De l'écrit à l'Information: langue et langage

Karën Fort

karen.fort@univ-lorraine.fr / https://members.loria.fr/KFort

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

Quelques sources d'inspiration

- B. Habert (anciennement ENS Lyon) : langues et langages

Retour sur le TD précédent

Caractéristiques des langues

Ambiguités à tous les étages (1/2) !

- Délimiter les unités : une même suite de sons ou de mots peut être découpée de plusieurs manières :

Caractéristiques des langues

Ambiguités à tous les étages (1/2) !

- Délimiter les unités : une même suite de sons ou de mots peut être découpée de plusieurs manières :

  mon beau-frère et ma soeur ; mon beau-frère est masseur ; mon beau-frère hait ma sœur

Caractéristiques des langues

Ambiguités à tous les étages (1/2) !

- Délimiter les unités : une même suite de sons ou de mots peut être découpée de plusieurs manières :

  mon beau-frère et ma soeur ; mon beau-frère est masseur ; mon beau-frère hait ma sœur

  (il a cassé) (sa pipe) ; (il) (a cassé sa pipe)

Caractéristiques des langues

Ambiguités à tous les étages (1/2) !

- Délimiter les unités : une même suite de sons ou de mots peut être découpée de plusieurs manières :

  mon beau-frère et ma soeur ; mon beau-frère est masseur ; mon beau-frère hait ma sœur

  (il a cassé) (sa pipe) ; (il) (a cassé sa pipe)

- Délimiter les groupes :

Caractéristiques des langues

Ambiguités à tous les étages (1/2) !

- Délimiter les unités : une même suite de sons ou de mots peut être découpée de plusieurs manières :

  mon beau-frère et ma soeur ; mon beau-frère est masseur ; mon beau-frère hait ma sœur

  (il a cassé) (sa pipe) ; (il) (a cassé sa pipe)

- Délimiter les groupes :

  (Nadine couvre) (la corbeille de fleurs) ;
  (Nadine couvre la corbeille) (de fleurs)

Caractéristiques des langues

Ambiguités à tous les étages (1/2) !

- Délimiter les unités : une même suite de sons ou de mots peut être découpée de plusieurs manières :

  mon beau-frère et ma soeur ; mon beau-frère est masseur ; mon beau-frère hait ma sœur

  (il a cassé) (sa pipe) ; (il) (a cassé sa pipe)

- Délimiter les groupes :

  (Nadine couvre) (la corbeille de fleurs) ;
  (Nadine couvre la corbeille) (de fleurs)

- Déterminer la fonction des groupes :

Caractéristiques des langues

Ambiguités à tous les étages (1/2) !

- Délimiter les unités : une même suite de sons ou de mots peut être découpée de plusieurs manières :
  mon beau-frère et ma soeur ; mon beau-frère est masseur ; mon beau-frère hait ma sœur
  (il a cassé) (sa pipe) ; (il) (a cassé sa pipe)
- Délimiter les groupes :
  (Nadine couvre) (la corbeille de fleurs) ;
  (Nadine couvre la corbeille) (de fleurs)
- Déterminer la fonction des groupes :
  Il attend la nuit
  [que la nuit vienne | pendant la nuit]

Caractéristiques des langues

Ambiguités à tous les étages (2/2) !

- Déterminer les relations grammaticales entre groupes :

Caractéristiques des langues

Ambiguités à tous les étages (2/2) !

- Déterminer les relations grammaticales entre groupes :
  Georges admire Marie autant que Jean [Georges admire Jean | Jean admire Marie]

Caractéristiques des langues

Ambiguités à tous les étages (2/2) !

- Déterminer les relations grammaticales entre groupes :
  Georges admire Marie autant que Jean [Georges admire Jean | Jean admire Marie]
- Déterminer les relations sémantiques entre groupes :

Caractéristiques des langues

Ambiguités à tous les étages (2/2) !

- Déterminer les relations grammaticales entre groupes :
  Georges admire Marie autant que Jean [Georges admire Jean | Jean admire Marie]
- Déterminer les relations sémantiques entre groupes :
  Toutes les victimes n'avaient pas été vaccinées [il existe des victimes non vaccinées | l'ensemble des victimes n'a pas été vacciné] ;

Caractéristiques des langues

Ambiguités à tous les étages (2/2) !

- Déterminer les relations grammaticales entre groupes :

  Georges admire Marie autant que Jean [Georges admire Jean | Jean admire Marie]

- Déterminer les relations sémantiques entre groupes :

  Toutes les victimes n'avaient pas été vaccinées [il existe des victimes non vaccinées | l'ensemble des victimes n'a pas été vacciné] ;

  Chacun cherche son chat [tous le même | chacun le sien]

Caractéristiques des langues

Ambiguités à tous les étages (2/2) !

- Déterminer les relations grammaticales entre groupes :

  Georges admire Marie autant que Jean [Georges admire Jean | Jean admire Marie]

- Déterminer les relations sémantiques entre groupes :

  Toutes les victimes n'avaient pas été vaccinées [il existe des victimes non vaccinées | l'ensemble des victimes n'a pas été vacciné] ;

  Chacun cherche son chat [tous le même | chacun le sien]

- Déterminer le but pratique visé :

Caractéristiques des langues

Ambiguités à tous les étages (2/2) !

- Déterminer les relations grammaticales entre groupes :

  Georges admire Marie autant que Jean [Georges admire Jean | Jean admire Marie]

- Déterminer les relations sémantiques entre groupes :

  Toutes les victimes n'avaient pas été vaccinées [il existe des victimes non vaccinées | l'ensemble des victimes n'a pas été vacciné] ;

  Chacun cherche son chat [tous le même | chacun le sien]

- Déterminer le but pratique visé :

  Cela ne se dit pas [constat | obligation]

Ambiguïté volontaire

L'ambiguïté n'est pas forcément un défaut des langues, mais aussi la possibilité de dire « plusieurs choses » à la fois :

Dans le milieu (c'est le cas de le dire) du football...
[entourage | réseau basé sur la corruption]

On joue aussi avec la langue, en poésie, dans les jeux de mots :

Un membre de l'Assemblée nationale perd son calme :
La moitié de ce parlement est composée d'imbéciles !
On exige des excuses.
Les voici : Je retire ce que j'ai dit. La moitié de ce Parlement n'est pas composée d'imbéciles !

La langue vit

Le sens d'un mot, d'une construction n'est pas univoque :
Exemple : le rôle de l'impératif dans

Empruntez le passage souterrain / Buvez Coca-Cola

Une langue « bouge » : arrivée de nouveaux mots, de nouvelles constructions. Les règles d'une langue évoluent nécessairement, pas celles d'un langage artificiel

- *Monsieur, n'est--ce pas vous qui vous appelez Sganarelle ?*
- *Eh ! qui ?*
- *Je vous demande si ce n'est pas vous qui se nomme Sganarelle [Molière, 17ème siècle]*

La place des détachements multiples en français oral :

Moi, ma mère, le salon, c'est de la moquette [José Deulofeu]

Des règles

On ne joue pas avec la langue comme on joue aux échecs

- Personne ne connaît toutes les règles d'une langue.
- Les usuels (dictionnaires et grammaires) ne notent pas l'intégralité des règles d'une langue.
- Les dictionnaires et les grammaires notent des mots/règles que ne connaissent pas tous les locuteurs, voire qui ne sont plus usités.

Des règles

Les règles d'une langue comportent flou et exceptions :

> *Je m'en vais ou je m'en vas, puisque l'un ou l'autre se dit ou se disent [le grammairien Vaugelas, au 17ème siècle, en mourant]*

On ne peut pas toujours dire si un énoncé appartient ou non à une langue :

> *Le son de sa voix est une cicatrice [A. Breton & P. Eluard]*

. . .alors qu'on peut toujours dire si une suite de signes appartient à un langage artificiel donné, par exemple si un programme est une « phrase » valide d'un langage de programmation.

et al.

- La compréhension peut dépendre de la connaissance de qui a communiqué à qui, où et quand (contexte) :
  je viens ici demain
- Le vocabulaire et les règles d'une langue sont très nombreux.
- On peut écorcher une langue et arriver à se faire comprendre :
  si vuos pvueoz lrie ccei, vuos aevz nu dôrle de cvreeau

Wooclap


Comment participer ?
1
Allez sur wooclap.com
2
Entrez le code d'événement dans le bandeau supérieur
Code d'événement
OBGCYT
Activer les réponses par SMS

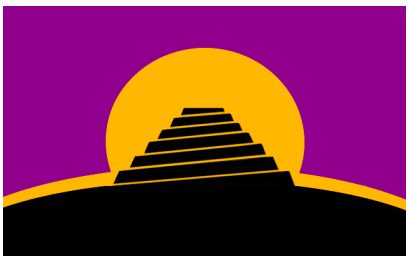
Venu, ni iru (wooclap)

Exercice

Citez des noms de langues artificielles

Langues artificielles

ou construites



Christian Thalmann, Jan van Steenbergen, Leland Paul, and David Peterson

- "langue[s] créée[s] par une ou plusieurs personnes dans un temps relativement bref" (Wikipédia)
- en général, dans un but précis
- incluent les langues de fiction (but artistique)
- très peu avec un nombre significatif de locuteurs :
  - espéranto (1887) : 100 000 à 3 millions
  - interlingua (1951) : 2 000
  - volapük (1880) : moins de 50

Langages

http://www.cnrtl.fr/lexicographie/langage

B. − Système de signes vocaux et/ou graphiques (*cf. langue* II A).

1. [*Langages naturels :* les langues parlées dans le monde] *Langage écrit, parlé. Les mots qui composent cette classe* [*la classe des mots invariables*], *ont tous les mêmes raisons d'en être (...); c'est pourquoi ils sont les mêmes dans tous les langages* (Destutt de Tr., *Idéol. 2,*1803, p. 128).*Les amnésiques du verbe oublient d'abord ce qu'il y a de plus particulier dans le langage, les noms propres, les substantifs, les adjectifs; les parties du langage qui ont la vie la plus dure sont les phrases toutes faites, les locutions usuelles* (Gourmont, *Esthét. lang. fr.,*1899, p. 285).*J'ai reçu des lèvres de ma vieille servante le bon langage français* (France, *Pt Pierre,*1918, p. 196).*La grande influence qu'il semble que Descartes ait exercée sur nos Lettres; l'événement dont il est l'auteur, de la première production en langage français d'un ouvrage de philosophie* (Valéry, *Variété IV,*1938, p. 209).

2. [*Langages artificiels,* établis en fonction d'axiomes, de règles d'écriture] Système de symboles. *Langage documentaire* (v. ce mot B 2); *langage formel, logique. Il* [*Leibniz*] *concevait la notion de langage formalisé, pure combinaison de signes dont seul importe l'enchaînement, de sorte qu'une machine serait capable de fournir tous les théorèmes, et que toutes les controverses se résoudraient par un simple calcul* (Bourbaki, *Hist. math.*,1960, p. 16).*Construire des langages artificiels à composantes logiques pour servir aux tâches essentielles de la documentation : analyse, enregistrement des documents, et recherche documentaire* (Coyaud, *Introd. ét. lang. docum.*,1966, p. 67).

− *INFORMAT., PROGRAMMATION.* Ensemble de symboles et de règles permettant de combiner ces symboles afin de donner des instructions à un ordinateur.

Langages artificiels

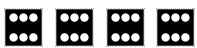
Exemples de :

- langages artificiels : l'algèbre, la logique
- langages de programmation : Java, Python

Un langage artificiel : le Yam's

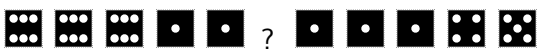
Les dés ont 6 faces, on jette 5 dés.

Certaines configurations ont une valeur (par ordre croissant) :

- brelan : 
- carré : 
- full :  (brelan) + (paire)
- petite suite (4 dés qui se suivent) : 
- grande suite (5 dés qui se suivent) : 
- Yams : 

Le Yam's : quizz

Quelle relation mathématique existe entre ces deux expressions ?



Un langage artificiel : le Shadok

(c) DVD Les shadoks, édition intégrale

Capacité de mémoire
Compter

Caractéristiques des langages artificiels

Pour toute suite de signes relevant d'un langage artificiel, on peut dire si elle appartient ou non au langage en question :

$$(a+b)^2 \qquad vs \qquad (a+b)(a-$$

Les langages artificiels sont :

- stabilisés (notion de *standard*) : Java 1.2 *vs* Java 1.7.
- explicitement conçus pour ne pas pouvoir être ambigus. Il a d'ailleurs fallu un certain temps avant qu'on comprenne comment empêcher un langage artificiel d'être ambigu (pour un jeu de paramètres donnés, un programme doit avoir un résultat (ou un comportement) unique).

Langages « finis » et « langages » infinis

Les langues autorisent la création d'énoncés à l'infini :

> *Maudit soit le père de l'épouse du forgeron qui forgea le fer de la cognée avec laquelle le bûcheron abattit le chêne dans lequel on sculpta le lit où fut engendré l'arrière-grand-père de l'homme qui conduisit la voiture dans laquelle ta mère rencontra ton père !*
>
> *[Robert Desnos, La colombe de l'arche, 1923]*

On ne peut pas faire l'inventaire des énoncés d'une langue.

Langages « finis » et « langages » infinis (wooclap)

Yam's

Combien de combinaisons de 5 dés sont possibles (au Yam's) ?

Langages « finis » et « langages » infinis

Le nombre de suites possibles dans certains langages artificiels est limité (même s'il peut être énorme, comme au jeu d'échec) : Yam's

D'autres langages artificiels permettent d'engendrer des suites à l'infini :

?

Langages « finis » et « langages » infinis

Le nombre de suites possibles dans certains langages artificiels est limité (même s'il peut être énorme, comme au jeu d'échec) : Yam's

D'autres langages artificiels permettent d'engendrer des suites à l'infini :

les langages de programmation

La langue de l'ordinateur

- Le texte n'existe pas en informatique

La langue de l'ordinateur

- Le texte n'existe pas en informatique
- Quelle langue ≪ parle ≫ l'ordinateur ? Quels sont ses mots ?

La langue de l'ordinateur

- Le texte n'existe pas en informatique
- Quelle langue « parle » l'ordinateur ? Quels sont ses mots ?
- Matériellement, un ordinateur ne comprend que le langage binaire, c'est-à-dire une suite de 0 et de 1
  Pour faire simple : du courant, pas de courant

La langue de l'ordinateur

- Le texte n'existe pas en informatique
- Quelle langue « parle » l'ordinateur ? Quels sont ses mots ?
- Matériellement, un ordinateur ne comprend que le langage binaire, c'est-à-dire une suite de 0 et de 1
  Pour faire simple : du courant, pas de courant
- On appelle **bit** (BInary digiT) cette unité élémentaire d'information, qui peut prendre comme valeur 0 ou 1

Allez, hop tous en combinaison (wooclap)

Exercice

Combien existe-t'il de possibilités de combiner 2 bits ?

Petite parenthèse : ça ne vous rappelle rien ?

(c) DVD Les shadoks, édition intégrale

Compter

Un peu plus loin dans les bases

- En fait, un processeur manipule plutôt des paquets de bits de taille fixe
- Les premiers processeurs fixèrent la taille de ces paquets à huit bits, soit un **octet**
- De huit bits (processeur 8086), les processeurs sont passés à 32 bits (Pentium 4), pour arriver aujourd'hui à 64. Ces paquets correspondent en fait à la capacité de transport (dans ses bus) et de traitement de la machine (taille des registres)

À noter

1 octet = 8 bits = 1 byte (en anglais)

Dites « A » !

- « A » est en fait une entité abstraite dont le nom est « a majuscule » et dont le **glyphe** ressemble à un triangle dont on aurait raccourci et remonté le côté bas [dessin]

Dites « A » !

- « A » est en fait une entité abstraite dont le nom est « a majuscule » et dont le **glyphe** ressemble à un triangle dont on aurait raccourci et remonté le côté bas [dessin]
- Comment dit-on « A » à un ordinateur ?

Dites « A » !

- « A » est en fait une entité abstraite dont le nom est « a majuscule » et dont le **glyphe** ressemble à un triangle dont on aurait raccourci et remonté le côté bas [dessin]

- Comment dit-on « A » à un ordinateur ?
- 01000001

Dites « A » !

- « A » est en fait une entité abstraite dont le nom est « a majuscule » et dont le **glyphe** ressemble à un triangle dont on aurait raccourci et remonté le côté bas [dessin]

- Comment dit-on « A » à un ordinateur ?
- 01000001

À noter

1 caractère = 1 octet (attention, simplification !)

Le byte et ses bits (wooclap)

Exercice

Combien de valeurs peut-on coder sur un octet ?

Quelques exemples

Caractère	Code Binaire	Description
A	01000001	Caractère « A »
a	01100001	Caractère « a »
T	01010100	Caractère « T »
t	01110100	Caractère « t »
3	00110011	Caractère « 3 »
$	00100100	Caractère « $ »
BEL	00000111	Bip

À noter

- pour passer de la majuscule à la minuscule, il suffit de mettre le troisième bit à 1
- les caractères chiffres ont leur propre code
- il n'existe qu'un « A », qui s'affiche différemment selon les polices ou le style (voir Définitions)

Dites « Bonjour » à la dame (wooclap)

Exercice

Combien de bits dans « Bonjour » ?

Donc, l'ordinateur cause octet, so what ?

Il faut un traducteur pour que notre texte soit « codé » et « décodé » proprement, de manière **standardisée**

C'est là qu'interviennent les tables de conversion, ou les encodages

Récapitulons

- un ordinateur ne comprend que le langage binaire, c'est-à-dire une suite de 0 et de 1
- l'objet qui prend comme valeur 0 ou 1 est appelé bit
- en simplifiant : 1 caractère = 1 octet = 8 bits

Da ASCII code

- Au début de l'informatique, on estimait que coder les caractères sur 7 bits, ça suffisait bien, puisque ça permet de représenter 2^7=128 caractères différents. « A » est donc codé 1000001
- C'est ainsi que fut créée la table ASCII (American Standard Code for Information Interchange), publiée en 1968
- Très longtemps ce fut LA table de référence en informatique, à tel point qu'elle devint une norme : ISO-646

Table ASCII : ça suffit, non ?

	0	1	2	3	4	5	6	7
0	NUL	DLE	space	0	@	P	`	p
1	SOH	DC1 XON	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3 XOFF	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	del

Non, ça ne suffit pas

Big Blue invente l'ASCII étendu

- 1981, IBM sort son premier PC et ajoute un bit à l'ASCII
- L'ASCII étendu OEM ((Original Equipment Manufacturer) permet donc de coder 256 caractères (2^8) et certains sont contents de pouvoir fêter Noël
- Ce code ASCII étendu n'est pas unique et dépend fortement de la plate-forme utilisée

Table ASCII étendu OEM

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	Ç	ü	é	â	ä	à	å	ç	ê	ë	è	ï	î	ì	Ä	Å
9	É	æ	Æ	ô	ö	ò	û	ù	ÿ	Ö	Ü	¢	£	¥	₧	ƒ
A	á	í	ó	ú	ñ	Ñ	ª	º	¿	⌐	¬	½	¼	¡	«	»
B	░	▒	▓	│	┤	╡	╢	╖	╕	╣	║	╗	╝	╜	╛	┐
C	└	┴	┬	├	─	┼	╞	╟	╚	╔	╩	╦	╠	═	╬	╧
D	╨	╤	╥	╙	╘	╒	╓	╫	╪	┘	┌	█	▄	▌	▐	▀
E	α	ß	Γ	π	Σ	σ	µ	τ	Φ	Θ	Ω	δ	∞	ø	∈	∩
F	≡	±	≥	≤	⌠	⌡	÷	≈	°	∙	·	√	ⁿ	²	■	

On imagine la tête de ceux qui ont besoin d'écrire водка. . .

La famille ISO

. . .d'où l'idée de créer différents jeux de caractères au gré des besoins de chaque langue

- À partir de 1987 la table ASCII étendue fut déclinée en de multiples variations (toujours codées sur un octet)
- Les 128 premiers caractères de tous les jeux ISO 8859 correspondent aux caractères ASCII
- La norme ISO 8859 contient aujourd'hui 16 tables, numérotées de 1 à 16 :1 pour les langues dites occidentales, 2 pour les langues d'Europe centrale/de l'Est, 5 pour le cyrillique, 6 pour l'arabe, 7 pour le grec, 8 pour l'hébreu, etc

Exemple : ISO-8859-1

La table ISO-8859-1 définit ce qu'elle appelle l'alphabet latin numéro 1 ou latin-1 : 191 caractères de l'alphabet latin

ISO/CEI 8859-1																
	x0	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**xA**	**xB**	**xC**	**xD**	**xE**	**xF**
0x	*inutilisés*															
1x																
2x		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
8x	*inutilisés*															
9x																
Ax		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬		®	¯
Bx	°	±	²	³	´	µ	¶	•	¸	¹	º	»	¼	½	¾	¿
Cx	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Dx	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
Ex	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
Fx	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Exemple : ISO-8859-5

Remarquez les 128 premiers caractères. . .

ISO/IEC 8859-5																
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	*unused*															
1x																
2x	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
8x	*unused*															
9x																
Ax	*NBSP*	Ё	Ђ	Ѓ	Є	S	І	Ї	J	Љ	Њ	Ћ	Ќ	*SHY*	Ў	Џ
Bx	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
Cx	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
Dx	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
Ex	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
Fx	№	ё	ђ	ѓ	є	s	і	ї	j	љ	њ	ћ	ќ	§	ў	џ

Exemple : ISO-8859-15

Aussi connue sous le nom de Latin-9 (?), c'est une extension directe d'ISO 1 (mais plus tardive), à l'exception de 8 caractères

Différences ISO 8859-1 / ISO 8859-15 :

Position	0xA4	0xA6	0xA8	0xB4	0xB8	0xBC	0xBD	0xBE
8859-1	¤	¦	¨	´	¸	¼	½	¾
8859-15	€	Š	š	Ž	ž	Œ	œ	Ÿ

Vous en aviez assez ?

Évidemment, parmi les encodages les plus courants se trouve des encodages « maison » :

- Microsoft : Windows1252 et al.
- Apple : MacRoman

Limitations

- Problèmes d'incomplétude ou d'affichage pour certaines langues
- Impossible d'écrire du russe et du français (hors ASCII) dans un seul et même fichier
- Problèmes d'erreurs dues à la quasi-superposition de certains encodages ($ se transformant en £, par exemple)
- Et le milliard de Chinois ?

Parenthèse sur le chinois

- Il existe deux jeux d'écriture du (des) chinois : le **simplifié**, utilisé dans la République Populaire de Chine et à Singapour et le **traditionnel**, plus répandu dans la diaspora, à Taïwan Hong Kong, Macao.

Parenthèse sur le chinois

- Il existe deux jeux d'écriture du (des) chinois : le **simplifié**, utilisé dans la République Populaire de Chine et à Singapour et le **traditionnel**, plus répandu dans la diaspora, à Taïwan Hong Kong, Macao.
- A chaque type d'écriture son encodage, en particulier : GB 2312-80 (ou Guobiao) pour le chinois simplifié, avec 6 763 caractères seulement (! !), et Big5 pour le traditionnel, avec 13 053 caractères.

Parenthèse sur le chinois

- Il existe deux jeux d'écriture du (des) chinois : le **simplifié**, utilisé dans la République Populaire de Chine et à Singapour et le **traditionnel**, plus répandu dans la diaspora, à Taïwan Hong Kong, Macao.
- A chaque type d'écriture son encodage, en particulier : GB 2312-80 (ou Guobiao) pour le chinois simplifié, avec 6 763 caractères seulement (! !), et Big5 pour le traditionnel, avec 13 053 caractères.
- 13 053 caractères ? ! Mais ils les mettent où ?

Parenthèse sur le chinois

- Il existe deux jeux d'écriture du (des) chinois : le **simplifié**, utilisé dans la République Populaire de Chine et à Singapour et le **traditionnel**, plus répandu dans la diaspora, à Taïwan Hong Kong, Macao.
- A chaque type d'écriture son encodage, en particulier : GB 2312-80 (ou Guobiao) pour le chinois simplifié, avec 6 763 caractères seulement (! !), et Big5 pour le traditionnel, avec 13 053 caractères.
- 13 053 caractères ? ! Mais ils les mettent où ?
- Les Chinois ont tout simplement plus de bits pour coder leurs jeux de caractères : 16 exactement (soit 2 octets)

Les éléphants d'Asie

Exercice

Combien de valeurs peut-on représenter sur 16 bits ?

Les éléphants d'Asie

Exercice

Combien de valeurs peut-on représenter sur 16 bits ?

→ $2x2x2x2x2x2x2x2x2x2x2x2x2x2x2x2 = 2^{16} = 65\ 536$

→ Ça suffit donc pour le chinois, même traditionnel. D'ailleurs, même 14 bits auraient suffi...($2^{14} = 16\ 384$)

Alors pourquoi 16 ?

→ Parce que c'est incomparablement plus pratique, étant donné que l'ordinateur gère des octets

Récapitulons

Récapitulons

- 1968 - ASCII (7 bits) : c'est pas Noël !

Récapitulons

- 1968 - ASCII (7 bits) : c'est pas Noël !
- 1981 - ASCII étendu (8 bits) : c'est Noël sans la водка

Récapitulons

- 1968 - ASCII (7 bits) : c'est pas Noël !
- 1981 - ASCII étendu (8 bits) : c'est Noël sans la водка
- 1987 - Les ISO (8 bits) : Noël avec водка à part, mais toujours pas d'€

Récapitulons

- 1968 - ASCII (7 bits) : c'est pas Noël !
- 1981 - ASCII étendu (8 bits) : c'est Noël sans la водка
- 1987 - Les ISO (8 bits) : Noël avec водка à part, mais toujours pas d'€
- 1997 - ISO-8859-15 (8 bits) : mise à jour d'ISO-8859-1, on passe à l'€ !

Récapitulons

- 1968 - ASCII (7 bits) : c'est pas Noël !
- 1981 - ASCII étendu (8 bits) : c'est Noël sans la водка
- 1987 - Les ISO (8 bits) : Noël avec водка à part, mais toujours pas d'€
- 1997 - ISO-8859-15 (8 bits) : mise à jour d'ISO-8859-1, on passe à l'€ !
- En parallèle, des encodages « maison » (8 bits) : mêmes défauts qu'ISO. Ne sont pas reconnus comme normes

Récapitulons

- 1968 - ASCII (7 bits) : c'est pas Noël !
- 1981 - ASCII étendu (8 bits) : c'est Noël sans la водка
- 1987 - Les ISO (8 bits) : Noël avec водка à part, mais toujours pas d'€
- 1997 - ISO-8859-15 (8 bits) : mise à jour d'ISO-8859-1, on passe à l'€ !
- En parallèle, des encodages « maison » (8 bits) : mêmes défauts qu'ISO. Ne sont pas reconnus comme normes
- On en reste à 8 bits = 256 caractères possibles, or le chinois en compte beaucoup plus !

Le projet Unicode

En 1988 (?) est élaboré un projet un peu fou : recenser tous les caractères de toutes les langues écrites existantes ou ayant existé et mettre au point une table de référence universelle capable de coder tout ça

L'entreprise colossale est aussitôt baptisée projet Unicode

Unicode se veut :

- universel : toutes les langues doivent être couvertes (même les plus rares)
- efficace : simple à analyser
- uniforme : nombre fixe de bits
- non-ambigu : une valeur = un seul caractère (une fois codé, éternel !)

La couverture d'Unicode

- Les 255 premiers caractères de la table Unicode sont ceux de la table ISO-5589-1

La couverture d'Unicode

- Les 255 premiers caractères de la table Unicode sont ceux de la table ISO-5589-1
- Première étape : les 63 586 caractères les plus utilisés sont réunis dans le Plan Multilingue de Base (PMB ou BMP en anglais)

La couverture d'Unicode

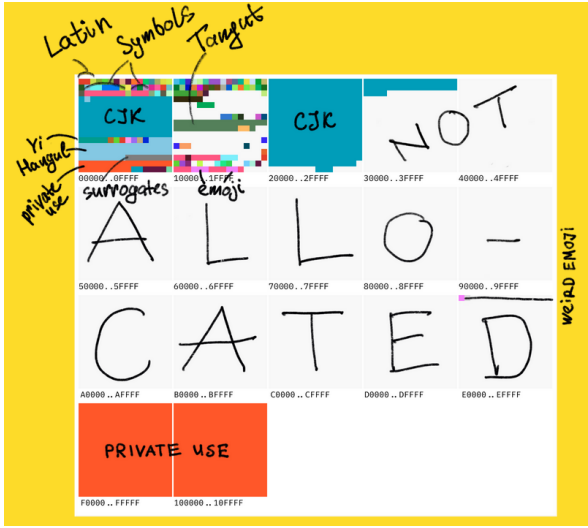
- Les 255 premiers caractères de la table Unicode sont ceux de la table ISO-5589-1
- Première étape : les 63 586 caractères les plus utilisés sont réunis dans le Plan Multilingue de Base (PMB ou BMP en anglais)
- Première publication de la norme Unicode, en 1991 : 65 536 caractères sont recensés et encodés

La couverture d'Unicode

- Les 255 premiers caractères de la table Unicode sont ceux de la table ISO-5589-1
- Première étape : les 63 586 caractères les plus utilisés sont réunis dans le Plan Multilingue de Base (PMB ou BMP en anglais)
- Première publication de la norme Unicode, en 1991 : 65 536 caractères sont recensés et encodés
- Aujourd'hui, Unicode version 12.1 (mai 2019) contient 137 994 caractères (dont 137 766 graphiques)

Ce qu'il y a dans Unicode

https ://tonsky.me/blog/unicode/



Le point de code Unicode

- En Unicode, une lettre correspond à quelque chose appelé un point de code qui n'est qu'un concept théorique
- Toute les lettres de tous les alphabets se sont vues attribuer un point de code par le consortium Unicode
- Ce point de code s'écrit : U+XXXX. Le U+ signifie « Unicode », et les X derrière sont en héxadécimal. U+FEC9 est la lettre arabe *Ain*. La lettre « A » correspond à U+0041
- Exemple : « Bonjour »

  U+0042 U+006F U+006E U+006A U+006F U+0075 U+0072

  → Exercice : que représente U+1F4A9 ?

C'est bien beau tout ça, mais l'ordinateur il en fait quoi ?

- Le consortium Unicode a prévu trois principaux formats « transformés » pour l'encodage des point de code en binaire. Ces formats de transformation sont baptisés UTF(Unicode Transformation Format)
- UTF-32, UTF-16, UTF-8
- Le principe intangible derrière ces transformations étant que tout point de code Unicode doit pouvoir être retrouvé sans ambiguïté à partir de sa version transformée

- langue naturelle vs langage artificiel
- langage de l'ordinateur