

Ch. 3

Caractères et chaînes de caractères

B. Quoitin
(bruno.quoitin@umons.ac.be)

Table des Matières

Codes de caractères

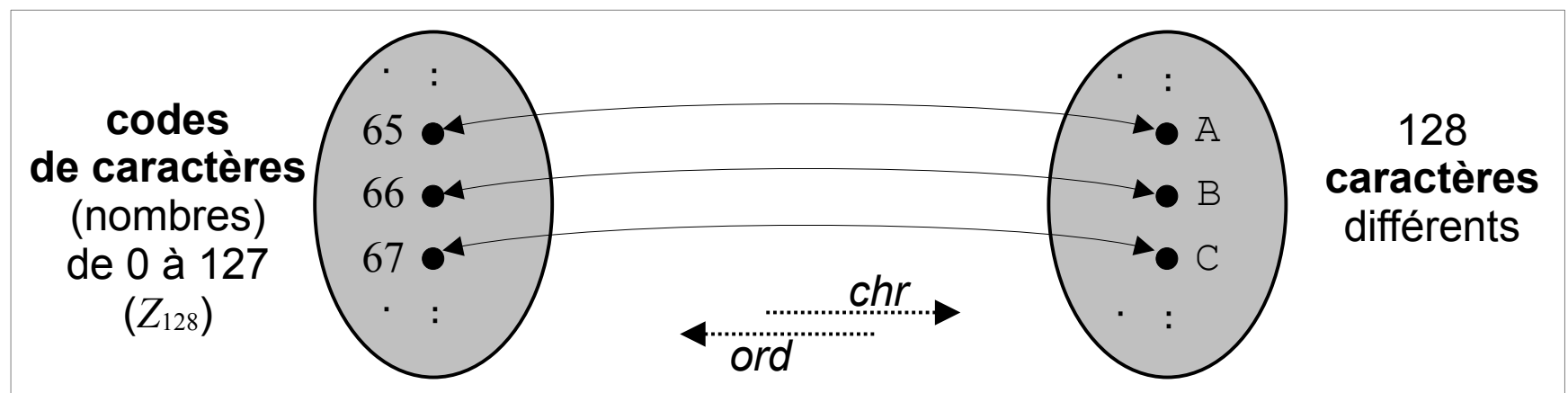
- Code ASCII
- Extensions ISO 8859
- Unicode

Chaînes de caractères

Représentation des Caractères

Code de caractères

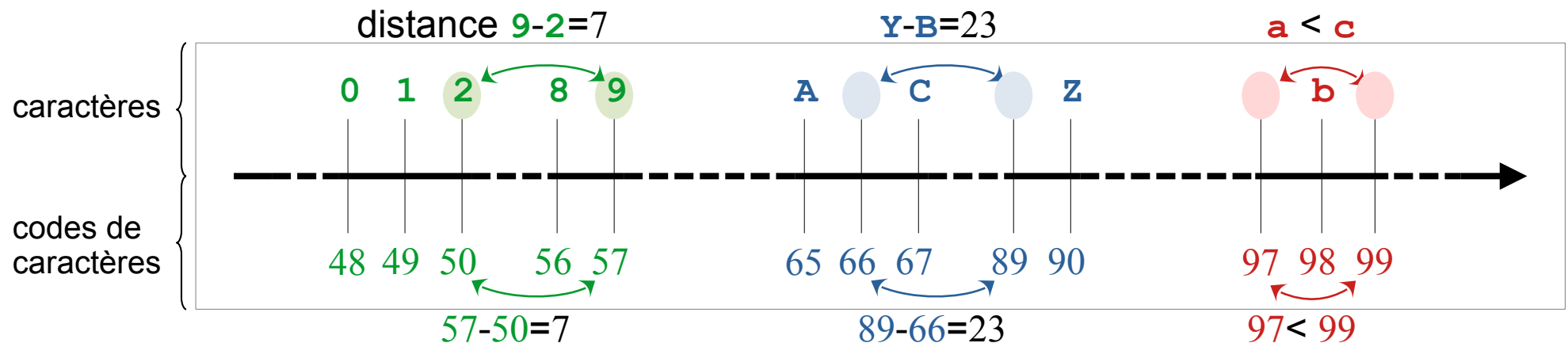
- Les ordinateurs peuvent manipuler des caractères.
 - Exemple : 26 *caractères de l'alphabet latin* en minuscule (a – z) et majuscule (A – Z), 10 *chiffres décimaux* (0 – 9) et *symboles spéciaux* (espace, point, virgule, signe moins, retour à la ligne, etc).
- Un caractère est représenté sous forme d'un nombre appelé **code de caractère** ou **code point**.
 - Exemple : le caractère A pourrait être représenté par 65.
- On désigne par **code**, un ensemble de correspondance entre caractères et codes de caractères. A chaque code de caractère correspond généralement un seul caractère, et vice versa (bijection).



Représentation des Caractères

Propriété des codes de caractères

- Certains sous-ensembles de caractères sont naturellement ordonnés. Par exemple,
 - les caractères qui correspondent aux **chiffres décimaux** (0 à 9)
 - les caractères qui correspondent aux lettres de l'alphabet latin **minuscules** (a à z) et **majuscules** (A à Z).



- Une propriété intéressante d'un code est que les code points correspondant aux caractères de ces sous-ensembles soient dans le **même ordre**

Table des Matières

Codes de caractères



- **Code ASCII**
- Extensions ISO 8859
- Unicode

Chaînes de caractères

Code ASCII

American Standard Code for Information Interchange (ASCII)

- Codes standard représentés **sur 7 bits** → codes de 0 à 127.
- Encode les caractères de l'alphabet latin en minuscule et majuscule, les chiffres décimaux, des signes de ponctuation et de codes de contrôle.

0	nul	1	soh	2	stx	3	etx	4	eot	5	enq	6	ack	7	bel	caractères de contrôle
8	bs	9	ht	10	nl	11	vt	12	np	13	cr	14	so	15	si	
16	dle	17	dc1	18	dc2	19	dc3	20	dc4	21	nak	22	syn	23	etb	
24	can	25	em	26	sub	27	esc	28	fs	29	gs	30	rs	31	us	
32	sp	33	!	34	"	35	#	36	\$	37	%	38	&	39	'	
40	(41)	42	*	43	+	44	,	45	-	46	.	47	/	
48	0	49	1	50	2	51	3	52	4	53	5	54	6	55	7	
56	8	57	9	58	:	59	;	60	<	61	=	62	>	63	?	
64	@	65	A	66	B	67	C	68	D	69	E	70	F	71	G	caractères imprimables
72	H	73	I	74	J	75	K	76	L	77	M	78	N	79	O	
80	P	81	Q	82	R	83	S	84	T	85	U	86	V	87	W	
88	X	89	Y	90	Z	91	[92	\	93]	94	^	95	_	
96	`	97	a	98	b	99	c	100	d	101	e	102	f	103	g	
104	h	105	i	106	j	107	k	108	l	109	m	110	n	111	o	
112	p	113	q	114	r	115	s	116	t	117	u	118	v	119	w	
120	x	121	y	122	z	123	{	124		125	}	126	~	127	del	

Notes: cette table peut être obtenue avec “man ascii”; le caractère ht est la tabulation horizontale (« tab »)


Code ASCII

Limitation - encodage prévu pour l'anglais

- Le code de caractères ASCII a été conçu pour l'encodage de documents en langue anglaise. Il n'est pas approprié pour l'encodage de caractères tels que “ç” en français et encore moins pour des caractères chinois, russes ou grecs, par exemple.
- Il existe des extensions (non standard) au code ASCII avec une représentation sur 8 bits. 128 codes de caractères supplémentaires sont alors disponibles. Ces codes de caractères ne sont cependant pas universels.

Table des Matières

Codes de caractères

- Code ASCII
-  • **Extensions ISO 8859**
- Unicode

Chaînes de caractères

Code ASCII

Tables ASCII étendues ISO 8859

- Ensemble de plusieurs extensions standard, majoritairement pour les formes d'écriture utilisées en Europe. Définies par l'ECMA et standardisées par l'ISO.
- Encodage sur **8 bits**; les 128 premiers codes correspondent à ASCII

- Exemple: ISO-8869-1 (Latin-1)

- ISO-8859-15 est similaire à ISO-8859-1 mais remplace 8 caractères et permet l'introduction du symbole Euro (€ 0xA4)

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL	STX	SOT	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
10	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
20	SP	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL
80																
90																
A0	NBSP	í	ñ	¿	¡	£	¥	¦	§	¨	©	ª	«	¬	®	¯
B0	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	À
	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï	
	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß	
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

Annotations:

- 0xA4: Euro symbol (€)
- 0xF6: Umlaut O (ö)
- 0xFB: Umlaut U (û)

ECMA : European Computer Manufacturer Association
ISO : International Organization for Standardization

Source: <https://www.charset.org/charsets/iso-8859-1>

Code ASCII

Tables ASCII étendues ISO 8859

- D'autres extensions existent pour supporter d'autres langues.
- Par exemple
 - ISO 8859-2 : langues slaves dérivées du Latin (Tchèque, Slovaque, Polonais, Hongrois, ...)
 - ISO 8859-3 : Turque, Maltais, Esperanto, ...
 - ISO 8859-4 : Estonien, Letton, Lituanien, ...
 - etc.



Table des Matières

Codes de caractères

- Code ASCII
- Extensions ISO 8859
- • **Unicode**

Chaînes de caractères

Unicode

Unicode

- Code destiné à remplacer ASCII et ses extensions.
- Supporte un **très grand nombre de caractères** différents et a l'ambition de supporter toutes les formes d'écriture. Supporte également l'écriture de certaines langues mortes.

和平

("paix")

الحكمة

("sagesse")



("Ptolémée")

- Les **code points d'Unicode vont de 0x000000 à 0x10FFFF**. Tous ne sont pas utilisés à ce stade et certains d'entre eux sont réservés. Les 256 premiers code points sont les mêmes qu'ISO-8859-1.
- Unicode est un standard qui continue à évoluer. La version 14 (Septembre 2021) définissait **144697** code points différents pour **159** formes d'écritures (appelées « blocs »).


Unicode

Unicode Transformation Format (UTF)

- Unicode supporte l'encodage d'un code point sur un nombre variable d'octets. Plusieurs encodages sont standardisés.
 - **UTF-32** : encodage de taille fixe, sur 4 octets (32 bits); représente exactement le *code point*
 - **UTF-16** : encodage de taille variable, sur 2 octets (16 bits) ou 4 octets (32 bits)
 - **UTF-8** : encodage de taille variable sur de 1 à 4 octets; les 128 premiers codes sont les mêmes qu'ASCII
- Aujourd'hui, l'encodage le plus utilisé est UTF-8 car il est le plus efficace en terme d'occupation d'espace.
- Note: le fonctionnement détaillé des encodages UTF-8 et UTF-16 sort du cadre de ce cours.

Unicode

Exemples d'encodages

	Caractère	Code ASCII	Code point Unicode	Encodage UTF-8	Encodage UTF-16	Encodage UTF-32
identique ASCII	H	48	48	48	00 48	00 00 00 48
	7	37	37	37	00 37	00 00 00 37
	CR	0a	a	0a	00 0a	00 00 00 0a
identique ISO-8859-1	û	/	fb	c3 bb	00 fb	00 00 00 fb
	ö	/	f6	c3 b6	00 f6	00 00 00 f6
	€	/	20ac	e2 82 ac	20 ac	00 00 20 ac
	☢	/	2622	e2 98 a2	26 22	00 00 26 22
	⌘	/	2318	e2 8c 98	23 18	00 00 23 18
	Ѓ	/	40a	d0 8a	04 0a	00 00 04 0a
	蠓	/	8813	e8 a0 93	88 13	00 00 88 13
		/	f09380a3	f0 93 80 a3	d8 0c dc 23	00 01 30 23

obtenu avec par exemple: `echo -n "Ѓ" | iconv -t utf-16be | hexdump -C`

Table des Matières

Codes de caractères

- Code ASCII
- Extensions ISO 8859
- Unicode



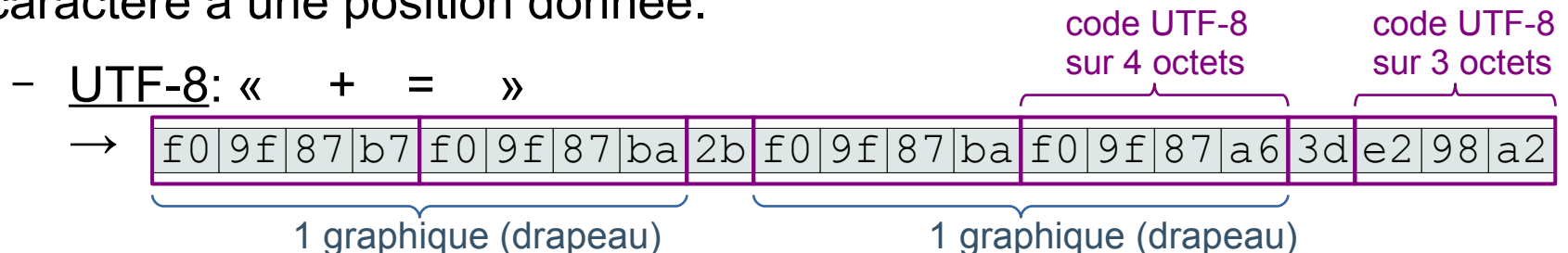
Chaînes de caractères

Chaînes de caractères

Chaîne de caractères

- Une chaîne de caractères est représentée comme la suite des codes des caractères qui la composent. Ces codes sont placés de manière consécutive en mémoire.
 - Exemple: la chaîne « **HELLO** » pourrait être représentée en mémoire par la suite des codes ASCII **72**, **69**, **76**, **76** et **79**.
- Si tous les codes ont la même taille (p.ex. 8 bits en ASCII), alors la représentation de la chaîne en mémoire peut être vue comme un tableau de codes et chaque caractère peut facilement être indexé.
 - ASCII: « HELLO » →

48	45	4C	4C	4F
----	----	----	----	----
- Dans le cas contraire, (p.ex. en UTF-8), il est plus difficile de trouver un caractère à une position donnée.



Chaînes de caractères

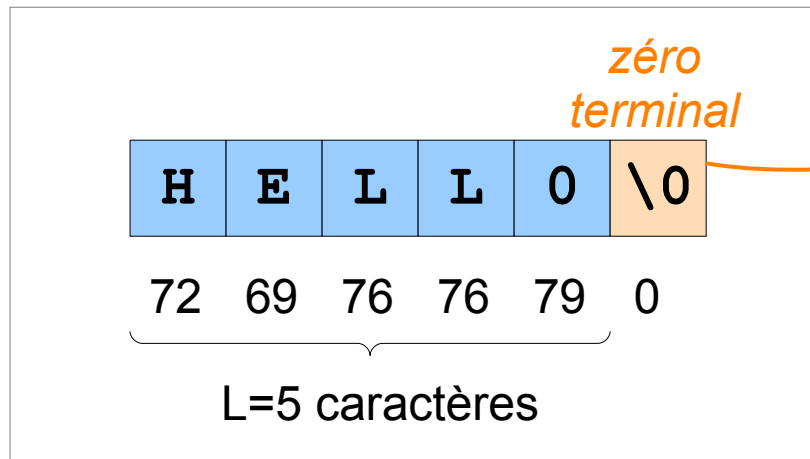
Chaîne de caractères

- Il existe plusieurs variantes de la représentation des chaînes de caractères en mémoire
 - sans longueur explicite: utilisation d'une sentinelle (p.ex. zéro terminal)
 - avec longueur explicite: utilisation d'un en-tête ou d'une variable contenant la longueur

Chaînes de caractères

Chaîne à zéro terminal

- Une chaîne à zéro terminal (*nul-terminated string*) est une représentation dans laquelle les caractères de la chaîne sont suivis d'un *caractère spécial de code 0* (sentinelle). Il n'y a donc pas de représentation explicite de la longueur.
- Le caractère 0 terminal est souvent noté '`\0`' dans les langages de programmation.



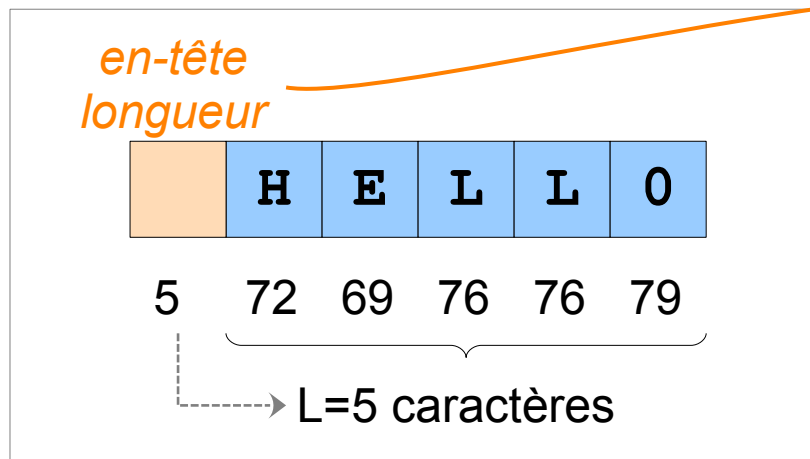
```
void str_print(char [] str) {  
    unsigned int i= 0;  
    while (str[i] != 0) {  
        putchar(str[i]);  
        i++;  
    }  
}
```

- Avantage: une telle chaîne peut avoir une longueur illimitée (seulement limitée par la mémoire disponible).
- Inconvénient: déterminer la longueur de la chaîne nécessite de la parcourir entièrement.

Chaînes de caractères

Chaîne avec en-tête longueur

- La chaîne est associée à, une zone mémoire contenant une représentation (explicite) de sa *longueur*. Cette zone mémoire peut être composée d'un ou plusieurs octets.
- La longueur peut soit précéder la chaîne en mémoire ou se trouver dans une variable associée.



```
void str_print(char [] str) {  
    unsigned int i = 0;  
    while (i < str[0]) {  
        putchar(str[i+1]);  
        i++;  
    }  
}
```

- Avantage: il est possible de connaître rapidement la longueur de la chaîne.
- Inconvénient: la longueur maximale d'une telle chaîne est limitée par la taille réservée à sa longueur. Par exemple, si l'en-tête fait 8 bits, la longueur maximale est 255 caractères.

Références

- **Computer Organization and Design: The Hardware/Software Interface, 4th Edition, D. Patterson and J. Hennessy, Morgan-Kaufmann, 2009**
- **<http://www.unicode.org>**
- **<http://www.charset.org>**

Remerciements

Merci à toutes les personnes qui ont permis par leurs remarques de corriger et d'améliorer ces notes de cours.