

Formaliser l'alphabet

max.silberztein@univ-fcomte.fr

Rappel : le codage

- On représente les diverses valeurs que peut avoir une information par des nombres, ex. « feu vert » = 1, « feu rouge » = 18, etc.
- Chaque valeur doit correspondre à un et un seul nombre
- Chaque nombre doit correspondre à une et une seule valeur
- On représente les nombres en binaire

=> on a numérisé l'information

un codage pour l'alphabet latin

Code	bits	lettre	Co de	bits	lettr e	cod e	bits	lettr e	cod e	bits	lettre
0	00000	inutilisé	8	01000	H	16	10000	P	24	11000	X
1	00001	A	9	01001	I	17	10001	Q	25	11001	Y
2	00010	B	10	01010	J	18	10010	R	26	11010	Z
3	00011	C	11	01011	K	19	10011	S	27	11011	inutilisé
4	00100	D	12	01100	L	20	10100	T	28	11100	inutilisé
5	00101	E	13	01101	M	21	10101	U	29	11101	inutilisé
6	00110	F	14	01110	N	22	10110	V	30	11110	inutilisé
7	00111	G	15	01111	P	23	10111	W	31	11111	inutilisé

Que manque-t-il ?

- Compléter l'alphabet, i.e. l'ensemble des caractères nécessaires pour écrire un texte français

Que manque-t-il ?

- Lettres minuscules
- Lettres accentuées
- Ponctuations
- Caractères de contrôle

Les lettres accentuées et diacritiques

- Les 13 lettres françaises accentuées

à â ç é è ê ë î ï ô ù û ü

Les lettres accentuées et diacritiques

- Les 13 lettres françaises accentuées

à â ç é è ê ë î ï ô ù û ü

- « ü » (L'Haÿ-les-Roses)
- « œ » dans *l'œuf*
- « æ » dans *Vitæ*



Appartements en vente à



Devenir propriétaire à

En vente à L'Haÿ-les-Roses (94)

Trouvez votre bonheur à L'Haÿ-les-Roses parmi ces beaux appartements en Offre Spéciale

Les lettres accentuées et diacritiques

- Les 13 lettres françaises accentuées

à â ç é è ê ë î ï ô ù û ü

- « ÿ » (L'Haÿ-les-Roses),
- « œ » dans *l'œuf*
- « æ » dans *Vitæ*
- Autres ligatures : fi (*fichier*), fl (*fleur*), ffi (*difficile*), ffl (*effleurer*)

Les lettres accentuées et diacritiques

- Les 13 lettres françaises accentuées

à â ç é è ê ë î ï ô ù û ü

- « ÿ » (L'Haÿ-les-Roses),
- « œ » dans *l'œuf*
- « æ » dans *Vitæ*
- Autres ligatures : fi (*fichier*), fl (*fleur*), ffi (*difficile*), ffl (*effleurer*)
- Les lettres étrangères souvent utilisées : ã, å, ñ,

Les ligatures françaises

- La ligature **æ** est utilisée dans une vingtaine de mots d'origine latine :
ad vitam æternam, ægagropile, ægosome, æpyornis, æschne, æthuse, althæa, cæcal, cæcum, cæsium, chamærops, curriculum vitæ, ex æquo, intuitu personæ, lapsus linguæ, nævus, præsidium, tædium vitæ, tænia, uræus.
- La ligature **œ** est utilisée dans une vingtaine de mots d'origine grecque ou latine :
bœuf, chœur, cœlacanthe, cœlentéré, cœur, écoeurement, écoeurer, fœtus, mœurs, nœud, œcuménique, œdème, œdipien, œil, œillère, œillet, œnologie, œnologue, œsophage, œstrogène, œuf, œuvre, rancœur, sœur, vœu

Ligatures dans les autres langues

- D'autres ligatures ou digraphes existent aussi pour d'autres langues : le « ll » espagnol, le *esszett* allemand « ß », le « ij » hollandais, le caractère « 8 » pour représenter la suite « ou » en grec, le double vav « II » en hébreu, etc.
- Certains systèmes d'écriture utilisent les ligatures de façon productive : par exemple en devanāgarī (utilisé pour certaines langues indiennes dont le hindi), deux consonnes qui se suivent sont systématiquement ligaturées.

Les codages sont arbitraires

... donc fondamentalement incompatibles :

- si je décide de coder la lettre « A » ainsi : 0011
- mais que mon interlocuteur décide de coder la lettre « A » ainsi : 1010
- et que le code 0011 correspond chez lui à la lettre « W »
=> alors à chaque fois que je lui enverrai un « A », il recevra un « W »

Standardisation des codages

- avec l'arrivée du télégraphe, le codage Baudot (1874) utilisait 5 bits, donc 32 caractères différents
- qui a évolué en codage Murray, puis ITA2 (International Telegraph Alphabet n°2) utilisé par les télex et radio-télétypes
- qui a évolué en codage ASCII (American Standard Code for Information Interchange), qui utilise 7 bits



		1	2	3	4	5			1	2	3	4	5
A		●	●				Q	1	●	●	●		●
B	?	●			●	●	R	4		●		●	
C	(●	●	●		S	'	●		●		
D	²	●			●		T	5					●
E	3	●					U	7	●	●	●		
F	/	●		●	●		V)		●	●	●	●
G	³ /		●		●	●	W	2	●	●			●
H	⁵ /			●		●	X	£	●		●	●	●
I	8		●	●			Y	6	●		●		●
J	⁷ /	●	●		●		Z	.	●				●
K	⁹ /	●	●	●	●		FIG	FIG	●	●		●	●
L	/		●			●	SPACE	SPACE			●		
M	'			●	●	●	LTR	LTR					
N	-			●	●		SPACE	SPACE					
O	9				●	●	LINE	LINE				●	
P	O		●	●		●	PAGE	PAGE					
							✱	✱	●	●	●	●	●
							COL	COL		●			

Codage Baudot



Code	Caractère	Code	Caractère	Code	Caractère	Code	Caractère
0	NUL	32	SP	64	@	96	`
1	SOH	33	!	65	A	97	a
2	STX	34	"	66	B	98	b
3	ETX	35	#	67	C	99	c
4	EOT	36	\$	68	D	100	d
5	ENQ	37	%	69	E	101	e
6	ACK	38	&	70	F	102	f
7	BEL	39	'	71	G	103	g
8	BS	40	(72	H	104	h
9	HT	41)	73	I	105	i
10	LF	42	*	74	J	106	j
11	VT	43	+	75	K	107	k
12	FF	44	,	76	L	108	l
13	CR	45	-	77	M	109	m
14	SO	46	.	78	N	110	n
15	SI	47	/	79	O	111	o

Code	Caractère	Code	Caractère	Code	Caractère	Code	Caractère
16	DLE	48	0	80	P	112	p
17	DC1	49	1	81	Q	113	q
18	DC2	50	2	82	R	114	r
19	DC3	51	3	83	S	115	s
20	DC4	52	4	84	T	116	t
21	NAK	53	5	85	U	117	u
22	SYN	54	6	86	V	118	v
23	ETB	55	7	87	W	119	w
24	CAN	56	8	88	X	120	x
25	EM	57	9	89	Y	121	y
26	SUB	58	:	90	Z	122	z
27	ESC	59	;	91	[123	{
28	FS	60	<	92	\	124	
29	GS	61	=	93]	125	}
30	RS	62	>	94	^	126	~
31	US	63	?	95	_	127	DEL

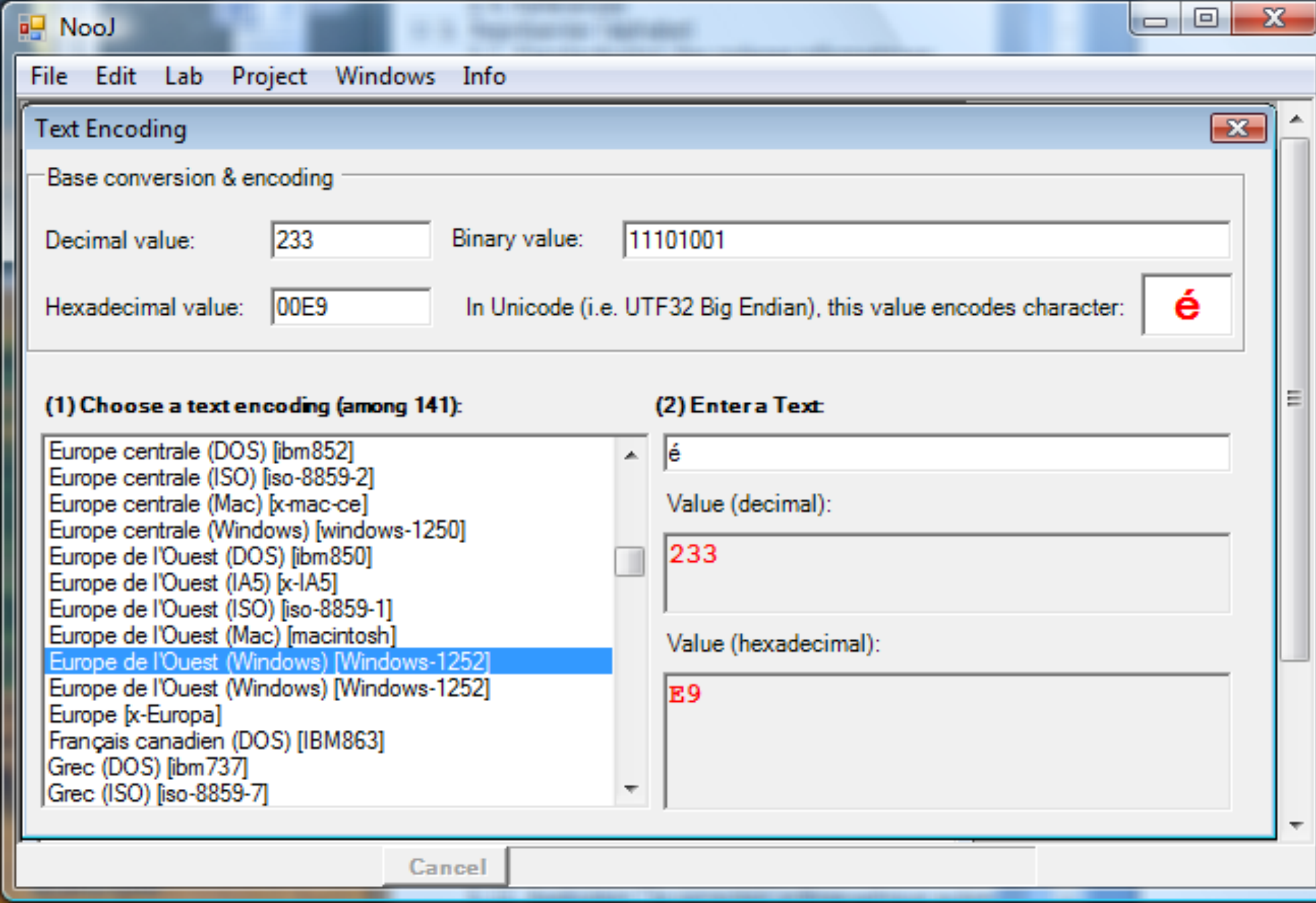
ASCII

Codages ASCII étendus

- 7 bits, i.e. 128 caractères, ne suffisent pas pour les langues autres que l'anglais
- Donc on a étendu le codage ASCII à 8 bits pour y ajouter les lettres françaises, ou alors l'alphabet arabe, ou alors l'alphabet grec, ou alors l'alphabet russe, etc.
- Les constructeurs ne se sont pas mis d'accord, donc on a un codage ASCII grec Mac, un codage ASCII russe IBM, un codage ASCII arabe Windows, etc.

Variantes orthographiques

- En chinois traditionnel, plus de 1.000 caractères possèdent une ou plusieurs variantes
- De plus, le chinois utilise des caractères contractés, utilisés comme abréviations
- Les textes grecs contiennent de nombreuses confusions entre caractères grecs et latins, ex. NATO écrit avec un alpha et un omicron
- Les textes égyptiens contiennent de nombreuses confusions entre les caractères arabes hamza et alifs



141 codages
ASCII
étendus

Situation chaotique

sont à 0,10E/SMS.
Les SMS reçus sont
gratuits. Pour plus

Agence Imagine R

26/09/14



image R : recharger votre passe d~~?~~s maintenant

Bonjour, Je tiens à vous informer que le forfait image R de

#region Code g~~?~~n~~?~~r~~?~~ par le Concepteur Windows Form

Incompatibilité

- Pour envoyer un fichier texte créé sur un Mac à un collègue qui a un IBM PC, il fallait le convertir
- Impossible d'envoyer un fichier contenant du texte français à un collègue grec ou russe...

Le codage Unicode

- Le consortium Unicode créé en 1991 pour établir un codage unique, répertorie tous les caractères de toutes les langues, y compris les langues mortes (ancien égyptien, maya)
- Le site www.unicode.org/charts donne accès aux tables de tous les systèmes d'écriture
- Le codage Unicode ne fournit qu'une correspondance entre caractères et nombres entiers ; il existe une dizaine de façons pour représenter ces nombres entiers dans les ordinateurs (UTF7, UTF8, UTF16L, UTF32H, etc.)

Implémentations d'Unicode

- L'implémentation la plus naturelle est UTF32H : on représente chaque entier en l'écrivant en binaire sous la forme d'un nombre de 32 bits (du haut vers le bas).
- Avec 32 bits, on a la possibilité de représenter plus de 4 milliards de codes !
- Actuellement, Unicode répertorie environ 200.000 caractères ; le gouvernement chinois collabore avec le consortium Unicode pour répertorier tous les caractères chinois traditionnels.

Exercice

- La lettre « A » a le code Unicode : 65
- Comment cette lettre est-elle représentée dans l'ordinateur en UTF32H ? En UTF32L ?
- La lettre « ü » a le code Unicode : 252
- Comment cette lettre est-elle représentée dans l'ordinateur en UTF32H ?

Exercice

- La lettre « A » a le code Unicode : 65
- Comment cette lettre est-elle représentée dans l'ordinateur en UTF32H ? En UTF32L ?

0000 0000 0000 0000 0000 0000 0100 0001

1000 0010 0000 0000 0000 0000 0000 0000

- La lettre « ü » a le code Unicode : 252
- Comment cette lettre est-elle représentée dans l'ordinateur en UTF32H ? En UTF32L ?

Exercice

- La lettre « A » a le code Unicode : 65
- Comment cette lettre est-elle représentée dans l'ordinateur en UTF32H ? En UTF32L ?

0000 0000 0000 0000 0000 0000 0100 0001

1000 0010 0000 0000 0000 0000 0000 0000

- La lettre « ü » a le code Unicode : 252
- Comment cette lettre est-elle représentée dans l'ordinateur en UTF32H ?

0000 0000 0000 0000 0000 0000 1111 1100

UTF32H, problème

- UTF32H a l'avantage d'être très simple à comprendre et à implémenter : un caractère est codé sous la forme d'une suite de 4 octets ; pour accéder au 17^e caractère d'un texte, il suffit de rechercher dans le fichier les 4 octets à partir de $16 \times 4 = 64^{\text{ème}}$ caractère.
- Mais du coup, l'ensemble des textes qui étaient représentés en ASCII (la très grande majorité) correspondent maintenant à des fichiers 4 fois plus gros ; les ordinateurs doivent travailler 4 fois plus pour les lire, ils sont donc 4 fois moins rapides...

UTF8

- Aujourd'hui, la quasi-totalité des systèmes utilisent UTF8 pour coder les textes : HTML5, Mac OSX, Windows, LINUX, etc.
- Pour représenter les caractères qui ont un code ASCII, on utilise un seul octet (identique)
- Pour représenter les caractères qui n'avaient pas de code ASCII, on utilise soit 2 octets (ex. Les alphabets arabe, arménien, cyrillique, hébreu, etc.), soit 3 octets (ex. Les alphabets chinois simplifié, japonais et coréen), soit même 4 octets (caractères chinois traditionnels, phénicien, etc.).
- AVANTAGE : très efficace pour la grande majorité des textes sur Internet

Unicode : problèmes

- Codage des caractères composés
- Codage des caractères chinois
- Incomplétude
- Unification

Codage des caractères composés

- Plusieurs façons de représenter les caractères composés :
 - Soit avec un code Unicode, ex. La lettre "ü" est codée par 00FC
 - Soit avec une séquence de codes Unicode, ex. La lettre "ü" est codée par la séquence de deux codes : 0075 (pour la lettre "u") suivi de 0308 (pour le tréma).
 - Les alphabets qui contiennent des lettres qui peuvent avoir plusieurs signes diacritiques posent de nombreux problèmes

Lettres composées dans Unicode

- Pour coder la lettre hébraïque "שׁ" (lettre shin/sin avec un point "sh" et un daguash de gémination), on a alors deux possibilités :
 - code du shin, du daguash, du point : 05E9 05BC 05C1 ;
 - code du shin, du point, du daguash : 05E9 05C1 05BC.
- Pour coder la lettre vietnamienne "ế" ("e" avec un accent aigu et un accent circonflexe), on a trois possibilités :
 - code du caractère composé : 1EBF ;
 - code de « e », de l'accent circonflexe, de l'accent aigu : 0045 0302 0301 ;
 - code de « e », de l'accent aigu, de l'accent circonflexe : 0045 0301 0302.

Incomplétude d'Unicode

- De nombreux caractères chinois ne sont pas répertoriés dans Unicode

馱

- Certains caractères ont plusieurs codes Unicode :

Traditional Chinese	Simplified Chinese	Japanese
氣(6C23)	气(6C14)	気(6C17)

- De nombreux caractères chinois sont utilisés en coréen, en japonais ou en vietnamien... mais avec des glyphes différents

Simplified Chinese	Traditional Chinese	Japanese	Korean
漢	漢	漢	漢

Analyse automatique de textes

- Lorsqu'on décrira une langue, on utilisera une orthographe et un codage spécifique
- Mais il y a peu de chances que les textes qu'on voudra analyser utilisent le même codage, de la même façon et de façon cohérente
- La lecture d'un texte à partir d'un fichier est alors compliqué, puisqu'il faut construire un logiciel simplement pour pouvoir reconnaître les lettres accentuées et composées ainsi que les ligatures

Ordre lexicographique

- Dans les dictionnaires français :
 - On supprime les accents
 - On compare les mots de gauche à droite
 - Si il y a une différence, on compare les lettres différentes
 - Sinon, on rétablit les accents et on compare les mots de droite à gauche :
sans accent < accent aigu < accent grave < accent circonflexe < tréma
 - Par exemple, relève < relevé car e < é
 - La casse (majuscule < minuscule),
 - Les ligatures (résolues)
 - Les abréviations (ste < santé)
 - Les nombres ordinaux (Louis XIII < Louis XIV)
- D'autres ordres sont utilisés pour chaque application

Chaque langue a son propre ordre lexicographique

- en allemand, la lettre « ß » et les lettres avec umlaut (« ä », « ö », « ü ») sont remplacées par les digrammes : « ss », « ae », « oe » et « ue »
- en suédois, le caractère « ä » est traité comme une lettre à part entière, rangée après le « z »
- en espagnol, le caractère « ñ » est considéré comme une lettre à part entière, rangée entre le « n » et le « o » ; le mot *señor* est donc rangé après *sensa*. Dans les dictionnaires traditionnels (publiés avant la réforme de 1994), les digrammes « ch » et « ll » considérés comme des lettres indépendantes (le mot *llanos* était donc rangé après le mot *los*)

Chaque langue a son propre ordre lexicographique

- en danois, les lettres « æ », « ø » et « å » sont des lettres à part entière, rangées après le « z » dans l'alphabet
- en néerlandais, le digramme « ij » est considéré comme une lettre à part entière, rangée entre le « y » et le « z ».
- Les alphabets non latins ont des règles particulières ; ex. en russe, on regroupe « е » et « ё » tandis que les lettres « ъ » et « ь » sont ignorées lors de la comparaison lexicographique.
- L'ordre thaï implique un calcul en 7 étapes : on compare la première consonne des deux mots, puis les voyelles implicites, puis les doubles consonnes, puis les autres voyelles, etc.

Ordre lexicographique chinois

- Impossible d'apprendre par coeur un ordre alphabétique puisqu'il y a plus de 3.000 caractères en chinois simplifié, voir 100.000 caractères en chinois traditionnel
- Deux méthodes pour trier en chinois :
 - Méthode graphique, basé sur l'écriture du caractère
 - Méthode phonétique, basée sur l'alphabet latin

Méthode graphique

- On trie le caractère en fonction de son radical. Il y a 214 radicaux en chinois, donc un dictionnaire chinois contient 214 sections.
- Pour trier deux radicaux, on compte le nombre de coups de pinceaux (traits) utilisés pour dessiner le radical, ex. 乚 (2 traits) < 面 (9 traits)
- Chacune des 214 sections contient tous les mots qui commencent par un caractère qui contient le radical en question
- Pour trier les mots à l'intérieur d'une section, on compte le nombre de traits à ajouter au radical pour compléter le caractère, ex. : dans la section 乚, on a les deux mots 北 (*nord*, 3 traits) et 阜 (*surpasser*, 6 traits).

Méthode phonétique

- On utilise la prononciation en mandarin, que l'on transcrit en caractères latins selon la transcription pinyin.
- On range ensuite les mots en utilisant l'ordre alphabétique latin.
- L'inconvénient, c'est qu'il faut savoir comment les mots se prononcent en mandarin pour savoir les ordonner

Conclusion

- Il est possible de représenter l'alphabet de n'importe quelle langue dans un ordinateur ; on utilise pour cela un codage
- Les codages étaient jusqu'à présent incompatibles entre eux ; le codage Unicode a pour vocation de les remplacer tous.
- Le codage Unicode n'est pas parfait ; en particulier il nécessite des calculs d'équivalence, n'est pas pratique pour les langues asiatiques, et n'est pas complet pour le chinois traditionnel.