Cahiers GUTenberg

CONVENTIONS CONCERNANT LES POLICES DC ET LES LANGUES NATURELLES

Yannis HARALAMBOUS

Cahiers GUTenberg, nº 15 (1993), p. 53-61.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1993___15_53_0>

Conventions concernant les polices DC et les langues*

Yannis HARALAMBOUS

Chair, Technical Working Group on Multiple Language Coordination† *(***WG-92-03***)*
187, rue Nationale, F–59800 Lille, France
email: yannis@gat.citilille.fr*, fax: +33 20 40 28 64*

... chaque langue trace
autour des hommes à qui elle appartient
un cercle magique d'où ils ne peuvent sortir
que pour tomber dans un autre.
Marshall MCLUHAN[1], *La galaxie Gutenberg* (1962)

Que sont les polices DC ?

Quand j'ai commencé à l'utiliser pour composer ma thèse, j'ai remarqué que TEX sortait un nombre démesuré de marques noires indiquant des dépassements de marge. En essayant de percer ce mystère, je me suis aperçu que TEX (en réalité, le bon vieux *Textures* version 1.0) n'était pas très coopératif en matière de césure ; plusieurs mots n'étaient jamais coupés ! Une recherche plus approfondie dans le *TEXbook* a clarifié la situation : la primitive `\accent` – depuis devenue illustre... – qui faisait partie de l'expansion des macros d'accentuation introduisait un crénage explicite qui à son tour bloquait la procédure de césure.

Il fallait donc utiliser des caractères déjà accentués plutôt que superposer les signes 'caractère' et 'accent' par le biais de la primitive `\accent`. Cela est possible depuis la sortie de TEX 3 qui permet une saisie à 8 bits, et donc l'utilisation de polices autres que *Computer Modern.* Bien-sûr, il a fallu

*. Cet article est paru dans *TEX and TUG NEWS (TTN)*, vol.1, n° 4, décembre 1992, pages 3–10, sous le titre *TEX Conventions Concerning Languages.* Il est reproduit ici avec l'aimable autorisation de Christina THIELE, rédactrice en chef de *TTN* et de l'auteur, Yannis HARALAMBOUS, qui en a assuré lui-même la traduction française. {Ndlr}

†. Groupe technique sur la coordination multilingue – voir l'appendice de cet article.

1. En réalité citant Wilhelm VON HUMBOLDT, et sa « stratégie qu'une culture doit suivre dans une semblable conjoncture ».

définir une police 8-bits « à la *Computer Modern*» pour maintenir la tradition de compatibilité et consistence interne de TEX. Jan Michael RYNNING et Norbert SCHWARZ entreprirent la tâche d'établir une nouvelle table de police. Lors de la conférence de Cork en 1990, le TUG et la communauté TEX internationale adoptèrent cette table comme le nouveau codage standard de sortie TEX, baptisé « codage DC », « codage de Cork » ou « codage TEX étendu »[2]. Norbert SCHWARZ écrivit le code d'une famille de polices similaires à *Computer Modern*, et les appela *polices DC*.

Il faudrait peut-être souligner le fait que Norbert ne considère nullement ce code comme définitif – de petites corrections de forme de caractère peuvent encore être nécessaires[3] ; néanmoins, le codage ne changera plus et la communauté TEX internationale est invitée à commencer la migration du codage *Computer Modern* décrit dans les cinq volumes de *Computers and Typesetting*, vers le nouveau codage DC montré dans la figure 1.

Cette migration consistera à

1. ajouter les nouvelles polices à votre système (les polices *Computer Modern* sont encore indispensables, à cause des lettres grecques majuscules : celles-ci ne font pas partie du codage texte DC – elles seront incluses dans les polices DC mathématiques (à paraître) ;

2. remplacer `plain.tex`, `lplain.tex` par les nouvelles versions ad hoc (le système NFSS, hautement recommandé, est d'ores et déjà compatible avec les polices DC : utiliser les fichiers `fontdef.dc`, `preload.dc` et `dclfont.tex` lors de la compilation du format) ;

3. faire attention, quand on écrit du code TEX, à n'utiliser que des macros de haut niveau pour certaines tâches qui font intervenir le codage de caractères : `\'a`, `\oe`, `$\Sigma$` produiront toujours le résultat souhaité : à, œ, Σ (par contre le code jadis équivalent `\accent'022a`, `\char'033`, `\char"06` produira à, œ, Σ dans un environnement de polices DC mais a¸ , ff, ° dans un environnement DC.

2. *TUGboat* 11 (1990), no. 4, pp. 514–516; *Cahiers GUTenberg*, n° 7, novembre 1990, 29–31.

3. Notamment en ce qui concerne la hauteur de certains accents.

TAB. 1 - : *Fonte* `dcr10`

	'0	*'1*	*'2*	*'3*	*'4*	*'5*	*'6*	*'7*
'00x	`	´	ˆ	˜	¨	˝	˚	ˇ
'01x	˘	¯	˙	¸	˛	,	≺	≻
'02x	“	”	„	«	»	–	—	
'03x	0	ı	j	ff	fi	fl	ffi	ffl
'04x	␣	!	"	#	$	%	&	’
'05x	(	)	*	+	,	-	.	/
'06x	0	1	2	3	4	5	6	7
'07x	8	9	:	;	<	=	>	?
'10x	@	A	B	C	D	E	F	G
'11x	H	I	J	K	L	M	N	O
'12x	P	Q	R	S	T	U	V	W
'13x	X	Y	Z	[	\	]	ˆ	_
'14x	‘	a	b	c	d	e	f	g
'15x	h	i	j	k	l	m	n	o
'16x	p	q	r	s	t	u	v	w
'17x	x	y	z	{	|	}	˜	¨
'20x	Ă	Ą	Ć	Č	Ď	Ě	Ę	Ğ
'21x	Ĺ	Ľ	Ł	Ń	Ň	Ŋ	Ő	Ŕ
'22x	Ř	Ś	Š	Ş	Ť	Ţ	Ű	Ů
'23x	Ÿ	Ź	Ž	Ż	IJ	İ	đ	§
'24x	ă	ą	ć	č	d’	ě	ę	ğ
'25x	Í	l’	ł	ń	ň	ŋ	ő	ŕ
'26x	ř	ś	š	ş	t’	ţ	ű	ů
'27x	ÿ	ź	ž	ż	ij	¡	¿	£
'30x	À	Á	Â	Ã	Ä	Å	Æ	Ç
'31x	È	É	Ê	Ë	Ì	Í	Î	Ï
'32x	Đ	Ñ	Ò	Ó	Ô	Õ	Ö	Œ
'33x	Ø	Ù	Ú	Û	Ü	Ý	Þ	SS
'34x	à	á	â	ã	ä	å	æ	ç
'35x	è	é	ê	ë	ì	í	î	ï
'36x	ð	ñ	ò	ó	ô	õ	ö	œ
'37x	ø	ù	ú	û	ü	ý	þ	ß

Recommendations à propos de certains noms de macros et de fichiers concernant les langues

Quelques mots

Toutes les macro-commandes qu'on trouve dans `plain.tex` sont définies et décrites dans le *TEXbook*, la bible de TEX. Ces macros font partie de TEX ; cela garantit la portabilité de TEX. Les polices DC nécessitent de nouvelles macros, qui à leur tour doivent être standardisées.

Ces macros concernent les caractères accentués et spéciaux qui ne sont pas disponibles dans les polices *Computer Modern.*

Outre les macros, il faut aussi standardiser les noms des fichiers de motifs de césure et d'exceptions de césure. Pour cela, il faut un préfixe indiquant la langue et que ce préfixe soit facile à apprendre, court, et universellement acceptable. Pour établir la liste que le lecteur trouvera ci-dessous (table 2), on s'est basé sur le code ISO 639[4], qui définit les langues, en deux lettres. En particulier, le fichier `hyphen.tex` de Fred LIANG, sera dorénavant appelé `ushyphen.tex`.[5]

4. Avec les exceptions suivantes: `us` et `gb` ont été choisis pour l'anglais américain et britannique; `sb` pour le sorabe et `se` pour le lappon. Les codes `la` (latin), `sa` (sanskrit), `en` (anglais), et `sh` (serbocroate) ne sont pas utilisés.

5. À ce jour (avril 1993), des motifs de césure existent pour les langues suivantes (β signifie qu'ils sont encore en tests) : l'allemand, l'anglais britannique et l'anglais américain, l'arménien, le bulgare (β), le catalan, le croate, le danois, l'espagnol, l'espéranto, l'estonien, le finnois, le français, le grec (moderne et ancien), le hongrois, l'islandais, l'italien, le kirundi (β), le latin, le lithuanien, le néerlandais, le norvègien, le polonais, le portugais, le russe, le slovaque, le suédois et le tchèque; les motifs de césure souahili et yiddish sont en préparation.

Macros pour caractères accentués et spéciaux

À part les caractères accentués et spéciaux définis dans le *TEXbook* (p. 52, 339) les macros suivantes sont introduites :

tapez	*pour obtenir*	(*autrement dit*)
`\r u`	ů	(l'accent rond tchèque)
`\k e`	ę	(le signe ogonek polonais)
`\v d \v D`	d' Ď	(la lettre tchèque d D avec hačekk)
`\v t \v T`	t' Ť	(la lettre tchèque t T avec hačekk)
`\v l \v L`	l' L'	(la lettre slovaque l L avec hačekk)
`\th \TH`	þ Þ	(le thorn islandais)
`\dh \DH`	ð Ð	(le eth islandais)
`\dj \DJ`	đ Đ	(le dj serbocroate)
`\ng \NG`	ŋ Ŋ	(le ng lappon)
`<< >>`	« »	(les guillemets français)
`,, ‘‘`	„ “	(les Gänsefüßchen allemands)

Notes

– Le caractère néerlandais ij IJ (qui se prononce « aye »), utilisé par exemple dans le petit poème enfantin :

Blijf, wijl 'k mijn tijd
Blij 't ij-rijm wijd,
Blijf, blijf mij bij, gij IJ,
Stijf vrij, wijl 'k lijm,
Mijn ij-rijk rijm;
Mijn wijs, mijn prins zijt gij

est une ligature et, au même titre que « ffi » ou « ffl », il faut y accéder directement ; à l'aide de polices nérlandaises spéciales (par exemple virtuelles) TEX pourra composer cette ligature automatiquement.

– Le caractère scandinave å Å peut être obtenu indifféremment par la commande `plain.tex` standard `\aa \AA` ou par `\r a \r A`.

– La macro `\c` sert aussi bien pour la cédille française, catalane, lettonne, turque et roumaine, que pour la lettre lettonne ģ, qui est surmontée d'une cédille inversée : `\c g` pour 'ģ'.

– Les lettres turques ı I, i İ sont obtenues par `\i I`, `i \.I`.

– Les guillemets français et allemands « » ” “ *ne nécessitent pas l'utilisation de macro !* Ils sont obtenus par des ligatures, tout comme les guillemets américains “ ”.[6]

Préfixes de langues

La table 2 indique les codes à deux lettres qui correspondent aux langues les plus importantes.

Notes

– Les langues ont été classées selon les trois critères suivants : (a) motifs de césure, (et valeurs de `left-` et `righthyphenmin`), (b) nécessité de polices particulières et (c) direction d'écriture.

  Ainsi par exemple, l'anglais britannique nécessite des motifs de césure différents de l'anglais américain ; le lithuanien, le letton, le gallois et l'espéranto nécessitent des polices virtuelles qui peuvent être basées sur les polices DC ; le vietnamien et la plupart des langues africaines nécessitent d'autres polices latines ; le grec et le russe nécessitent des polices spéciales ; l'arabe et l'hébreu sont écrits de droite à gauche, etc.

  Si une différenciation plus poussée est souhaitée (par exemple pour distinguer le néerlandais du flamand, le portuguais du brésilien, l'allemand de l'autrichien ou du suisse allemand etc.) *toujours basée sur ces critères*, alors on pourra ajouter des codes supplémentaires.[7]

– Malgré les déchirures politiques, le serbe et le croate ne sont en fait qu'une seule langue (le « serbocroate ») ; cette langue possède deux variantes : l'écave et le yécave. La première est plus utilisée en Serbie et écrite en alphabet cyrillique, alors que la seconde est utilisée en Croatie et écrite en alphabet latin. Néanmoins cette classification est très grossière. Dans notre contexte, les macros `\SR` et `\HR` signifieront respectivement : écave écrit en cyrillique et yécave écrit en latin.

6. Il n'est question ici que d'accéder aux glyphes des guillemets français et/ou allemands. La typographie française traditionnelle française veut que l'utilisation des guillemets affecte la mise en page globale (par exemple en insérant un guillemet ouvrant à chaque début de paragraphe de la citation, en chaque début de ligne de citation de deuxième ordre). Dans ce cas on est bel et bien obligé de passer par des macros, comme cela est réalisé dans le style `french.sty` de Bernard Gaulle.

7. Par contre si les différences se situent, par exemple, au niveau de la ponctuation, alors la différenciation peut être faite au niveau des macros (La)TEX.

TAB. 2 - : *Codes pour préfixer les principales langues*

Abkhazian	ab	Féroïen	fo	Letton	lv	Singhalais	si
Afan oromo	om	Fidji	fj	Lingala	ln	Siswati	ss
Afar	aa	Finnois	fi	Lithuanien	lt	Slovène	sl
Afrikaans	af	Français	fr	Macédonien	mk	Slovaque	sk
Albanais	sq	Frison	fy	Malais	ms	Somali	so
Allemand	de	Géorgien	ka	Malayalam	ml	Sorabe	sb
Amharique	am	Galicien	gl	Malgache	mg	Swahili	sw
Anglais américain	us	Gallois	cy	Maltais	mt	Soundanais	su
Anglais britannique	gb	Guarani	gn	Maori	mi	Suédois	sv
Arabe	ar	Goujrati	gu	Marathe	mr	Tagal	tl
Arménien	hy	Grec	el	Moldave	mo	Tadjik	tg
Assamais	as	Groenlandais	kl	Néerlandais	nl	Tamoul	ta
Aymara	ay	Hébreu	he	Népalais	ne	Tatar	tt
Azéri	az	Haoussa	ha	Nauri	na	Tchèque	cs
Bashkir	ba	Hindi	hi	Norvégien	no	Télougou	te
Basque	eu	Hongrois	hu	Occitan	oc	Thaï	th
Bengali	bn	Indonésien	id	Oriya	or	Tibétain	bo
Bhoutani	dz	Inuktitut	iu	Ouigoure	ug	Tigrinya	ti
Bihari	bh	Inupiak	ik	Ourdou	ur	Tonga	to
Birman	my	Interlingua	ia	Pashto	ps	Tsonga	ts
Bislama	bi	Interlingue	ie	Persan	fa	Turc	tr
Breton	br	Irlandais	ga	Polonais	pl	Turkmène	tk
Bulgare	bg	Islandais	is	Portugais	pt	Tchi	tw
Biélorusse	be	Italien	it	Pendjabi	pa	Ukrainien	uk
Cambodgien	km	Japonais	ja	Quechua	qu	Uzbéki	uz
Catalan	ca	Javanais	jw	Rhéto-roman	rm	Vietnamien	vi
Chinois	zh	Kannara	kn	Roumain	ro	Volapük	vo
Coréen	ko	Kashmiri	ks	Russe	ru	Wolof	wo
Corse	co	Kazakh	kk	Samoan	sm	Xhosa	xh
Croate	hr	Kinyarwanda	rw	Sango	sg	Yidich	yi
Danois	da	Kirghiz	ky	Serbe	sr	Yorouba	yo
Écossais	gd	Kiroundi	rn	Sesotho	st	Zhouang	za
Espéranto	eo	Kurde	ku	Setchwana	tn	Zoulou	zu
Espagnol	es	Laotien	lo	Shona	sn		
Estonien	et	Lappon	se	Sindhi	sd		

– Par `\HE` on entend le "Ivrit" (hébreu moderne). Il est écrit dans le même alphabet que le yiddish[8] mais possède des régles de césure différentes.

– La translitération et la césure de l'arménien suivent la prononciation « de l'ouest », utilisée par les linguistes et par les communautés arméniennes de l'Europe et des États-Unis. L'utilisation de polices virtuelles est recommandée pour la saisie d'arménien translitéré d'après la prononciation « de l'est ».

– Cette liste ne couvre que les langues vivantes. L'utilisation de codes à deux lettres pour les langues classiques, telles que latin, grec ancien, hébreu classique, araméen, syriaque, sanscrit, turc ottoman etc., n'est pas recommandée. Vous êtes invité(e) à contacter le Groupe Technique de Coordination Multilingue pour de plus amples informations.

Noms de fichiers de césure

Pour composer le nom d'un fichier de césure on utilisera le code à deux lettres ci-dessus, suivi de `hyphen.tex`.

Ainsi par exemple, les fichiers de césure français, allemand, britannique, seront nommés respectivement `frhyphen.tex`, `dehyphen.tex`, `ukhyphen.tex`.

Le fichier anglais américain standard sera dorénavant nommé `ushyphen.tex`.

Appendice

Cet article a été publié dans *TEX and TUG News* (TTN) en décembre 1992. Entre temps les travaux du TWGMLC (Groupe technique sur la coordination multilingue) ont avancé : plus de 25 langues y sont aujourd'hui représentées. Le but de TWGMLC est de préparer pour chaque langue un TLP *(TEX Language Package)* qui consistera en :

– « matériel TEX3 » : motifs de césure ;

– « matériel LATEX3 » : style linguistique ; puisque LATEX3 n'est pas encore prêt, les styles préparés par les membres du TWGMLC sont basés sur une version mise à jour du système Babel de Johannes Braams. Il est prévu intégrer cette dernière dans le noyau de LATEX3 ;

8. Ceci dit, on n'utilise pas les mêmes polices pour composer le yiddish et l'hébreu moderne.

– fichier de configuration MakeIndex3 ; cette extension de MakeIndex, écrite par Joachim SCHROD et présentée lors du congrès TEX à Paris en 1991[9], permettra la création d'index pour toute langue ;

– le cas échéant, des polices (réelles ou virtuelles) ;

– éventuellement un *filtre*, c'est-à-dire une application qui pourrait ultérieurement faire partie de TEX et qui permettrait de convertir des caractères 8-bits d'après un fichier de configuration normalisé ;

– de la documentation : une notice expliquant le contenu du TLP en anglais (américain) et dans la langue-cible, et une introduction à (La)TEX dans la langue-cible.

L'avancement des travaux du TWGMLC ainsi que les sorties β seront annoncés au fur et à mesure par la presse TEXnique. Une mise au point (ainsi qu'une discussion ouverte à ce sujet) sera faite lors de l'Assemblée Générale de TUG à Aston (Angleterre), en juillet 1993.

9. « An International Version of MakeIndex », *Cahiers GUTenberg*, n° 9, septembre 1991, 81–90. Cette version est en β-test.