Cahiers GUTenberg

EN CHINOIS DANS LE TEXTE

Alain Cousquer

Cahiers GUTenberg, nº 6 (1990), p. 15-24.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1990___6_15_0>

原文是中文
En chinois dans le TEXte

Alain COUSQUER

GEDIS, LIFL, Université de Lille I.

Faisons une hypothèse : alors que vous êtes en train de rédiger à l'aide de votre formateur de texte favori un article de haute tenue sur votre auteur préféré, le très célèbre 白居易, poète bien connu de la dynastie des Tang (VII^{ème} VIII^{ème} siècle après J.C.), vous vous rendez compte avec horreur que, malgré tout votre talent, vous n'arriverez jamais à rendre dans votre traduction les délicates et subtiles nuances du texte original qui en font tout le charme ; respectueux de votre lecteur, une seule solution s'offre à vous : la citation dans la langue originale[1]. Nanti donc de cette bonne résolution, vous vous attelez au problème, et voici ce que ça donne[2] :

古原草　　白居易

离离原上草
一岁一枯荣
野火烧不尽
春风吹又生
远芳侵古道
晴翠接荒城
又送王孙去
萋萋满别情

Bien sûr, j'entend déjà les puristes : «Oui, mais ce sont des caractères simplifiés ! et c'est écrit dans n'importe quel sens ! Comment peut-on rendre toute la beauté de ce texte dans ces conditions et avec de tels caractères ? Autant éditer Ronsard à la ronéo avec une IBM à boule[3] ! Et d'ailleurs, rien ne vaut la calligraphie !». À quoi je leur répondrai qu'il vaut mieux éditer avec les moyens dont on dispose que pas du tout, que Gutenberg leur a déjà répondu d'un geste éloquent dans une publicité bien connue, et que rien ne les empêche d'arrêter ici leur lecture.

Pour ceux qui continuent à me lire, il peut être intéressant de voir comment on peut arriver à ce résultat. Ce sera également une bonne occasion de plonger dans la mécanique interne de votre formateur favori et de mettre ainsi à l'épreuve toutes les connaissances que les meilleurs auteurs de cette revue se sont efforcés de vous fournir dans les derniers numéros. En route donc pour une visite au monde des PXL, TFM et autres FIXes qui hantent

[1] ceux qui, tels Vatel attendant ses poissons, ont pensé au suicide devraient plutôt prendre quelques vacances ...

[2] Si, par hasard, vous ne lisiez pas couramment le chinois classique, en voici une traduction de F. Cheng et E. Simion [1] :

Herbes sur la plaine antique — *Bái Jūyi*

Herbes tendres à travers la plaine
chaque année se fanent et repoussent,
les feux sauvages n'en viennent pas à bout,
au souffle du printemps elles renaissent,
de leurs senteurs parfument l'antique voie,
gerbes d'émeraudes dans les ruines anciennes,
agitées et frémissantes de nostalgie,
elles disent adieu au seigneur qui s'en va ...

[3] mille pardons aux honorables sociétés citées ici, mais ce n'est pas moi qui parle, ce sont les puristes.

le cœur des polices TEXiennes.

Je ne parlerai pas ici de problèmes de saisie, ni du codage informatique des caractères chinois, que je nommerai dorénavant *idéogrammes*, même si cette appellation fait l'objet de controverses sémantiques (voir [2] sur ces questions). Le *glyphe* d'un idéogramme (ou d'une lettre latine) sera son dessin. Il nous suffira de savoir que le codage informatique des idéogrammes utilise deux octets, quelle que soit la norme utilisée, qu'un éditeur de texte adapté à la saisie du chinois produit ces deux octets pour chaque idéogramme sélectionné et qu'il y a deux types de méthodes utilisées pour distinguer, dans un texte multilingue, ce qui appartient à une langue de ce qui appartient à une autre :

- les méthodes de balisage du texte, qui insèrent à chaque changement de langue une indication codée de la langue qui suit, et donc de la longueur, un ou deux octets, avec laquelle on doit lire les données suivantes (c'est le principe utilisé par TEX 3.0 avec la commande `\language`).

- le marquage de chaque caractère lui-même : une des langues (en général celle qui utilise les lettres latines par exemple) codée sur un octet avec le premier bit à 0, l'autre langue, codée sur deux octets, a le premier bit du premier octet au moins mis à 1[4].

  On peut aussi tout coder sur deux octets comme cela se fait sous CCDOS, version chinoise de MSDOS. Dans ce cas, un premier octet égal au caractère # (*"23*) indique une lettre latine dans l'octet suivant, sinon on a affaire à un idéogramme chinois.

En possession de votre fichier contenant votre texte en français et en chinois, il ne vous reste plus qu'à l'imprimer. Ceci implique d'abord la disponibilité de polices chinoises, ensuite la «traduction» de chaque paire d'octets représentant un idéogramme en une séquence d'instructions compréhensible par TEX nécessaire à l'affichage dudit idéogramme. Ce sont ces deux tâches, création des polices chinoises et traduction en TEX des fichiers multilingues que je vais présenter ici, étant bien entendu qu'il ne s'agit pas de créer une version chinoise de TEX (comme il peut exister JTEX et JLATEX, versions japonaises de TEX), mais uniquement de disposer d'un mécanisme simple d'insertion d'idéogrammes dans un texte en lettres latines. Pour ceux qui ne disposent pas d'un éditeur de texte produisant des idéogrammes chinois (ou au moins leur code informatique), on peut insérer «à la main» les commandes de TEX qui permettent de les imprimer ; cela nécessite, outre la disponibilité des fontes chinoises (c'est à dire des fichiers `.tfm` et `.pxl/.pk`), l'extraction du nom de la police et le calcul du rang du caractère dans celle-ci à partir de son code sur deux octets.

La création des polices de caractères

L'idéal, pour écrire en chinois avec TEX, serait de disposer de polices de caractères chinois définies et crées avec METAFONT. C'est un luxe encore inaccessible aujourd'hui, bien que des études pour la génération de telles polices aient été menées depuis plusieurs années [3, 4, 5, 6]. La raison en est simple : comme ses locuteurs, les caractères chinois sont nom-

[4] dans ce cas, il vaut mieux parler en réalité de texte bi-alphabet. Cette méthode a l'inconvénient de limiter à 128 le nombre de caractères disponibles dans la langue codée sur un octet.

breux : en se limitant à ceux qui sont définis dans les normes officielles (et qui disposent ainsi d'un code informatique normalisé sur deux octets), on en recense 6763 pour le code GB 2312-80 de la République populaire de Chine et 13 005 pour le code BIG-5 en usage à Taïwan. On conçoit qu'un tel volume ait fait reculer jusqu'ici tous les fanatiques chinois ou sinisants de TEX ou METAFONT, d'autant plus que les particularités de ces idéogrammes conduisent à envisager leur génération automatique à partir des sous-ensembles qui les composent tous : les traits de base (quelques dizaines) et les radicaux (environ 200, formés eux-mêmes de traits de base).

La composition des caractères chinois obéit en effet à des règles rigoureuses ; enfermé dans un carré idéal (qui peut devenir rectangle pour des effets de style ou de mise en relief), tout idéogramme est une juxtaposition de radicaux, comme le montre la figure (1). L'idée simple qui vient alors tout naturellement à l'esprit pour cette génération automatique serait de définir (dessiner) les traits de base, puis de fournir les règles de création des radicaux à partir de ces traits de base (positionnement et transformations géométriques ou déformations éventuelles) et enfin de fournir dans les mêmes conditions les règles de formation des idéogrammes eux-mêmes à partir des radicaux. Comme il s'agit d'un avenir qui, quoique prometteur, reste encore potentiel, on ne s'étonnera pas que nous ayons suivi une autre voie pour obtenir des résultats plus immédiats.

De quoi peut-on disposer à l'heure actuelle ? Essentiellement de polices matricielles obtenues par lecture optique et numérisation, et qui sont, pour certaines, disponibles dans le domaine public dans des formats divers. Nous disposons ainsi dans notre laboratoire, pour le jeu de caractères correspondant à la norme GB 2312-80 (jeu que nous désignerons par *hanzi simplifié*), de deux fontes au format BDF (format défini au MIT[5] et utilisé dans le monde X-WINDOW) avec des matrices de 16×16 et 24×24, et d'un jeu utilisant des matrices 28×28 fournies avec un terminal multilingue conçu au laboratoire CATAB de l'université J. Moulin à Lyon : ce sont elles qui sont utilisées dans cet article[6]. En voici un exemple : 原文是中文[7].

Parler, dans le monde TEX, de polices matricielles, c'est parler des fichiers `.pxl`, ce (déjà) vieux format utilisé dans les anciennes implémentations de TEX, et qui avait sur son successeur `.pk` un avantage et un inconvénient : il était moins obscur, et c'est ce qui nous intéresse ici, mais par contre il avalait avec une belle gloutonnerie les précieux mégaoctets de nos disques durs : on ne peut pas tout avoir... Comme il existe des programmes de conversion d'un format dans l'autre, les utilisateurs potentiels du chinois peuvent se rassurer : ils réussiront peut-être à conserver quelques octets disponibles sur leur disque dur[8] !

L'utilisation des fichiers `.tfm` et `.pxl`

Un petit mot de rappel sur la manière dont travaille TEX : à partir du fichier source produit par l'utilisateur (fichier ASCII, 7 bits/caractère ou 8 bits/carac-

[5] Massachusset Institute of Technologie.

[6] Elles ne sont pas du domaine public, mais ce sont les plus belles...

[7] *en chinois dans le texte*, bien sûr.

[8] si du moins ils n'ont pas, comme moi, la malchance d'avoir une visionneuse sur écran qui n'accepte que des polices `.pk` et un pilote d'imprimante qui ne jure lui que par les `.pxl`. Je dispose heureusement d'un bon disque dur.

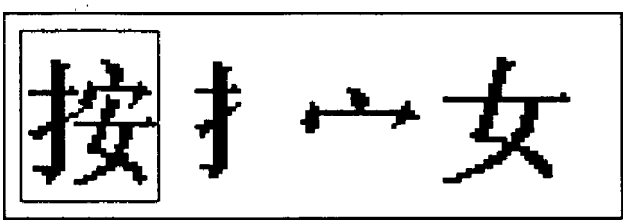

Figure 1: Le caractère *àn* et ses trois radicaux.

tère avec la nouvelle version 3.0), la mécanique interne de TEX — son appareil digestif dirait Knuth — produit un fichier `.dvi` qui décrit un document virtuel, formé de boîtes dimensionnées, imbriquées ou juxtaposées, pour composer les pages (corps de pages, en-tête et pied de page), les paragraphes, les lignes et enfin les caractères qui composent ces dernières. Pour chaque boîte, TEX n'a besoin que de ses dimensions et de la valeur symbolique de son contenu, par exemple la lettre «A» dans la fonte courante si cette boîte doit contenir ce signe typographique : c'est le rôle des fichiers `.tfm` ; ils fournissent à TEX les dimensions dont celui-ci a besoin pour construire ses boîtes.

Le dessin du caractère lui-même, c'est à dire la matrice des pixels blancs ou noirs, ne sera fourni qu'au programme (pilote d'imprimante ou d'écran) qui transformera le document virtuel au format DVI en bon papier (ou écran) bien écrit et lisible par n'importe qui. Bref, du point de vue des dimensions globales, outre les informations propres à chaque caractère, un fichier `.tfm` n'a besoin que du corps (`design_size`) dans lequel cette police sera utilisée et du facteur d'agrandissement (`magnification`) utilisé au moment de sa génération à partir de METAFONT. On ne peut qu'admirer ici la hauteur de vue dont TEX fait preuve en appliquant ainsi le principe « l'intendance suivra » !

L'intendance, c'est justement ce dont sont chargés, via le pilote d'imprimante, les fichiers `.pxl` (ou `.pk`, qui sont la même chose dans un format compressé). À eux de fournir les bons dessins aux bonnes dimensions quand on le leur demande ! Dépendant du dispositif d'impression terminal, ils sont propres à celui-ci, ou, plus exactement, à sa résolution ; un carré de 12 points[9] de côté ne requiert pas le même nombre de pixels pour être affiché à 300 ppp[10] ou à 72 ppp. À l'inverse, une matrice de 28 × 28 pixels correspond respectivement à un carré de 2,37 mm de côté sur une imprimante laser à 300 ppp, et de 9,88 mm sur un écran à 72 ppp. Pour afficher le même caractère, en «vraie grandeur», on a besoin de deux fichiers `.pxl` (ou `.pk`) différents, utilisables chacun par le pilote *ad-hoc*.

Avant de passer à la description détaillée de ces fichiers et de leur construction, il est bon de faire un petit calcul de la place requise par les fontes que l'on s'apprête à construire ; en ne prenant en compte que celles qui sont nécessaires à une imprimante laser (300 ppp — certains programmes d'épreuvage à l'écran étant capables de les utiliser au prix d'une qualité douteuse) et en se limitant aux corps **6**, **7**, **8**, **12** et **14**[11], on obtient

[9] pt, au sens dimension de TEX.

[10] ppp : pixels par pouce.

[11] Ces corps ne sont pas choisis au hasard : un corps de 7 (pt) correspond à des lettres de hauteur totale 2,47 mm ; c'est à peu près ce que donne une matrice de 28 x 28 pixels. Comme nous disposons de

les chiffres suivants : en format PXL, il faut respectivement 550 ko pour le corps 6, 750 ko pour le 7, 1 Mo pour le 8, 2,2 Mo pour le 12 et 3 Mo pour le 14, soit au total 7,5 Mo. Si on veut ajouter un style différent, par exemple l'équivalent du **gras**, on peut l'obtenir par un facteur d'agrandissement horizontal supérieur au facteur d'agrandissement vertical : 4,5 Mo supplémentaires pour des caractères gras en 8, 12 et 14 points sont alors nécessaires. Au total, 12 Mo, que l'on peut heureusement ramener à un peu plus de 3 Mo en transformant tous ces fichiers `.pxl` en fichiers `.pk`.

La création des fichiers .pxl et .tfm

Le programme de création des fichiers de polices chinoise travaille automatiquement ; plusieurs paramètres étant communs aux deux fichiers, il crée les deux fichiers simultanément, mais, pour des raisons de commodité, je vais présenter leur création séparément.

Les fichiers `.tfm`

Divisés en trois parties, la première contient douze entiers dans six mots. Les deux premiers entiers, `lf` et `lh`, longueurs en mots du fichier et de l'en-tête, sont remplis automatiquement par le programme de création. Les deux suivants, `bc` et `ec`, index du premier et du dernier caractère disponibles, sont égaux respectivement à 0 et 127 : tous les caractères sont utilisés.[12]

Sur les sept entiers suivants, seuls les trois premiers (nombre de largeurs `nw`, hauteurs `nh` et profondeurs `nd` des caractères) nous intéressent : les tables dont ils indiquent les tailles contiennent chacune deux valeurs, la première étant toujours 0 et la deuxième la valeur effective : `nw`, `nh` et `nd` sont donc égaux à 2 ; les quatre valeurs suivantes sont mises à 0, les tables correspondantes n'existant pas dans les polices chinoises.

Le dernier demi-mot de l'en-tête contient le nombre de paramètres finaux du fichier `.tfm` : il est au moins égal à 7 ; c'est la valeur effectivement retenue ici.

A la suite de cette première partie viennent en principe dix tableaux de données ; ils ne sont que six ici, les tableaux 6, 7, 8 et 9 ayant été déclarés de longueur nulle dans l'en-tête ; ce sont respectivement :

`header` : tableau de douze mots ; le premier contient un entier qui est mis à 0 : aucun contrôle de conformité ne sera effectué entre les fichiers `.tfm` et `.pxl` ; le deuxième mot contient le corps, en `FIX`es, de la police pour lequel elle a été conçue et dessinée ; pour un corps de 14, on prendra donc comme valeur (entière) 14×2^{20}. Les dix mots suivants peuvent contenir n'importe quoi.

`char_info` : tableau de 128 mots, à raison de un mot pour chacun des caractères de la police ; pour le chinois, ce mot est le même pour tous, les quatre octets le composant étant :

- *octet 1 :* (index de la largeur du caractère dans le tableau `width`) : tou-

caractères chinois dans des matrices de 16 x 16, 24 x 24 et 28 x 28, le corps 6 correspond aux matrices 24 x 24 sans facteur d'expansion, le 12 aux mêmes agrandies deux fois, les corps 7 et 14 aux matrices 28 x 28 dans les mêmes conditions et le corps 8 aux matrices 16 x 16 doublées (ie corps 4 serait à peu près illisible). L'absence des corps 10 et 11 est évidemment des plus ennuyeuse, mais on se limite à ce qui est le plus simple et disponible.

[12] En théorie — et sans doute en pratique — on peut utiliser 256 caractères par police. Par une bizarrerie sur laquelle je n'ai pas encore eu le temps de me pencher, le programme `pxtopk` que j'utilise pour fabriquer les fichiers `.pk` à partir des `.pxl` ne veut travailler correctement que sur des fichiers contenant au plus 128 caractères ; il ne devrait pas être trop difficile de retrouver dans le programme de conversion la variable en cause et de la modifier.

jours égal à 1, tous les idéogrammes ayant la même largeur.

- *octet 2 :* (index, codés chacun sur un demi octet, des hauteurs et profondeurs dans les tableaux **height** et **depth**) : leur valeur étant également 1, il contient la valeur binaire 00010001, soit 17.
- *octets 3 et 4 :* ils sont mis à 0, les index qu'ils contiennent correspondant à des tableaux qui n'existent pas pour les polices chinoises.

width, height, depth : trois tableaux composés de deux mots chacun ; comme on l'a vu, les deux valeurs possibles sont toujours 0 pour la première et la valeur effective, en **pt**, de la dimension correspondante pour la seconde ; cette valeur étant une valeur réelle (au sens mathématique, non-entière), on mettra dans le mot la partie entière de son produit par 2^{16} conformément à la représentation des réels dans TEX ;

italic, lig_kern, kern et **exten** : ces tableaux n'existent pas pour les fontes chinoises.

param : dernier tableau du fichier **.tfm**, il est formé de sept mots :

- *mot 1 :* (décalage relatif dû à l'inclinaison) pas d'inclinaison, égal à 0.
- *mot 2 :* (espace entre les mots : valeur **réelle** en **pt**) La valeur retenue est égale au quart de la largeur d'un idéogramme.
- *mot 3* et *4 :* ces deux valeurs sont mises à 0 (pas d'expansion ou de compression de l'espace inter-mot, pour conserver un alignement vertical des idéogrammes des lignes chinoises successives).
- *mot 5 :* sans utilité ici, mis à 0.
- *mot 6 :* (valeur du cadratin (**\quad**) égal à la largeur d'un caractère.
- *mot 7 :* (espace additionnel en fin de phrase) égal au cadratin.

Ce dernier paramètre, tout comme le précédent, n'est pas réellement utilisé par le préprocesseur ; ils sont mis là «au cas où» ...

Les fichiers **.pxl**

Pour le fichier **.pxl**, les choses sont plus simples, puisqu'il doit contenir essentiellement les glyphes des caractères sous forme de matrices de points. Composés lui aussi de mots de 32 bits, il comprend les quatre parties suivantes (les exemples numériques sont donnés pour la création d'une police dans le corps 14, avec une matrice de 56 × 56 pixels) :

- un premier mot contenant l'identificateur de contrôle **pxl_id** : 1001 (**entier**) ;
- une suite de mots contenant les glyphes des 128 caractères de la police ;
- le répertoire de la fonte composé de 128 blocs de quatre mots ; on a dans chacun de ces quatre mots :

  mot 1 : (largeur l et hauteur h (en pixels) du glyphe) : 56 (**entier**) pour le corps **14**.

  mot 2 : position en pixels du point de référence : déplacement horizontal dans le premier demi-mot, égal à 0, déplacement vertical dans le deuxième, égal à 42 ($\simeq 56 \times 0,75$, selon le choix fait dans le fichier **.tfm**) soit 42 (**entier**).

  mot 3 : l'adresse du début du dessin du caractère est calculée par le programme de génération.

mot 4 : (valeur de la largeur, en pt, de la boîte contenant le caractère) : 15,3 (**réel**).

- la dernière partie, formée de cinq mots, contient :

  checksum : (code de contrôle) : 0 , pas de contrôle.

  magnification : (facteur d'agrandissement de la fonte) 1000 (**entier**), agrandissement égal à 1.

  design_size : (corps de la fonte) : 14,0 , (**réel**).

  dir_ptr : adresse du premier mot du répertoire de la fonte, fournie par le programme.

  pxl_id : (code de contrôle) : 1001 .

L'écriture du programme de génération des fontes chinoises n'est plus alors qu'une question de programmation ; bien sûr, beaucoup de chose sont assez arbitraires, comme le nom des fontes, la position du point de référence ou le corps à leur attribuer : j'ai ainsi considéré qu'un caractère chinois devait avoir une hauteur à peu près égale à la hauteur maximum d'une majuscule au dessus de la ligne de référence, et descendre un peu moins que le jambage maximun d'une police latine pour s'accorder avec cette dernière. Voici à titre d'exemple dans la figure (2) ce que donne une fonte «normale» de 56 × 56 pixels (on a ici les 128 premiers caractères du code GB 2312-80) :

Comment écrire en chinois ?

Cette question qui peut paraître saugrenue mérite cependant quelques explications : dans la calligraphie chinoise, le sens traditionnel d'écriture est de haut en bas et de droite à gauche ; le fait pour cette écriture d'être composée d'idéogrammes porteurs de sens a pour conséquence que les mots y sont composés de peu de caractères : la plupart d'entre eux sont formés de un ou deux idéogrammes, parfois quatre ; la reconaissance individuelle de chaque idéogramme a comme conséquence que le sens de parcours du texte écrit pèse moins sur la vitesse de lecture et la compréhension que dans le cas de nos langues alphabétiques où elles reposent en grande partie sur la reconnaissance globale de groupes de lettres, et donc sur une habitude visuelle liée au sens d'écriture.

Le découpage en lignes

Il n'y a pas en chinois, même pour des mots composés de plusieurs caractères, de marque typographique indiquant une liaison ou inversement une séparation entre ceux-ci ; une coupure de ligne (ou de colonne en cas d'écriture verticale) peut apparaître n'importe où entre deux idéogrammes. Lorsque plusieurs lignes se suivent, la typographie chinoise a pour habitude de respecter également un alignement vertical entre les idéogrammes : question d'esthétique sans doute, mais aussi facilité de lecture ; dans le même but, les signes typographiques de ponctuation occupent, en chinois moderne[13], la même largeur que les idéogrammes. C'est pour respecter autant que faire se peut cet alignement que les fontes chinoises sont générées avec une valeur de l'espace inter-mot égale à 0, sans expansion (ni compression) possible. Le préprocesseur insère ainsi après chaque code un espace qui a pour seul rôle d'autoriser une césure à cet endroit ; chacun d'entre eux est considéré comme un mot. La mécanique interne de TEX peut ainsi jouer pleinement, les césures s'effectuant toujours entre deux mots. Ceci conduit

[13] ces signes sont d'usage récent et importés d'occident ; le chinois classique ignorait tout découpage, y compris celui des phrases.

	'0	'1	'2	'3	'4	'5	'6	'7	
'00x	啊	阿	埃	挨	哎	唉	哀	皑	"0x
'01x	癌	蔼	矮	艾	碍	爱	隘	鞍	
'02x	氨	安	俺	按	暗	岸	胺	案	"1x
'03x	肮	昂	盎	凹	敖	熬	翱	袄	
'04x	傲	奥	懊	澳	芭	捌	扒	叭	"2x
'05x	吧	笆	八	疤	巴	拔	跋	靶	
'06x	把	耙	坝	霸	罢	爸	白	柏	"3x
'07x	百	摆	佰	败	拜	稗	斑	班	
'10x	搬	扳	般	颁	板	版	扮	拌	"4x
'11x	伴	瓣	半	办	绊	邦	帮	梆	
'12x	榜	膀	绑	棒	磅	蚌	镑	傍	"5x
'13x	谤	苞	胞	包	褒	剥	薄	雹	
'14x	保	堡	饱	宝	抱	报	暴	豹	"6x
'15x	鲍	爆	杯	碑	悲	卑	北	辈	
'16x	背	贝	钡	倍	狈	备	惫	焙	"7x
'17x	被	奔	苯	本	笨	崩	绷	甭	
	"8	"9	"A	"B	"C	"D	"E	"F	

Figure 2: la police *hzaa14.1500pxl*

évidemment, pour des lignes chinoises pleines, à des **overfull** ou **underfull \hbox** ... qu'il serait possible d'éliminer en calculant judicieusement les longueurs de ligne ou en insérant des «ressorts» en fin de ligne quand elles sont pleines[14].

[14] Je n'ai pas encore d'idées sur la façon de réaliser cela, et me demande même si c'est nécessaire dans les conditions de fonctionnement que je me suis fixées ; l'insertion automatique d'un tel «ressort» à chaque changement de langue doit résoudre la plupart des cas.

L'écriture verticale

La question de l'écriture verticale (et de droite à gauche) d'un texte comporte deux aspects qui n'entrent pas dans le cadre de cet article et pour lesquels je me contenterai de donner quelques idées générales. Le premier est celui d'un texte entièrement composé de cette manière ; on peut alors considérer qu'il y a une similitude profonde avec l'écriture horizontale, au moins dans les principes de découpage et de présentation en pages,

paragraphes et lignes : la dimension fondamentale est ici la hauteur de page, le haut de page jouant le rôle de la marge gauche pour les textes horizontaux. Le deuxième aspect est celui de l'insertion d'un paragraphe[15] de texte «vertical» dans un texte «horizontal» ; la dimension fondamentale de référence est maintenant la largeur de la ligne alors que l'écriture verticale requiert une hauteur pour pouvoir effectuer ses césures. Des considérations esthétiques interviennent alors : il faut que le texte soit équilibré en largeur, autrement dit, que cette hauteur ne soit pas telle qu'il y ait un grand espace vide à gauche d'une colonne verticale. La hauteur de la boîte verticale est ainsi fonction de la longueur du texte à inclure ; on voit le genre de problèmes à résoudre dans ce cas.

Le préprocesseur

Comment passer du code à deux octets d'un idéogramme aux commandes TEX qui l'impriment ? Essentiellement en appelant la police qui le contient, puis en fournissant son index dans cette dernière. C'est finalement tout ce que fait le préprocesseur : il calcule le nom de la police, `hz`*xy* par exemple, puis le rang de l'idéogramme dans cette police, soit *abc* et génère alors la commande {`\hz`*xy*{`\char`*abc*} } ; le fonctionnement réel est en fait un peu plus complexe pour tenir compte des ponctuations chinoises : l'espace entre les deux '}' finaux n'est inséré que si le code chinois suivant n'est pas un code de ponctuation, ceci pour éviter une césure entre les deux idéogrammes.

[15] le cas de plusieurs paragraphes (qui pourraient s'étendre sur plus d'une page) ne se pose pas : comment faudrait-il tourner les pages dans ce cas ?

Le nom des polices

Alors qu'une police ordinaire contient tous les signes typographiques utilisés dans un jeu de lettres et symboles donnés, plusieurs polices sont nécessaires pour un seul jeu chinois dans un même style et corps. Leurs noms doivent à la fois les distinguer, indiquer leur appartenance au même jeu et se prêter le mieux possible au bon fonctionnement du préprocesseur. Le nom est formé d'une base qui indique le corps et le style, et d'une extension pour le numéro de la police. J'ai ainsi choisi de les appeler `hz`*xyln*`.tfm`, `hz` pour hanzi[16], avec les variantes suivantes : `Hz` indique un idéogramme plus haut que large, `hZ` plus large que haut (deux formes de mise en relief, la seconde étant une sorte de gras). L'extension *xy* est composée de deux lettres (soit `aa`, `ab`, ..., `ca` pour les 53 polices du jeu *hanzi simplifié*) et de deux chiffres *ln* qui indiquent le corps nominal de la fonte. Ces noms sont générés automatiquement par le programme de création des fontes chinoises.

[16] *hànzì* : idéogramme en chinois.

ANNEXE

Types, dimensions et unités en TEX

TEX est riche en de nombreux domaines, et parmi ceux-ci, les dimensions (et les unités correspondantes) ne sont pas les plus pauvres.[17] Commençons donc par le début : TEX connait les unités de longueur courantes : centimètre (`cm`), pouce (`in`), *etc...* En typographie, l'unité de base est le *point* ; TEX utilise donc le point comme unité de longueur ; la seule différence entre le point TEXien et le point typographique est que, comme le dit Knuth, «les unités de mesure de TEX sont fermement fondées sur le système métrique (...)» ; moyennant quoi, 72,27 `pt` = 1 `in` et 1 `in` = 2,54 `cm`. J'utiliserai les notations `point` et `pt` pour indiquer qu'il s'agit de points au sens de TEX. Dans tous ses calculs internes, TEX utilise des valeurs entières : pour représenter un nombre réel, il utilise une représentation binaire de ce nombre dont les seize bits de poids faible constituent la partie décimale ; il appelle cette représentation des `scaled points`. Si on a besoin de donner dans un fichier, par exemple, une valeur égale à 7,0, il faudra fournir la valeur (entière) $\mathcal{E}(7,0 \times 2^{16})$[18]. Pour définir une police nous avons besoin de son *corps* : c'est en gros la taille des lettres où elle rendra le mieux dans un texte ; TEX appelle cette taille (qui est un nombre réel compris entre 1,0 et 256,0) le `design_size`[19] et utilise une autre unité pour la mesurer de façon interne : c'est le `FIX` ; 1 `FIX` est égal au corps nominal (`design_size`) multiplié par 2^{-20}. Le `FIX` est ainsi une unité relative au corps nominal ; c'est dans cette unité que seront exprimées les dimensions diverses d'un caractère telles que la largeur, la profondeur (des jambages), la hauteur au dessus de la ligne de base, *etc...*

[17] La recherche des informations à ce sujet est d'ailleurs un vrai jeu de piste à travers les divers manuels, les notices des utilitaires, certains articles du TUGboat ou les comptes rendus des conférences annuelles du TEX Users Group.

[18] en notant $\mathcal{E}(x)$ la partie entière du réel x.

[19] c'est plutôt de *taille nominale* qu'il faudrait parler, car, par le mécanisme des `\font\xyz=...scaled ...` ou `\font\xyz=...at ...`, on peut obtenir un autre corps que le corps de création.

Références bibliographiques

[1] Cheng François : *L'écriture poétique chinoise*. Seuil éditeur, 1985.

[2] Cousquer Alain, Cousquer Éliane : *Informatisation du chinois*. AFCET/Interfaces N° 64, 1988.

[3] Tung Yunmei : *LCCD, A Language for Chinese Character Design*. SOFTWARE - PRACTICE AND EXPERIENCE, Vol 11, 1981.

[4] Hobby John, Gu Guoan : *A chinese MetaFont*. Tugboat Vol. 5, n° 2, 1984.

[5] Li Jiarong : *Generation of some chinese character with MetaFont*, TEX for scientific documentation. pp.161,170. Addison-Wesley editeur, 1985.

[6] Hosek D. : *Design of oriental characters with MetaFont*. Tugboat Vol. 10, n° 4, 1989.

[7] Knuth D.E. : *TEX : The program*. Computers and typesetting, Vol. B. Addison-Wesley ed. 1986.