 塚, U+FA0F 垈, U+FA11 崎, U+FA13 栖, U+FA14 桒, U+FA1F 薗, U+FA21 蛙, U+FA23 趁, U+FA24 辶, U+FA27 錡, U+FA28 錞 et U+FA29 陽). Les autres caractères appartenant à ce bloc sont des doublons ou des variantes unifiables avec d'autres caractères provenant d'un ou de plusieurs autres blocs ; Unicode les inclut pour des raisons de compatibilité bijective.

Caractéristiques générales des idéogrammes han

Le dictionnaire japonais *Kouzien* qui fait autorité définit les caractères han de la façon suivante : « Caractères originaires de Chine servant à écrire la langue chinoise. On les utilise aujourd'hui en Chine, au Japon et en Corée. De nature logographique (chaque caractère

représente un mot plutôt qu'un son simple), ils tirent leur origine de principes pictographiques et idéographiques. Ils peuvent également s'employer phonétiquement. Au Japon, on les nomme généralement *kanzi* (caractère han, c'est-à-dire chinois) ils regroupent aussi les « caractères nationaux » (*kokuzi*) comme *touge* (col de montagne) créés selon les mêmes principes. On les appellent aussi *mana* (« vrais noms », par rapport à *kana*, nom erronés ou empruntés). »

Pendant plusieurs siècles, le chinois fut l'écriture incontestée de l'Extrême-Orient. L'influence de la langue chinoise et de son écriture sur les des langues modernes de l'Extrême-Orient est comparable à l'influence du latin sur le vocabulaire et les formes écrites des langues de l'Occident. Cette influence se perçoit clairement en japonais et en coréen où les caractères han se mêlent aux écritures phonétiques indigènes (kana, au Japon, hangûl, en Corée). (Voir *Tableau 11-2*).

Au cours des siècles, résultat d'une évolution à la fois graphique et sémantique, les formes et le sens des caractères changèrent. Par exemple, le caractère chinois U+6E6F 湯 *t'ang* (japonais *tou* ou *yu*, coréen *t'ang*) signifiait à l'origine « eau chaude » alors qu'il a le sens de « soupe » en chinois. « Eau chaude » demeure l'acception de ce caractère en japonais et en coréen, alors qu'il acquiert le sens de « soupe » dans les emprunts récents au chinois comme « nouilles à soupe » (japonais *tanmen*, coréen *t'angmyen*). Il n'en demeure pas moins que l'identité de forme et la similarité frappant de sens justifient amplement l'idée d'une unification de l'écriture han au-delà des langues qui l'emploient.

La « nationalité » des caractères han ne posa problème qu'à partir du moment où chaque pays créa ses propres jeux de caractères (par exemple, le GB 2312-80 en Chine, le JIS X 0208-1978 au Japon et le KS C 5601-87 en Corée) en ne considérant que ses besoins régionaux. La cause n'en semble pas délibérée, mais plutôt liée à des exigences locales ponctuelles et à un manque de coordination entre les différents pays. Il demeure que les caractères han sont fondamentalement indépendants de la langue, comme le confirment diverses définitions de dictionnaires, listes de vocabulaire et normes de codage.

Tableau 11-2. Caractères han fréquents

Caractère han	Chinois mandarin[a]	Japonais	Coréen	Traduction française
天	tian^{1} (*t'ien*)	ten, ame	tch'en	ciel
地	di^{4} (*ti*)	ti, tuti	tchi	sol, terre
人	ren^{2} (*jen*)	zin, hito	In	homme, personne
山	shan^{1} *(chan)*	san, yama	san	montagne
水	shui^{3} *(chouei)*	sui, mizu	swou	eau
上	shang^{4} *(chang)*	zyou, ue	sang	au-dessus
下	xia^{4} *(hsia)*	ka, sita	ha	en dessous

a. Les chiffres en exposant dans ce tableau représentent les marques de ton chinoises (en mandarin).

Terminologie. Il existe plusieurs termes courants pour désigner les caractères idéographiques extrême-orientaux. Parmi ceux-ci, on rencontre souvent *hanzi* ou *han-tseu* (chinois), *kanzi* (japonais), *kanji* (japonais populaire), *hanja* (coréen) et *chữ hán* (écriture « annamite », vietnamien). Leurs traductions françaises habituelles (caractère han, caractère idéographique han, caractère idéographique de l'Extrême-Orient ou caractère idéographique CJC) sont synonymes. Pour plus de clarté, cet ouvrage utilise un sous-ensemble des termes français pour désigner ces caractères. Le terme *kanzi* fait référence à une publication gouvernementale japonaise précise. Le terme non apparenté de *K'ang-hsi* ou *Kangxi* (nom d'une dynastie

chinoise, plutôt qu'une romanisation de plus des « caractères han ») ne fait référence qu'au dictionnaire qui servit de base à la version 2.0 du *Répertoire et classement unifié*.

Distinguer l'utilisation des caractères han selon la langue. L'unification des caractères han suscite quelques inquiétudes car les langues extrême-orientales les utilisent différemment. Au plan informatique, l'unification des caractères han ne crée pas plus de difficultés que l'emploi d'un seul jeu de caractères latins pour écrire des langues aussi dissemblables que l'anglais et le français. Les programmeurs ne s'attendent pas que les caractères « c », « h », « a » et « t » suffisent à établir déterminer s'il s'agit du mot français *chat* ou du mot anglais signifiant *bavarder*. De même, un Français utilise-t-il le contexte pour interpréter le mot québécois *char* comme équivalent à *voiture.* Peu d'utilisateurs s'étonneront que l'on puisse également utiliser l'ASCII pour représenter des mots comme le terme gallois *ynghyd* qui peut paraître bizarre à des francophones. Même s'il était commode, pour des logiciels de correction orthographique et grammaticale, d'identifier la langue des mots, il ne serait ni pratique ni productif de créer un jeu de caractères latins différent pour chaque langue utilisant ces caractères.

De la même manière, on assemble souvent des caractères han pour « épeler » des mots dont le sens n'est pas nécessairement évident si on considère isolément ses caractères constitutifs. Ainsi, les deux caractères « couper » et « main » signifient-ils « timbre » en japonais, alors que cette forme composée n'aura aucun sens pour un Chinois ou un Coréen (voir *Figure 11-1*).

Figure 11-1. Composition de mots han

切	+	手	=	1. « timbre » en japonais 2. « couper la main » en chinois
couper		main		

Même au sein d'une même langue, un ordinateur doit connaître le contexte pour distinguer les sens des mots représentés par les caractères codés. Le mot japonais *tchoûgokou*, par exemple, peut désigner la Chine ou une région du centre-ouest de Honshū (voir *Figure 11-2*).

Figure 11-2. Contexte des caractères

中	+	国	=	1. Chine (« Empire du milieu ») 2. Tchoûgokou, région de Honshū (Honchoû ou Hondo)
milieu		pays		

Coder 4 fois ces caractères, au lieu des 2 retenus, aurait probablement créé davantage de confusion sans pour autant apporter une solution d'ensemble. Unicode délègue les problèmes d'identification de langue[5] ou de reconnaissance de mots à des logiciels de niveau supérieur et ne prétend pas coder la langue des caractères han.

Tri des idéophonogrammes han. Le standard Unicode ne définit pas de méthode pour trier les caractères idéophonographiques. En effet, le résultat de ce tri dépend du profil culturel (la *locale*) et de l'application. Parmi les ordres lexicographiques possibles, citons l'ordre phonétique, le radical et le nombre de traits utilisés (*K'ang-hsi*, *Xinhua Zidian*, etc.), la méthode des quatre coins et le nombre total de traits. En règle générale, la simple utilisation du numéro de caractère ne permet pas de trier correctement les données selon un des ordres lexicographiques mentionnés ci-dessus ; il faut habituellement recourir à des données auxiliaires. (Voir *Tableau 11-5.*)

[5] Ce n'est plus tout à fait le cas depuis l'ajout par Unicode 3.1 d'étiquettes linguistiques, voir *Chapitre 14-7, Étiquettes.*

Glyphes. Les caractères han sont à chasse fixe. Chaque caractère occupe le même espace vertical et horizontal, peu importe la complexité de la forme. Cette pratique, résultat d'une longue tradition typographique chinoise, inscrit chaque caractère dans une cellule carrée. Écrits verticalement, les caractères han peuvent être dessinés dans un certain nombre de styles cursifs, toutefois ces formes cursives restent inaccoutumées et on ne retrouve pas de polices numériques han habituelles qui les mettent en œuvre.

Les glyphes peuvent fortement varier selon les différents pays et les applications. La police la plus utilisée dans un pays peut ne pas l'être dans un autre.

La diversité des types de glyphes qui illustre les caractères du répertoire han dans cet ouvrage a été limitée par la disponibilité des polices de caractères han. Avant de choisir une poilice han, il est recommandé de consulter des sources fiables et adaptées aux applications et aux marchés visés. On suppose que la plupart des mises en œuvre Unicode permettront aux utilisateurs de choisir la police de caractères (ou un mélange de plusieurs polices de caractères) appropriée à un profil culturel (une *locale*) donné.

Principes

Modèle conceptuel tridimensionnel. Afin de permettre la définition de règles explicites d'unification, on a conçu un cadre de modélisation des caractères han. Ce modèle exprime ces caractères en fonction de trois attributs de base : la sémantique (le sens, la fonction), la forme abstraite (la forme générale) et la forme concrète (l'aspect en contexte, l'œil). On représente ces attributs à l'aide de trois dimensions selon les axes X, Y et Z.

L'attribut sémantique (représenté le long de l'axe X) différencie les caractères selon leur sens et leur utilisation. On distingue des caractères qui n'ont aucun rapport comme 澤 (marais) et 機 (machine) ainsi que des emprunts ou des extensions de sens au-delà du groupe sémantique d'origine comme dans 机$_{1}$ (un emprunt phonétique utilisé comme forme simplifiée de 機) distinct de 机$_{2}$ (table, le sens originel).

L'attribut forme abstraite (représenté le long de l'axe Y) différencie les variantes d'une forme d'un même caractère ayant un seul attribut sémantique (c'est-à-dire un caractère occupant une seule position sur l'axe X).

L'attribut forme concrète (l'œil), représenté le long de l'axe Z, spécifie les caractéristiques de la police de caractères utilisée pour représenter chaque variante de forme.

Seuls les caractères ayant la même forme concrète (c'est-à-dire occupant une seule position sur les axes X et Y) peuvent être unifiés. On ignore habituellement les différences stylistiques et de type d'œil (l'axe Z).

Règles d'unification. On applique les règles suivantes lors de l'unification des caractères han issus de différentes sources de jeux de caractères.

R1 Règle de la séparation des sources. Si une source normative de premier ordre distingue deux idéogrammes, ils restent distincts.

On appelle parfois cette règle la règle *aller-retour* parce qu'elle permet une conversion aller-retour des données de caractères entre la source du GRI et le standard Unicode, et ce sans perte d'informations.

Cette règle fut d'application uniquement lors de la constitution du premier *Répertoire et classement unifiés* (RCU). En 1992, le GRI décida qu'elle ne serait dorénavant plus employer.

Ainsi, les idéophonogrammes provenant du RCU, illustrés à la *Figure 11-4,* auraient été unifiés conformément à la règle R3; néanmoins, puisqu'ils sont distincts dans la source normative J0 (JIS X 0208-1990), on ne les a pas unifiés.

R2 **Règle d'éloignement. En règle générale, si deux idéogrammes ne dérivent pas d'un caractère d'une même origine (les caractères ne sont pas apparentés), on ne les unifie pas.**

Figure 11-3. Conservation des variantes

劍 劍 劒 劒 劒 釼

épée, poignard, sabre

Par exemple, les idéophonogrammes suivants (de la *Figure 11-4*), quoique visuellement semblables, ne sont pas unifiés en raison de différences étymologiques et sémantiques.

Figure 11-4. Non apparentés, non unifiés

土 ≠ 士

terre guerrier, lettré

R3 **On établit la forme abstraite de chaque idéogramme à l'aide d'une classification à deux niveaux (décrite ci-dessous). Toute paire d'idéogrammes qui possède la même forme abstraite est unifiée pour peu que son unification ne contrevienne ni à la règle de la séparation des sources ni à la règle d'éloignement.**

Classification à deux niveaux. À l'aide du modèle tridimensionnel, on classe les caractères en deux niveaux. La classification à deux niveaux différencie les caractères selon la forme abstraite (axe Y) et la forme concrète d'un œil particulier (axe Z). La différence des formes abstraites permet d'identifier les variantes de forme.

Afin d'établir les différences entre forme abstraite et forme concrète, la structure et les caractéristiques de chaque composant idéographique s'analysent selon la méthode présentée ci-dessous.

Structure des composants idéographiques. On analyse la structure des composants de chaque idéophonogramme est étudiée. Un composant est une combinaison géométrique d'éléments primitifs. En assemblant ces composants, il est possible de former divers idéogrammes. Des composants peuvent s'ajouter à d'autres pour créer des structures plus complexes. Un idéophonogramme se définit donc comme un arbre de composants dont le racine représente l'idéogramme au complet et les feuilles les éléments de base (voir *Figure 11-5* et *Figure 11-6*).

Figure 11-5. Structure des composants

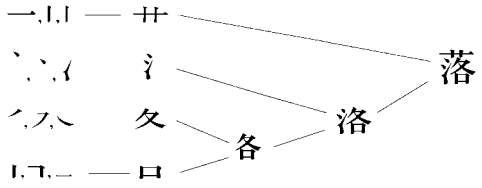
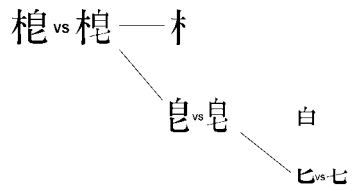


Figure 11-6. Nœud supérieur d'un composant



Caractéristiques des idéogrammes. On compare les idéogrammes en fonction des paramètres suivants :

- leur nombre d'éléments ;
- la position relative des éléments au sein de chaque idéogramme complet ;
- la structure des éléments correspondants ;
- le traitement qu'en font les jeux de caractères sources (l'idéogramme est-il séparé ou non d'un autre qui serait identique selon les critères présentés ci-dessus) ;
- la clé (radical) comprise dans un des éléments.

Unicité. Si un ou plusieurs paramètres des idéophonogrammes comparés diffèrent, on considère que les idéophonogrammes possèdent des formes abstraites différentes. On les traite alors comme des caractères dissemblables qui ne peuvent être unifiés.

Unification. Si tous les paramètres des idéophonogrammes coïncident, les idéogrammes possèdent la même forme abstraite et doivent être unifiés.

Le *Tableau 11-3* recense quelques différences typiques entre formes abstraites de caractères. On n'unifie donc pas ces idéophonogrammes.

Tableau 11-3. Idéogrammes non unifiés

Caractères	Raisons
崖 ≠ 厓	Nombre différent de composants
台 ≠ 峯	Même composants placés dans des positions relativement différentes
拡 ≠ 擴	Même nombre de composants et mêmes positions relatives, la structure des composants correspondants diffère toutefois
区 ≠ 區	Caractères considérés comme différents dans un jeu de caractères source
祕 ≠ 秘	Caractères avec un composant à radical différent
爲 ≠ 為	Même forme abstraite, œils dissemblables

Le *Tableau 11-4* illustre les différences d'œil d'idéophonogrammes Unicode unifiés.

Tableau 11-4. Idéogrammes unifiés

Caractères	Raisons
周 ≈ 周	Ordre différent des traits
雪 ≈ 雪	Différence de dépassement du trait
酉 ≈ 酉	Différence dans le contact des traits
巨 ≈ 巨	Saillie différente des coins de crochet
覀 ≈ 西	Trait vertical plutôt que courbé
朱 ≈ 朱	Différence de terminaison du trait
父 ≈ 父	Accent supplémentaire au début du trait
八 ≈ 八	Trait faîtier modifié
半 ≈ 半	Différence dans la rotation des traits / points

Agencement des idéogrammes han. L'agencement des caractères han dans Unicode prend en considération la position de ces caractères dans les quatre dictionnaires les plus importants. Le dictionnaire *Kangxi Zidian (K'ang-hsi Tseu-tien)* constitue le principal dictionnaire de référence car il contient la plupart des caractères sources, il est largement utilisé et les principes selon lesquels les caractères s'y présentent sont communs à tout l'Extrême-Orient.

Le codage des idéophonogrammes suit l'ordre des dictionnaires mentionnés au *Tableau 11-5*, conformément à leur priorité.

Tableau 11-5. Ordre des idéogrammes han

Priorité	Dictionnaire	Ville	Éditeur	Édition
1	Dictionnaire K'ang-hsi	Pékin	Librairie Zhonghua, 1989	7e
2	Daikanwa Jiten	Tôkyô	Taishuukan Shoten, 1986	9e
3	Hanyu Dazidian	Tch'eng-tou	Sichuan Cishu, 1986	1ère
4	Dèjawon	Séoul	Samsong, 1988	1ère

Lorsqu'un caractère se trouve dans le *Kangxi*, il suit l'ordre prescrit dans le *Kangxi.* Absent du *Kangxi*, mais présent dans le Daikanwa Jiten, on extrapole une position en considérant, dans le *Kangxi*, le caractère le précédant dans le *Daikanwa Jiten. Absent à la fois du Kangxi* et du *Daikanwa*, on a recours aux dictionnaires *Hanyu Dazidian* et *Dèjawon* de la même manière.

On regroupe les idéophonogrammes au radical *Kangxi* simplifiés à la suite des radicaux Kangxi traditionnels dont ils dérivent. Ainsi, les caractères au radical simplifié 讠, lequel correspond au radical Kangxi traditionnel 言, suivent-ils les caractères non simplifiés ayant le radical 言. L'ordre de ces caractères simplifiés suit celui du *Hanyu Dazidian.*

Les quelques caractères ne se trouvant dans aucun des quatre dictionnaires se placent à la suite des caractères ayant le même radical *Kangxi* et le même nombre de traits.

L'ordre radical-trait[6] qui résulte de cette fusion est culturellement neutre. Il ne s'agit pas d'une réplique exacte de l'ordre des dictionnaires courants. L'information nécessaire au tri de tous les idéophonogrammes CJC selon la méthode radical-trait se trouve sur le cédérom. On devrait l'utiliser pour intercaler dans l'ordre correct les caractères contenus dans les différents blocs idéographiques.

Correspondances entre les normes

Le cédérom comprend les correspondances définies par le GRI entre les idéophonogrammes du standard Unicode et les sources du GRI. Ces correspondances constituent une partie normative de l'ISO/IEC 10646-1 ; ceci signifie que les caractères sont définis pour permettre la conversion vers ces différents jeux de caractères normalisés.

Ces correspondances trouvent leur source dans les versions des normes nationales présentées au GRI par ses membres. Dans certains cas, ces versions diffèrent de l'édition publiée de ces normes.

Ces correspondances, établies par le GRI, forme une partie normative de l'ISO/CEI 10646-1 et le sont également pour le standard Unicode. Ces correspondances n'étant pas toujours exactement identiques à celles dérivées à partir des éditions publiées de ces normes sources, les concepteurs de tableaux de correspondances destinées à convertir des caractères provenant de jeux de caractères « réels » peuvent décider d'utiliser d'autres correspondances plus fidèles aux normes publiées de ces jeux de caractères. Des tableaux Unicode précisent où ces variantes de correspondances peuvent s'avérer utiles.

Les systèmes spécialisés dans la conversion peuvent utiliser des mécanismes de correspondances plus complexes : la conversion sémantique, la normalisation des cariantes ou encore la conversion entre forme simplifiée et forme traditionnelle de l'écriture chinoise.

Le Consortium Unicode fournit également des informations de correspondances supplémentaires à celles définies par le GRI. Ces correspondances assurent la transformation entre Unicode et les normes de jeux de caractères de la source U (y compris les caractères doublons du KS C 5601:1987), elles permettent aussi l'association avec des parties de normes de jeux de caractères omises par les sources du GRI et comprennent des renvois à des dictionnaires de référence et d'autres décomptes de traits par caractère.

Idéogrammes de compatibilité CJC : U+F900 – U+FAFF

La norme nationale coréenne KS C 5601 :1987, l'un des jeux sources de référence pour la version 2.0 du *Répertoire et classement unifié* CJC, contient en double 268 idéogrammes identiques afin de représenter d'autres prononciations possibles. En d'autres mots, cette norme code parfois un caractère plusieurs fois pour signaler plusieurs utilisations linguistiques. Cette méthode équivaut à coder 4 fois la lettre « o » pour indiquer ses différentes prononciations dans les mots *bon, mot, mort, coin.* La forme de ces idéogrammes est en tout point identique à celle du caractère de référence et c'est pourquoi le GRI avait exclut ces doublons. Afin de permettre des conversions bijectives, c'est-à-dire aller-retour, entre Unicode et KS C 5601-1987, ces formes se trouvent codées dans un bloc séparé du bloc principal qui regroupe les *Idéophonogrammes unifiés CJC*.

[6] C'est ainsi qu'on nomme l'ordre lexicographique traditionnel des dictionnaires. En effet, presque tous les dictionnaires et les encyclopédies récents classent les caractères d'abord par leur radical (d'où le terme de *clé* qu'on emploie aussi parfois). Ainsi, le caractère étoile (星) se trouve-t-il à la section du radical soleil (日) qui regroupe tous les caractères contenant ce radical. Dans cette section, les caractères sont ensuite classés par ordre croissant du nombre de traits contenus dans les composants restants. Le caractère étoile, dont l'autre élément est composé de cinq traits, prend place vers le milieu de la section. Le disque optique contient un index Unicode radical-trait sous la forme d'un fichier PDF.

En outre, ce bloc contient 34 autres idéogrammes de plusieurs normes régionales ou industrielles, essentiellement afin d'assurer une compatibilité de conversion bijective. Douze de ces idéogrammes (U+FA0E 塚, U+FA0F 垬, U+FA11 崎, U+FA13 栃, U+FA14 桺, U+FA1F 朗, U+FA21 蛙, U+FA23 赱, U+FA24 迈, U+FA27 錡, U+FA28 鐓 et U+FA29 陽) ne sont pas codés dans les zones d'idéogrammes CJC unifiés. Il ne s'agit de doublons. Ces 12 caractères doivent donc être considérés comme un petit complément à l'ensemble des idéogrammes unifiés.

Kanboun : U+3190 – U+319F

Ce bloc contient un jeu de signes kanboun utilisés en japonais pour indiquer l'ordre de lecture japonais des textes classiques chinois. Ils ne sont pas codés dans des normes de codage de caractères courantes, mais s'utilisent fréquemment en littérature. Ils s'écrivent d'ordinaire à la manière d'annotation, à la gauche de chaque ligne verticale du texte chinois.

Consultez aussi les blocs *Lettres et mois CJC cerclés* (U+3200…U+32FF) et *Compatibilité CJC* (U+3300…U+33FF)

Clés chinoises (Kangxi ou K'ang-hsi) : U+2E80 – U+2FD5

Les clés ou radicaux idéographiques, des idéogrammes ou des portions d'idéogrammes, servent de clé d'indexation pour les dictionnaires et les listes de mots, et de base pour la création de nouveaux idéophonogrammes. Le terme *radical,* dérivé du latin *radix* « *racine »,* désigne la partie du caractère qui sert à le classer dans les dictionnaires.

Il n'existe pas un seul jeu de clés idéographiques utilisé dans l'ensemble de l'Extrême-Orient ; Toutefois, les 214 radicaux du dictionnaire *K'ang-hsi* remontant au XVIII^{e} siècle sont pour ainsi dire universels.

Les radicaux peuvent revêtir des graphies très différentes selon qu'ils sont utilisés comme idéogrammes indépendants ou qu'ils font partie d'un idéogramme plus complexe. En effet, portions d'un caractère, les radicaux prennent souvent des formes graphiques différentes. Le radical *eau* est un exemple typique ; idéogramme indépendant il s'écrit水, alors qu'il prend généralement la forme 氵 quand il fait partie d'un idéophonogramme.

Unicode comprend deux blocs de radicaux : les *Clés chinoises (K'ang-hsi ou Kangxi)* (U+2F00 jusqu'à U+2FD5) – ce bloc contient les formes de base des 214 radicaux et les *Formes supplémentaires des clés* (U+2E80 jusqu'à U+2EF3) – celui-ci regroupe un jeu de variantes de formes utilisé lorsque ces clés font partie d'un caractère ou s'écrivent en chinois simplifié. Les dictionnaires et leur index font souvent apparaître ces variantes de forme comme des caractères indépendants. Les radicaux contenus dans ces blocs ne sont pas assujettis aux règles habituelles d'unification en vigueur pour les autres caractères Unicode.

La plupart des radicaux des blocs CJC et des clés chinoises (Kangxi) font également partie du bloc Unicode des *Idéophonogrammes unifiés CJC*. Les clés qui possèdent deux formes graphiques distinctes (une en tant qu'idéogramme indépendant et une autre comme partie d'un idéogramme) voient habituellement ces deux formes codées dans le bloc *Idéophonogrammes unifiés CJC* (tel est le cas pour U+6C34 水 et U+6C35 氵, le radical de l'eau).

Normes. La norme CNS 11643:1992 sépare son bloc de radicaux de son bloc idéographique. Toutefois ce bloc des radicaux ne retient que 212 des 214 clés Kangxi. Toutes ces clés se trouvent incluses dans le bloc Unicode des radicaux Kangxi.

Habituellement, les radicaux qui peuvent être considérés comme des idéophonogrammes à part entière ont un sens précis et on y fait référence par ce sens. De ce fait, le nom de la plupart des caractères du bloc des radicaux Kangxi reflète leur sens. Les autres clés portent un nom qui décrit leur forme.

Sémantique. Il est interdit d'utiliser les caractères provenant des blocs des clés Kangxi et CJC en tant qu'idéophonogrammes. Ils possèdent des propriétés et un sens différents. U+2F00 ⼀ CLÉ CHINOISE UN n'équivaut pas à U+4E00 一 IDÉOPHONOGRAMME UNIFIÉ CJC-4E00. Le premier doit être traité comme un symbole, le second comme un mot ou une partie de mot.

Il est nécessaire de distinguer sémantiquement les caractères utilisés comme idéogrammes de ces mêmes caractères utilisés comme radicaux. Dans le but d'accentuer cette différence, il se peut que les radicaux se voient affecter un style de police de caractères différents de leurs homologues idéophonographiques.

Description idéophonographique : U+2FF0 – U+2FFB

Malgré les quelque 70 000 idéophonogrammes codés dans Unicode[7], plusieurs milliers d'idéophonogrammes extrêmement rares manquent toujours. C'est le cas de près de la moitié des caractères du dictionnaire K'ang-hsi. Les recherches se poursuivent afin de compléter le recueil idéographique, néanmoins il est peu vraisemblable qu'Unicode ne contienne jamais l'ensemble des idéogrammes. En effet, les idéogrammes correspondent *grosso modo* à nos mots et il s'en crée sans cesse. Il est donc inévitable que les derniers néologismes idéographiques soient absents d'Unicode.

Les 12 caractères du bloc de *Description idéophonographique* fournissent un mécanisme normalisé et essentiel à l'échange de textes faisant référence à des idéophonogrammes non codés. Il est, en effet, possible de décrire les idéophonogrammes non codés à l'aide de ces caractères de description et des idéophonogrammes existants ; à partir de cette description, le lecteur peut alors se faire une idée de l'idéophonogramme.

Ce processus diffère du codage formel d'un idéophonogramme. Il n'existe pas de description canonique pour les idéophonogrammes non codés ; on ne leur associe pas de sens ; aucune correspondance vers d'autres caractères n'existe pour les caractères ainsi décrits. Conceptuellement, une description d'idéophonogramme s'apparente plus à la phrase « un « e » surmonté d'un accent aigu » qu'à la suite de caractères « U+006 U+0301 ».

Par ailleurs, la prise en charge des caractères du bloc de description idéophonographique n'implique pas que le moteur de rendu puisse afficher les caractères décrits de la sorte.

Notons également que plusieurs des idéogrammes que les utilisateurs pourraient se voir obligés de représenter aujourd'hui à l'aide des caractères de description idéophonographique feront officiellement partie des prochaines versions d'Unicode.

L'algorithme de la description idéophonographique repose sur le fait que pratiquement tous les idéophonogrammes CJC peuvent être décomposés en plus petites parties formant elles-mêmes des idéophonogrammes à part entière. Unicode codant une grande partie des idéophonogrammes, il est donc possible de représenter la plupart des idéophonogrammes non codés à l'aide des caractères de description idéophonographique.

[7] Unicode 3.1 ajoute 42.711 idéogrammes unifiés aux 27.484 déjà présents dans Unicode 3.0.

Suites de descriptions idéographiques. On définit les suites de description idéographique grâce à la grammaire suivante. Veuillez consulter la *Section 0.2, Conventions de notation* pour plus d'information sur la notation utilisée ci-dessous.

```
SDI ::= IdéogrammeUnifié | Clé | OpérateurDeDescriptionBinaire SDI SDI
       | OpérateurDeDescriptionTernaire SDI SDI SDI

OpérateurDeDescriptionBinaire ::= U+2FF0 | U+2FF1 | U+2FF4 | U+2FF5 | U+2FF6
                                | U+2FF7 | U+2FF8 | U+2FF9 | U+22FA | U+2FFB

OpérateurDeDescriptionTernaire ::= U+2FF2 | U+2FF3

Clé ::= U+2E80 .. U+2EF3 | U+2F00 .. U+2FD5

IdéogrammeUnifié ::= U+3400 .. U+4DB5 | U+4E00 .. U+9FA5 | U+FA0E | U+FA0F
                   | U+FA11 | U+FA13 | U+FA14 | U+FA1F | U+FA21 | U+FA23
                   | U+FA24 | U+FA27 .. U+FA29 | U+20000 .. U+2A6D6
```

Deux contraintes de longueur auxquelles sont soumises les suites de description idéophonographiques s'ajoutent à cette grammaire :

- aucune suite ne peut dépasser 16 valeurs scalaires Unicode en longueur;
- aucune suite ne peut contenir plus de six idéophonogrammes unifiés consécutifs sans caractère de description idéophonographique intermédiaire.

Une suite comprenant des caractères de description idéophonographique mais qui ne respecte ni la grammaire ci-dessus ni les contraintes de longueur ne forme pas une suite de description idéophonographique.

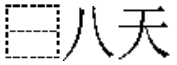
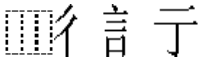
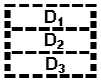
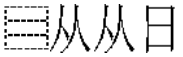
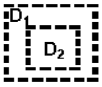
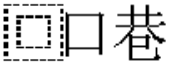
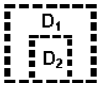
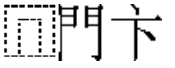
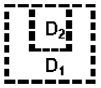
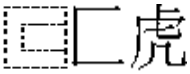
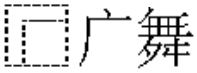
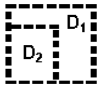
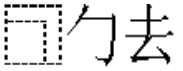
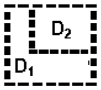
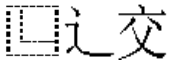
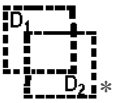
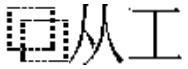
Remarquons que les idéogrammes-doublons de compatibilité (U+F900 à U+FA2D sauf U+FA0E, U+FA0F, U+FA11, U+FA13, U+FA14, U+FA1F, U+FA21, U+FA23, U+FA24, U+FA27, U+FA28 et U+FA29) ne sont pas considérés par cette grammaire comme faisant partie des idéogrammes unifiés, bien qu'ils possèdent la propriété idéographique (voir la *Section 4.9, Lettres et autres propriétés utiles*). Les idéogrammes-doublons ne sont pas admis au sein des suites descriptives idéographiques, afin de réduire progressivement l'ambiguïté de de ces suites de description.

La *Figure 11-7* illustre cette grammaire pour décrire des caractères non codés.

Un utilisateur désireux de représenter un idéophonogramme non codé devra analyser sa structure afin d'établir la manière de le décrire à l'aide des suites de description idéophonographiques. En général, il vaut mieux utiliser la division naturelle clé-élément phonétique de l'idéophonogramme[8], s'il y en a une, et la suite de description la plus courte possible. Toutefois ces règles ne sont pas absolues. Sinon, on préfère cependant la suite de description idéophonographique la plus courte.

8 Rappelons que les idéophonogrammes sont le plus souvent composés d'un radical qui indique une parenté entre tous les caractères qui le comportent et d'un autre composant souvent phonétique. Le caractère « cheval » illustre bien le rôle de l'élément phonétique. Seul on le prononce *ma* et on l'écrit 马. Associé à deux radicaux « bouche » en chef 骂, le sens devient « insulter » mais le résultat se prononce *mà*. Adjoint au radical « insecte » 蚂, nous obtenons « fourmi » qu'on prononce à nouveau *ma*.

Figure 11-7. Utilisation des caractères de description idéographique

Nom du caractère de description idéographique	Positions relatives des CDI	Exemple de SDI	Idéogramme Representé
DE GAUCHE À DROITE			
DE HAUT EN BAS			
DE LA GAUCHE AU MILIEU PUIS À LA DROITE			
DU HAUT AU MILIEU PUIS EN BAS			
TOUT AUTOUR			
AUTOUR À PARTIR DU HAUT			
AUTOUR À PARTIR DU BAS			
AUTOUR À PARTIR DE LA GAUCHE			
AUTOUR À PARTIR D'EN HAUT À GAUCHE			
AUTOUR À PARTIR D'EN HAUT À DROITE			
AUTOUR À PARTIR D'EN BAS À GAUCHE			
CHEVAUCHEMENT			

NOTE : D_1 et D_2 se chevauchent. Ce diagramme ne signifie pas que D_1 est au coin supérieur gauche et D_2 au coin inférieur droit.

Les contraintes de longueur permettent de borner les accès directs aux chaînes d'idéophonogrammes. Pour établir l'appartenance d'un caractère à une suite de description idéophonographique, il suffit d'examiner les quelques caractères qui le précèdent.

Une suite de description idéophonographique pouvant en contenir d'autres, il se peut que les mises en œuvre doivent tenir compte des profondeurs de récursivité et de pile.

La profondeur de récursivité d'une suite descriptive idéophonographique représente le nombre maximum d'opérations en cours permises lors de l'analyse syntaxique d'une suite de description idéophonographique. Dans l'exemple ci-dessous, la profondeur de récursivité

maximale vaut 3 car trois opérations sont en cours à la fin de la suite de description idéophonographique.

La profondeur de pile correspond au nombre maximum d'idéophonogrammes consécutifs sans caractères de description idéophonographique intermédiaire. Aucun des exemples de la *Figure 11-8* ne possèdent plus de trois idéophonogrammes consécutifs; la profondeur de pile totale vaut donc trois.

Officiellement, Unicode ne précise pas la profondeur de récursivité maximale d'une suite de description idéophonographique. Par contre, la profondeur de pile d'une suite de description idéophonographique valable doit être inférieure à six.

Équivalences. Il est possible de décrire de nombreux idéophonogrammes non codés de plusieurs façons en utilisant cet algorithme, soit parce que les éléments de la description peuvent être eux-mêmes davantage divisés, soit parce qu'Unicode code plus d'une fois le même composant (doublons du type U+6C34 水 et U+2F54 水 CLÉ CHINOISE EAU).

Unicode ne définit pas de correspondance entre deux suites de description idéophonographique non identiques (c'est-à-dire composées des mêmes caractères).

Les suites descriptives idéophonographiques ne doivent pas remplacer les idéogrammes déjà présents dans Unicode. En effet, la recherche, le tri et d'autres opérations sur les textes échoueraient.

Interaction avec l'indicateur de variation idéographique. Comme pour les idéophonogrammes, l'*indicateur de variation idéographique* (U+303E) peut se placer devant une suite de description idéophonographique pour indiquer que la description ne correspond qu'à une approximation de l'idéophonogramme souhaité. Une suite de caractères comprenant un indicateur de variation idéographique n'est pas une suite de description idéophonographique.

Rendu. Les caractères d'une description idéophonographique sont des caractères visibles. Ils ne doivent pas être traités comme des caractères de commande. La suite U+2FF1 ⿱ U+4E95 井 U+86D9 蛙 doit avoir une apparence distincte de la suite U+4E95 U+86D9.

Une suite de description idéophonographique valable peut être rendue en affichant ses caractères constitutifs séparément ou en en faisant l'analyse syntaxique pour en dessiner l'idéophonogramme ainsi décrit. Dans ce dernier cas, il faut considérer la suite de description idéographique comme une ligature lors du mouvement du curseur, de la détection de la position des caractères sous-jacents et des autres opérations d'interface utilisateur (Voir la *Section 5.12*, *Édition et sélection*).

Frontières de caractère. Les caractères d'une description idéophonographique ne sont pas des diacritiques et il n'est pas nécessaire qu'ils influent sur les limites de caractère ou de mot. Ainsi, la suite U+2FF1 U+4E95 U+86D9 peut-elle être traitée comme une suite de trois caractères ou même de trois mots.

Les mises en œuvre Unicode peuvent choisir d'analyser les suites de description idéophonographique lors de l'établissement des limites de caractère ou de mot, néanmoins ce choix augmentera considérablement la complexité et lenteur des algorithmes en question.

Normes. Les caractères de la description idéophonographique sont issus du GBK – un complément de caractères idéographiques au GB 2312:80. GBK constitue comme une annexe normative du GB 13000.1:93.

Supplément B aux idéophonogrammes CJC unifiés : U+20000 – U+2A6D6

Les idéophonogrammes du supplément B aux idéophonogrammes CJC unifiés constituent un ensemble de 42 711 idéogrammes qui complémente les 27 484 idéogrammes du plan multilingue de base.

Les idéogrammes de ce bloc sont issus de six sources du GRI : les sources G, H, T, J, K et V. Il n'existe pas de source U pour les idéophonogrammes du supplément B. La source H correspond à une nouvelle source du GRI non exploitée pour les blocs han du PMB, elle regroupe les caractères émanant de normes publiées par le gouvernement de Hongkong.

Le *Tableau 11-6*, ci-dessous, reprend la liste des normes et autres références bibliographiques associées à ces six sources du GRI. La deuxième colonne du tableau indique le nom abrégé de cette source, alors que la troisième fournit un nom descriptif. Le consortium Unicode et l'ISO/CEI identifient, dans plusieurs fichiers de données publiés, les sources précises utilisées par le GRI à l'aide de ces noms abrégés.

Tableau 11-6. Sources du supplément B aux han unifiés

Source G	G_KX	idéogrammes du dictionnaire K'ang-hsi (y compris l'additif) absent du PMB
	G_HZ	idéogrammes Hanyu Da Zidian (*Han-yu Ta Tseu-tien*) absents du PMB
	G_CY	Ci Yuan (*Ts'eu Yuan*)
	G_CH	Ci Hai (*Ts'eu Hai*)
	G_HC	Hanyu Da Cidian (*Han-yu Ta Ts'eu-tien*)
	G_BK	Encyclopédie chinoise
	G_FZ	Système *Founder Press*
	G_4K	Siku Quanshu (*Sseu-k'ou Ts'iuan-chou*)
Source H	H	Jeu de caractères complémentaires de Hongkong
Source T	T4	CNS 11643:1992 4e plan
	T5	CNS 11643:1992 5e plan
	T6	CNS 11643:1992 6e plan
	T7	CNS 11643:1992 7e plan
	TF	CNS 11643:1992 15e plan
Source J	J3	JIS X 0213:2000, niveau 3
	J4	JIS X 0213:2000, niveau 4
Source K	K4	PKS C 5700-3:1998
Source V	V0	TCVN 5773:1993
	V2	VHN 01:1998
	V3	VHN 02:1998

Comme pour les autres blocs idéographiques han, les idéophonogrammes du supplément B puisent à différentes versions des normes nationales présentées au GRI par ses membres. Ils peuvent occasionnellement différer quelque peu des versions publiées de ces normes.

À l'instar des autres idéogrammes CJC unifiés, le nom des idéogrammes du supplément B se définit à l'aide d'un algorithme rudimentaire. Ainsi, IDÉOGRAMME DE COMPATIBILITÉ CJC-20000 correspond au nom de l'idéogramme codé à U+20000.

Ces idéogrammes peuvent intervenir dans la composition de suites descriptives idéographiques.

Complément aux idéophonogrammes de compatibilité CJC : U+2F800 – U+2FA1D

Ce bloc est constitué des idéogrammes de compatibilité complémentaires nécessaires à une conversion aller-retour (bijective) entre Unicode et les plans 3, 4, 5, 6, 7 et 15 de la norme CNS 11643:1992. Il ne faut les utiliser à aucune autre fin et, plus particulièrement, ils ne peuvent intervenir dans la composition de suites de description idéographique. On définit les noms des idéogrammes de compatibilité à l'aide d'un algorithme rudimentaire. Ainsi, le nom de l'idéogramme de compatibilité U+2F800 est-il IDÉOGRAMME DE COMPATIBILITÉ CJC-2F800.

11.2 Hiragana

Hiragana : U+3040 – U+309F

L'hiragana est un syllabaire cursif utilisé pour transcrire les mots japonais ainsi que les particules de phrases et les terminaisons flexionnelles japonaises. Il s'utilise également pour indiquer la prononciation des mots japonais. Les syllabes hiragana ont la même prononciation que les syllabes katakana correspondantes.

Normes. Le bloc hiragana correspond à la norme JIS X 0208:1990, étendue à l'aide de la syllabe hors norme U+3094 ゔ VU, présente dans certaines normes de sociétés japonaises.

Diacritiques. Pour former des syllabes voisées et semi-voisées à partir des syllabes de base, l'hiragana et le katakana emploient respectivement U+3099 ◌゙ DIACRITIQUE KATAKANA-HIRAGANA SON VOISÉ et U+309A ◌゚ DIACRITIQUE KATAKANA-HIRAGANA SON SEMI-VOISÉ. Toutes les combinaisons courantes des formes de syllabes de base utilisant ces diacritiques sont déjà codées comme caractères indépendants ; la norme JIS utilise de préférence ces formes précomposées. Ces diacritiques doivent suivre le caractère de base auquel ils s'adjoignent. Comme la plupart des mises en œuvre et la norme JIS considèrent que ces signes chassent, Unicode comprend également deux versions non combinatoires (à chasse) de ces signes codés à U+309B et U+309C.

Marques d'itération. Les deux caractères U+309D ゝ MARQUE D'ITÉRATION HIRAGANA et U+309E ゞ MARQUE D'ITÉRATION VOISÉE HIRAGANA dénotent l'itération (la répétition) de la syllabe précédente selon que la syllabe répétée contient, respectivement, une consonne voisée ou non.

11.3 Katakana

Katakana : U+30A0 – U+30FF

Le katakana est un syllabaire à l'aspect carré pour transcrire en japonais les mots étrangers (habituellement européens). On l'utilise également pour faire ressortir visuellement certains mots japonais. Les syllabes katana ont la même prononciation que les syllabes hiragana correspondantes. Le syllabaire katakana comprend deux caractères, U+30F5 カ SYLLABE KATAKANA KA MINUSCULE et U+30F6 ケ SYLLABE KATAKANA KE MINUSCULE, pour lesquels il n'existe pas de correspondance directe en hiragana ; on les utilise pour respecter de rares conventions orthographiques (par exemple dans la graphie de certains toponymes qui ont recours à des particules japonaises archaïques).

Normes. Le bloc katakana s'inspire de la norme JIS X 0208 :1990.

Simili-ponctuation. On utilise U+30FB ・ POINT MÉDIAN KATAKANA pour séparer les mots des passages non japonais. U+30FC ー MARQUE KATAKANA-HIRAGANA DE SON PROLONGÉ s'utilise surtout en katakana et parfois en hiragana pour indiquer que la voyelle de la syllabe précédente est longue. Les deux marques d'itération U+30FD ヽ MARQUE D'ITÉRATION KATAKANA et U+30FE ヾ MARQUE D'ITÉRATION VOISÉE KATAKANA ont la même fonction en katakana que les deux marques hiragana d'itération correspondantes.

Formes de demi et pleine chasse : U+FF00 – U+FFEF

Dans le cadre de codages informatiques extrême-orientaux, un jeu de caractères à 2 octets (JC2O), tel que le JIS X 0208-1990 ou le KS C 5601-1987, s'utilise généralement de pair avec un jeu de caractères à un octet (JC1O), tel que l'ASCII ou une variante de l'ASCII. On affiche les textes codés à l'aide d'un JC2O et d'un JC1O de telle sorte que les glyphes représentant les caractères JC2O occupent deux cellules d'affichage, dans la mesure où les caractères du JC10 (ASCII) en occupent une. Dans ces systèmes, on désigne sous le nom de *pleine chasse* (*zenkaku*) la largeur de la double cellule d'affichage, on qualifie de *demi-chasse* (*hankaku*) celle de la cellule simple.

Puisque les largeurs d'affichage varient, certains caractères apparaissent souvent deux fois, une fois sous une forme à pleine chasse dans le répertoire JC2O et une fois sous une forme à demi-chasse dans le répertoire JC1O. Afin d'assurer une conversion bijective avec ces systèmes de codage mixtes, il est nécessaire pour certains caractères d'inclure à la fois la forme à pleine chasse et celle à demi-chasse. Ce bloc est composé des formes supplémentaires nécessaires à la conversion de textes employant les deux formes.

Lors de conversions impliquant des jeux de caractères codés sur différents nombres d'octets, il faut considérer tous les caractères de la zone des écritures générales comme des caractères à demi-chasse (*hankaku*) s'ils ont un équivalent à pleine chasse ailleurs dans le standard ou s'ils n'existent pas dans les codages à largeur mixte ; sinon il faut les considérer comme des caractères à pleine chasse (*zenkaku*). Plus particulièrement, la plupart des caractères provenant de la zone des symboles et signes phonétiques CJC, de la zone idéographique ainsi que les caractères des *Idéogrammes de compatibilité CJC*, des *Formes de compatibilité CJC* et des *Petites variantes de forme* doivent être considérés comme des caractères à chasse pleine (*zenkaku*). Pour une description complète des propriétés de largeur extrême-orientale, consultez le *Rapport technique Unicode n° 11*, « East Asian Width », sur le cédérom ou, pour une version tenue à jour, sur le site Internet du consortium Unicode.

Les caractères de ce bloc correspondent à des formes à chasse pleine des caractères du bloc ASCII (sauf ESPACE), à certains caractères du supplément Latin-1 et à quelques symboles monétaires. Ce bloc reprend, en outre, des formes à demi-chasse des jamos de compatibilité hangûl et des katakana. Enfin, quelques caractères provenant de la zone de symboles y sont reproduits (U+FFE8…U+FFEE) avec une propriété de demi-chasse explicite.

Comme pour les autres caractères de compatibilité, il est conseillé d'utiliser les caractères de référence équivalents et des balises styliques pour préciser la force et la chasse des glyphes en question.

Unifications. La forme à pleine chasse de U+0020 ESPACE est unifiée avec U+3000 ESPACE IDÉOGRAPHIQUE.

11.4 Hangûl

Jamos hangûl : U+1100 – U+11FF

On peut considérer le hangûl coréen comme une écriture syllabaire. Contrairement à plusieurs autres écritures syllabaires, les syllabes se forment à partir d'un jeu d'éléments alphabétiques, de manière régulière. Chacun de ces éléments alphabétiques se nomme *jamo*.

Le standard Unicode prévoit à la fois un jeu complet des blocs de syllabes hangûl précomposées modernes et un jeu de jamos jointifs. Ce jeu de jamos hangûl jointifs peut servir à coder toutes les syllabes coréennes tant modernes qu'anciennes. Pour une description des jamos jointifs et des syllabes précomposées hangûl, consultez la *Section 3.11, Comportement des jamos jointifs*, et la description du bloc des syllabes hangûl (U+AC00..U+D7A3).

Les jamos hangûl se divisent en trois classes : les *tch'ôsong* (consonnes initiales ou caractères syllabiques initiaux), les *djoungsong* (voyelles ou caractères syllabiques centraux) et *djôngsong* (consonnes finales ou caractères syllabiques finaux). Dans l'examen qui suit, on représente ces différentes classes à l'aide d'un I (consonne initiale), V (voyelle) ou F (consonne finale).

Pour la composition, deux bourres invisibles permettent de représenter un tch'ôsong et un djoungsong : U+115F BOURRE HANGÛL TCH'ÔSONG et U+1160 BOURRE HANGÛL DJOUNGSONG

Tri. Dans les textes coréens, l'unité de tri habituelle est le bloc syllabique hangûl. L'agencement des jamos jointifs permet de trier ces suites à l'aide de comparaisons binaires. Ainsi, quand on compare (a) IVFIV à (b) IVIV, il faut comparer les premiers groupes syllabiques des deux chaînes (c'est à dire IVF et IV). Si les deux premiers caractères sont identiques – toutes les consonnes finales ayant une valeur binaire supérieure à celle des consonnes initiales – le F sera donc supérieur au second I (b). Ce résultat produit un ordre correct entre les suites. Ce résultat correspond à un ordre de tri correct pour les chaînes de caractères. Le numéro de caractère des bourres a été choisi afin de conserver cet ordre.

- Comme pour tout autre caractère codé, un tri ne se satisfait pas d'une simple comparaison binaire. Des suites irrégulières, constituées par exemple de bourres superflues, produiront des tris incorrects. C'est également le cas quand un non-jamo suit une suite jamo (comme lors de la comparaison de IVF et IVx, où x est un caractères Unicode dont le numéro est supérieur à U+1FFF, par exemple un U+3000 ESPACE IDÉOGRAPHIQUE)

Si on désire trier un ensemble de blocs de syllabes précomposées et des jamos, il est plus facile de décomposer les blocs syllabiques précomposées en jamos jointifs avant de les comparer.

Jamos de compatibilité hangûl : U+3130 – U+318F

Ce bloc comprend des voyelles et consonnes hangûl (jamos) à chasse et non jointives. Ces caractères ne sont repris qu'afin d'assurer une compatibilité avec la norme KS C 5601. Contrairement aux caractères provenant du bloc de jamos hangûl (U+1100..U+11FF), les caractères jamos de ce bloc ne possèdent pas de sémantique jointive.

On considère les caractères de ce bloc comme des formes à pleine chasse, contrairement aux formes à demi-chasse des jamos de compatibilité hangûl codés à U+FFA0..U+FFDF.

Normes. Le standard Unicode respecte la norme KS C 5601 en ce qui à trait aux éléments jamos hangûl.

Syllabes hangûl : U+AC00 – U+D7A3

L'écriture hangûl utilisée par le système d'écriture coréen se compose de voyelles et de consonnes autonomes (jamos) qui se combinent visuellement à l'intérieur de cellules d'affichage carrées pour former des blocs de syllabe complète[9]. Les syllabes hangûl peuvent être codées directement comme des combinaisons précomposées de jamos ou des suites décomposées de jamos jointifs. La seconde méthode se code à l'aide de caractères du bloc des jamos hangûl (U+1100…U+11FF).

On peut exprimer les blocs de syllabes hangûl modernes à l'aide de deux ou trois jamos sous la forme *consonne* + *voyelle* ou *consonne* + *voyelle* + *consonne*. Il existe 19 consonnes initiales (tch'ôsong), 21 voyelles (djoungsong) et 27 consonnes finales (djôngsong). On peut donc former 399 blocs syllabiques de deux jamos et 10 773 blocs syllabiques de trois jamos possibles, pour un total 11 172 blocs de syllabes hangûl modernes. On appelle l'ensemble des 11 172 syllabes hangûl modernes codées dans ce bloc, le jeu Tchôhap *(Johab).*

Normes. Les syllabes hangûl proviennent de la norme KS C 5601:1992 (norme qui représente l'intégralité du jeu Tchôhap). Ce groupe constitue un surensemble des syllabes hangûl codées des normes coréennes antérieures (KS C 5601:1992, KS C 5657:1991).

Équivalence. Chaque syllabe hangûl de ce bloc peut également être représentée par une suite de jamos jointifs équivalente ; toutefois l'inverse est impossible car des milliers d'anciennes syllabes hangûl n'existent pas sous la forme de syllabes précomposées et ne peuvent être codées qu'à l'aide de suite de jamos jointifs.

Composition des syllabes hangûl. Les syllabes hangûl peuvent être composées à partir de jamos jointifs selon un processus régulier. La *Section 3.11, Comportement des jamos jointifs,* décrit par le détail l'algorithme de mise en correspondance entre une suite de jamos jointifs et le numéro de caractère d'une syllabe hangûl dans le jeu Tchôhap.

Décomposition des syllabes hangûl. Réciproquement, on peut décomposer chaque syllabe hangûl provenant du jeu Tchôhap en une suite de caractères de jamos jointifs. La *Section 3.11, Comportement des jamos jointifs*, décrit avec précision l'algorithme de décomposition.

Nom des syllabes hangûl. Le nom des syllabes hangûl est dérivé algorithmiquement à partir du nom de ses composants. (Pour plus d'information, consultez la *Section 3.11, Comportement des jamos jointifs.)*

Glyphe représentatif des syllabes hangûl. On peut de former le glyphe représentatif d'une syllabe hangûl à partir de ses jamos constitutifs en se fondant sur la classification des voyelles énuméré dans le *Tableau 11-7*.

[9] Rappelons que l'hangûl est un alphabet (on distingue les voyelles des consonnes) qui s'utilise comme un syllabaire (on regroupe toujours les lettres en syllabe) et adopte une mise en page chinoise (les syllabes s'inscrivent dans un carré idéal).

Tableau 11-7. Disposition des syllabes selon l'orientation des djoungsong

Verticale			Horizontale			Horizontale et verticale		
U+1161	ㅏ	A	U+1169	ㅗ	Ô	U+116A	ㅘ	WA
U+1162	ㅐ	È	U+116D	ㅛ	YÔ	U+116B	ㅙ	WÈ
U+1163	ㅑ	YA	U+116E	ㅜ	OU	U+116C	ㅚ	EU
U+1164	ㅒ	YÈ	U+1172	ㅠ	YOU	U+116F	ㅝ	WO
U+1165	ㅓ	O	U+1173	ㅡ	Û	U+1170	ㅞ	WÉ
U+1166	ㅔ	É				U+1171	ㅟ	WI
U+1167	ㅕ	YO				U+1174	ㅢ	ÛI
U+1168	ㅖ	YÉ						
U+1175	ㅣ	I						

Si la voyelle de la syllabe s'appuie sur une ligne verticale, la consonne initiale se met à gauche. Si la voyelle de la syllabe repose sur une ligne horizontale, la consonne précédente se place au-dessus. Enfin, si voyelle s'appuie sur une combinaison de lignes verticale et horizontale, on place la consonne initiale au-dessus de la ligne horizontale et à gauche de la ligne verticale. Dans tous les ces cas, la consonne finale, s'il y en a une, s'écrit en dessous et au milieu du groupe résultant.

mar m a r 말 môs m ô s 못

Quelle que soit la police de caractères, la disposition exacte, l'œil et la taille des éléments varieront en fonction de la forme des autres caractères et du style général de la police.

Consultez également *Lettres et mois CJC cerclés* (U+3200…U+32FF), *Compatibilité CJC* (U+3300…U+33FF), *Formes de demi et pleine chasse* (U+FF00…U+FFEF).

11.5 Bopomofo

Bopomofo : U+3100 – U+312F

Le bopomofo est un jeu de caractères utilisé pour noter ou enseigner la phonétique chinoise, principalement le mandarin. On rencontre ces caractères dans les dictionnaires et le matériel pédagogique, mais pas dans les textes chinois courants. Officiellement, cet alphabet porte les noms chinois de Zhuyin-Zimu (*Tchou-yin Tseu-mou* en EFEO[10], « alphabet phonétique ») et Zhuyin-Fuhao (*Tchou-yin Fou-hao* en EFEO, « symboles phonétiques »). Cependant le nom non officiel de « bopomofo [11]» (analogue à « abécédaire ») est plus commode en français, par surcroît il s'utilise également en Chine. Lors de sa création, le bopomofo faisait partie d'une campagne d'alphabétisation populaire qui suivit l'instauration de la république en 1911 ; il est donc admis par toutes les branches de la culture chinoise moderne, même si, en République Populaire de Chine, le système de romanisation *pinyin* l'a en grande partie remplacé.

Normes. En République populaire de Chine, la norme GS 2312 et, en République de Chine (Taïwan), la norme CNS 11643 incorporent le jeu mandarin du bopomofo.

Signes de ton mandarins. Le système bopomofo comprend de petites lettres modificatives qui notent les cinq tons mandarins. Dans Unicode, ces signes se retrouvent unifiés dans le bloc des lettres modificatives, comme l'indique le *Tableau 11-8*.

Tableau 11-8. Signes de ton mandarins

premier ton	U+02C9	ˉ	LETTRE MODIFICATIVE MACRON
deuxième ton	U+02CA	ˊ	LETTRE MODIFICATIVE ACCENT AIGU
troisième ton	U+02C7	ˇ	CARON
quatrième ton	U+02CB	ˋ	LETTRE MODIFICATIVE ACCENT GRAVE
ton léger	U+02D9	˙	POINT EN CHEF

Bopomofo mandarin normalisé. L'ordre des lettres bopomofo mandarines (U+3015..U+3129) est le même dans le monde entier. Le numéro du premier caractère (U+3105 ㄅ LETTRE BOPOMOFO P) correspond à un multiple de 16 du premier caractère de la norme GB-2312, inscrite au registre de l'ISO. Le caractère U+3127 ㄧ LETTRE BOPOMOFO I peut se rendre comme une barre verticale ou horizontale. On choisit souvent de le représenter perpendiculairement à la ligne de base du texte (c'est-à-dire une barre horizontale dans un texte composé à la verticale). On emploie cependant souvent d'autres conventions. Le standard Unicode représente ce caractère dans les tableaux à l'aide d'un trait horizontal ; on considère la forme verticale comme une variation de rendu. Cette variante d'œil ne possède pas de numéro de caractère particulier.

Bopomofo étendu. Pour représenter les sons propres aux dialectes chinois autres que le mandarin, on a ajouté des caractères phonétiques (U+3105..U+3129) au jeu bopomofo de base. Ce supplément ne jouit pas d'une reconnaissance aussi grande que le jeu mandarin de base. On retrouve les trois caractères bopomofo étendu U+312A..U+312C dans certains ouvrages de référence courants, telle que l'encyclopédie Xin Ci Hai (*Hsin Ts'eu Haï*). Un autre jeu de 24 caractères bopomofo étendu, codé à U+31A0…U+31B7 et conçu en 1948, sert à

[10] La transcription de l'École française d'Extrême-Orient, plus facile à lire pour les francophones non sinisants.
[11] Il s'agit en effet de la valeur des quatre premières lettres de cet alphabet transcrites à la pinyin (en EFEO, on transcrirait *pop'omofo*).

exprimer les sons des « dialectes »[12] hakka et min du Sud. Utilisés de pair avec le jeu bopomofo principal, ces caractères complémentaires permettent une transcription phonétique complète de ces langues. Il n'existe pas de lettres bopomofo normalisées pour transcrire le cantonnais et plusieurs autres langues méridionales de la Chine.

Les petits caractères codés à U+31B4…U+31B7 représentent les consonnes syllabiques finales absentes du mandarin standard et de ses dialectes. De même forme que les caractères bopomofo « p », « t », « k' » et « h », ils sont plus petits que les consonnes initiales correspondantes. En règle générale, on les accole à la voyelle médiane de la syllabe. On a codé séparément ces consonnes finales afin de pouvoir représenter clairement le min du Sud et le hakka dans les textes ordinaires sans devoir les souscrire ou avoir recours à d'autres mécanismes complexes de rendu.

Signes de ton du bopomofo étendu. En plus des signes de ton énumérés au *Tableau 11-8*, le bloc des lettres modificatives comprend également les signes de ton du *Tableau 11-9* destinés à transcrire les tons supplémentaires du min du Sud et du hakka. Le « ton de départ » renvoie au *qusheng* (*ts'iu-cheng*) dans l'analyse tonale chinoise traditionnelle, dont font partie où la variante *yin* découle historiquement d'initiales dévoisées, et la variante *yang* d'initiales voisées. De façon générale, les langues du Sud de la Chine ont conservé davantage de distinctions tonales que le mandarin.

Tableau 11-9. Signes de ton hakka et min du Sud

ton de départ yin	U+02EA	˪	LETTRE MODIFICATIVE SIGNE DE TON DE DÉPART YIN
ton de départ yang	U+02EB	˫	LETTRE MODIFICATIVE SIGNE DE TON DE DÉPART YANG

Rendu du bopomofo. Le bopomofo se rend de gauche à droite et à l'horizontale, on l'écrit cependant à l'occasion à la verticale. Seul, on peut l'écrire dans n'importe quelle direction, mais d'ordinaire il sert d'annotation interlinéaire dans les textes chinois (écrits en caractères han). Dans les livres pour enfants, il n'est pas rare que chaque caractère soit annoté de sa prononciation bopomofo. Cette annotation interlinéaire ressemble structurellement au système d'annotation japonais *ruby (furigana)*, toutefois l'utilisation explicite de signes de ton adjoints aux lettres bopomofo complique le système d'annotation chinois.

Dans les interlinéations horizontales, le bopomofo se place généralement au-dessus du ou des caractères han correspondants ; s'il y en a, les signes de ton apparaissent à la fin de chaque groupe syllabique de lettres bopomofo. Dans les interlinéations verticales, le bopomofo se place sur le côté droit du ou des caractères han correspondants ; s'il y en a, les signes de ton apparaissent sur une rangée interlinéaire séparée et placée à droite de la voyelle. Utilisé pour transcrire le min du Sud et le hakka, les signes de ton peuvent se mêler aux chiffres latins 0-9 pour exprimer les changements de valeurs tonales résultant d'une juxtaposition de tons de base.

[12] L'intelligibilité entre les dialectes mandarins et les « dialectes » du Sud-Est chinois est très faible, voire nulle. Ces « dialectes », parmi lesquels on retrouve le cantonais, le foukien (min), le wou et le hakka (arrière-pays de Canton), sont en fait des langues différentes.

11.6 Yi des Monts frais

Yi : U+A000 – U+A4CF

Le syllabaire yi s'utilise pour écrire la langue yi, un membre de la famille des langues sino-tibétaine. Cette écriture se nomme également cuan (*ts'ouan*) ou wei.

Les Yi, aussi connus sous le nom de Lolo ou Nouo-Sou, sont l'une des minorités non han les plus importantes en République populaire de Chine (RPC). La plupart des Yi vivent au sud-ouest de la Chine, d'autres cependant habitent en Birmanie, au Laos et au Viêt-nam. Le yi est l'une des langues officielles de la RPC.

Les plus anciens textes du yi classique, une écriture idéophonographique, remonte à quelque cinq siècles. Contrairement à d'autres écritures sinoformes, les idéophonogrammes ne semblent pas être des dérivés des idéophonogrammes han. L'écriture yi classique comprend quelque 8 000 à 10 000 caractères, bien que les idéophonogrammes précis varient d'une région à l'autre. L'écriture classique fut d'abord étudié par deux Français, le missionnaire Vial et un officier, d'Ollone. Les transcriptions que donnent d'un même signe Vial et d'Ollone, dont les terrains d'observation ont été différents, sont souvent fort éloignées les unes des autres.

Afin d'augmenter le taux d'alphabétisation chez les Yi, le syllabaire yi fut introduit en 1970. Unicode code ce syllabaire ; l'écriture idéophonographique yi classique ne fait partie d'Unicode à l'heure actuelle.

Parmi les langues lolo, on trouve des langues à trois tonèmes, comme le hani, le nahsi, le lisou, le lahou ; à quatre tonèmes, comme le yi des Monts frais ou à cinq tonèmes, comme le sani.

Chaque syllabe yi se compose d'une consonne initiale optionnelle, d'une voyelle et d'un ton. L'essentiel du syllabaire yi se compose de 820 signes pour les syllabes dotées d'un des trois premiers tons (haut, bas descendant, moyen) et d'un arc ajouté au signe du ton moyen, pour indiquer le quatrième ton (ton montant).

Normes. En 1991, la RPC adopta une norme nationale de codification pour le yi (GB 13134:91). Ce codage englobe 1 165 syllabes yi. Unicode, s'appuyant principalement sur cette norme, inclut tous ces caractères,.

Nomenclatures et ordres. Le nom des syllabes yi correspond à la romanisation de leur prononciation. On y indique le ton par l'ajout à la syllabe romanisée d'une lettre finale : « t » pour le ton haut uni (55) ; « p » pour le ton bas descendant (21) ; « x » pour le ton montant (34) et l'absence de lettre pour le ton moyen uni (33).

Exemples : U+A059 SYLLABE BBIP (prononcé *bì*), U+A410 SYLLABE QURX (*tç'oué*), U+A411 SYLLABE QUR (*tç'oue*) .

Rendu. Le yi respecte les règles d'écriture des idéophonogrammes han. Les caractères s'écrivent de gauche à droite ou, occasionnellement, du haut vers le bas. Les différents caractères de l'écriture yi n'interagissent pas au niveau typographique. Dans l'écriture yi, il n'y a pas d'interaction entre les caractères individuels.

Clés yi. Pour faciliter la recherche de caractères yi dans les dictionnaires, on a créé un ensemble de clés. Le répertoire yi se divise en plusieurs sous-ensembles, chacun partageant une clé (radical) commune. Le nom de cette clé correspond à celui du caractère yi dont la forme s'en rapproche le plus.