#### Représenter un texte

Yann Strozecki yann.strozecki@uvsq.fr

Université de Versailles St-Quentin-en-Yvelines

Année universitaire 2022-2023

On veut représenter des mots sur ordinateur. Qu'est-ce qu'un mot formellement ?

On veut représenter des mots sur ordinateur. Qu'est-ce qu'un mot formellement ?

On a un alphabet fini :  $\Sigma = \{a, b, c, d\}$ . C'est un ensemble de *symboles*.

On veut représenter des mots sur ordinateur. Qu'est-ce qu'un mot formellement ?

On a un alphabet fini :  $\Sigma = \{a, b, c, d\}$ .

C'est un ensemble de symboles.

On créé des mots par concaténation de symboles : abacadaba.

Un mot est une *suite finie* de symboles de l'alphabet.

**Exemple**: Les nombres entiers positifs sont des mots, dont les symboles sont les chiffres.

On veut représenter des mots sur ordinateur. Qu'est-ce qu'un mot formellement ?

On a un alphabet fini :  $\Sigma = \{a, b, c, d\}$ .

C'est un ensemble de symboles.

On créé des mots par concaténation de symboles : abacadaba. Un mot est une *suite finie* de symboles de l'alphabet.

**Exemple**: Les nombres entiers positifs sont des mots, dont les symboles sont les chiffres.

 $\Sigma^k$  est l'ensemble des mots de k lettres sur  $\Sigma$ .  $\Sigma^*$  est l'ensemble de tous les mots sur  $\Sigma$ .

**Exemple**:  $abc \in \Sigma^3$  et  $ab \notin \Sigma^4$ .

Notre objectif est de représenter toute information par un mot de  $\{0,1\}^*$ .

## Encodage de mots et décodage

Un codage est une fonction injective de  $\Sigma \to \{0,1\}^*$ . Par exemple E(a)=00, E(b)=01, E(c)=10, E(d)=11.

## Encodage de mots et décodage

Un codage est une fonction injective de  $\Sigma \to \{0,1\}^*$ . Par exemple E(a)=00, E(b)=01, E(c)=10, E(d)=11.

On étend cette fonction en une fonction de  $\Sigma^* \to \{0,1\}^*$  par concaténation :

$$E(w_1w_2\dots w_k) = E(w_1)E(w_2)\dots E(w_k)$$

Par exemple E(abacadaba) = 000100100011000100.

#### Décodage

L'application donnée par  $E(a)=0, E(b)=10, E(c)=01, E(d)=110, \ {\rm n'est\ pas\ acceptable}$  pour définir un codage.

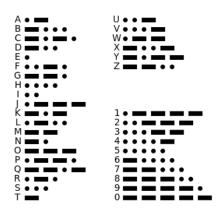
### Décodage

L'application donnée par  $E(a)=0, E(b)=10, E(c)=01, E(d)=110, \ {\rm n'est\ pas\ acceptable}$  pour définir un codage.

Pour résoudre ce problème, on utilise des codages à longueur fixe. Proposer un tel codage pour  $\{a,b,c,d,e\}$ . Est-il efficace en taille?

#### L'ancêtre de tous les "codages binaires", le *Morse* International Morse Code

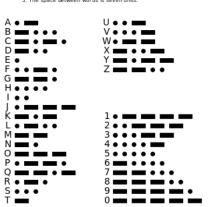
- 1. The length of a dot is one unit.
- 2. A dash is three units.
- 3. The space between parts of the same letter is one unit.
- 4. The space between letters is three units.
- The space between letters is three units.The space between words is seven units.



# L'ancêtre de tous les "codages binaires", le *Morse*

#### International Morse Code

- 1. The length of a dot is one unit.
- 2. A dash is three units.
- 3. The space between parts of the same letter is one unit.
- 4. The space between letters is three units.
- The space between letters is three units.
  The space between words is seven units.



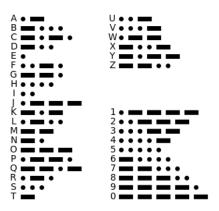
Le mot S.O.S. se note :

.. - - - ...

#### L'ancêtre de tous les "codages binaires", le Morse

#### International Morse Code

- 1. The length of a dot is one unit.
- 2. A dash is three units.
- 3. The space between parts of the same letter is one unit.
- 4. The space between letters is three units.
- 5. The space between words is seven units.



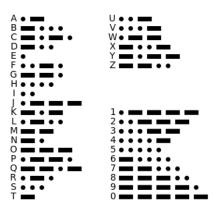
Le mot S.O.S. se note:

Code de longueur variable. La taille des caractères dépend de leur fréquence.

### L'ancêtre de tous les "codages binaires", le Morse

#### International Morse Code

- 1. The length of a dot is one unit.
- 2. A dash is three units.
- 3. The space between parts of the same letter is one unit.
- 4. The space between letters is three units.
- The space between letters is three units.
  The space between words is seven units.



Le mot S.O.S. se note :

... - - - ...

Code de longueur variable. La taille des caractères dépend de leur fréquence.

Non ambigu car il y a un troisième symbole, le silence.

Le code ASCII représente chaque caractère sur 7 bits, c'est un code de taille fixe. À chaque caractère est associée une configuration de 8 bits (1 octet), le bit de poids fort (le plus à gauche) est toujours égal à zero.

▶ Les codes compris entre 0 et 31 sont des caractères de contrôle. Ils sont utilisés pour indiquer des actions telles que passer à la ligne (CR, LF), émettre un bip sonore (BEL), etc.

- ► Les codes compris entre 0 et 31 sont des caractères de contrôle. Ils sont utilisés pour indiquer des actions telles que passer à la ligne (CR, LF), émettre un bip sonore (BEL), etc.
- ▶ Les lettres se suivent dans l'ordre alphabétique (codes 65 à 90 pour les majuscules, 97 à 122 pour les minuscules), ce qui simplifie les comparaisons et le passage d'une lettre à la suivante.

- ► Les codes compris entre 0 et 31 sont des caractères de contrôle. Ils sont utilisés pour indiquer des actions telles que passer à la ligne (CR, LF), émettre un bip sonore (BEL), etc.
- ► Les lettres se suivent dans l'ordre alphabétique (codes 65 à 90 pour les majuscules, 97 à 122 pour les minuscules), ce qui simplifie les comparaisons et le passage d'une lettre à la suivante.
- ► On passe des majuscules au minuscules en modifiant le 5ième bit, ce qui revient à ajouter 32 au code ASCII décimal.

- ► Les codes compris entre 0 et 31 sont des caractères de contrôle. Ils sont utilisés pour indiquer des actions telles que passer à la ligne (CR, LF), émettre un bip sonore (BEL), etc.
- ► Les lettres se suivent dans l'ordre alphabétique (codes 65 à 90 pour les majuscules, 97 à 122 pour les minuscules), ce qui simplifie les comparaisons et le passage d'une lettre à la suivante.
- ➤ On passe des majuscules au minuscules en modifiant le 5ième bit, ce qui revient à ajouter 32 au code ASCII décimal.
- ▶ Les chiffres sont rangés dans l'ordre croissant (codes 48 à 57), et les 4 bits de poids faible définissent la valeur en binaire du chiffre.

#### The table

Dec	Hex	Name	Char	Ctrl-char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	0	Null	NUL	CTRL-@	32	20	Space	64	40	0	96	60	
1	1	Start of heading	SOH	CTRL-A	33	21	1	65	41	A	97	61	a
2	2	Start of text	STX	CTRL-B	34	22	"	66	42	В	98	62	b
3	3	End of text	ETX	CTRL-C	35	23	#	67	43	C	99	63	c
4	4	End of xmit	EOT	CTRL-D	36	24	\$	68	44	D	100	64	d
5	5	Enquiry	ENQ	CTRL-E	37	25	%	69	45	E	101	65	е
6	6	Acknowledge	ACK	CTRL-F	38	26	8.	70	46	F	102	66	f
7	7	Bell	BEL	CTRL-G	39	27		71	47	G	103	67	g
8	8	B ackspace	BS	CTRL-H	40	28	(	72	48	н	104	68	h
9	9	Horizontal tab	HT	CTRL-I	41	29	)	73	49	I	105	69	i
10	OA.	Line feed	LF	CTRL-J	42	2A		74	4A	J	106	6.4	j
11	OB	Vertical tab	VT	CTRL-K	43	2B	+	75	4B	K	107	6B	k
12	OC.	Form feed	FF	CTRL-L	44	2C	,	76	4C	L	108	6C	1
13	OD.	Carriage feed	CR	CTRL-M	45	2D	-	77	4D	М	109	6D	m
14	Œ	Shift out	SO	CTRL-N	46	2E		78	4E	N	110	6E	n
15	0F	Shift in	SI	CTRL-O	47	2F	/	79	4F	0	111	6F	0
16	10	Data line escape	DLE	CTRL-P	48	30	0	80	50	P	112	70	р
17	11	Device control 1	DC1	CTRL-Q	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	DC2	CTRL-R	50	32	2	82	52	R	114	72	r
19	13	Device control 3	DC3	CTRL-S	51	33	3	83	53	S	115	73	s
20	14	Device control 4	DC4	CTRL-T	52	34	4	84	54	T	116	74	t
21	15	Neg acknowledge	NAK	CTRL-U	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	SYN	CTRL-V	54	36	6	86	56	V	118	76	٧
23	17	End of xmit block	ETB	CTRL-W	55	37	7	87	57	w	119	77	w
24	18	Cancel	CAN	CTRL-X	56	38	8	88	58	X	120	78	×
25	19	End of medium	EM	CTRL-Y	57	39	9	89	59	Υ	121	79	У
26	1A	Substitute	SUB	CTRL-Z	58	ЗА	:	90	5A	Z	122	7A	z
27	1B	Escape	ESC	CTRL-[	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	FS	CTRL-\	60	3C	<	92	5C	\	124	7C	1
29	1D	Group separator	GS	CTRL-]	61	3D	-	93	5D	j	125	7D	}
30	1E	Record separator	RS	CTRL-^	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	US	CTRL	63	3F	?	95	5F	_	127	7F	DEL

#### Un exemple

La phrase ASCII 7 bits s'écrit :

- ► En hexadécimal 41 53 43 49 49 20 37 20 62 69 74 73

#### La table

Mais où sont les caractères accentués?

#### La table

Dec	Hex	Char	Dec	Hex	Char	1	Dec	Hex	Char	Dec	Hex	Char
128	80	Ç	160	A0	á		192	СО	L	224	E0	α
129	81	ű	161	A1	lí		193	C1	1	225	E1	В
130	82	é	162	A2	ó		194	C2	Т	226	E2	ÌΓ
131	83	â	163	А3	ú		195	C3	ŀ	227	E3	İπ
132	84	ä	164	A4	ñ		196	C4		228	E4	Σ
133	85	à	165	A5	Ñ		197	C5	+	229	E5	σ
134	86	å	166	A6	<u>a</u>		198	C6		230	E6	μ
135	87	Ç	167	A7	0		199	C7	ŀ	231	E7	۲
136	88	ê	168	A8	Ĺ		200	C8	Ŀ	232	E8	φ
137	89	ë	169	Α9	-		201	C9	F	233	E9	8
138	8A	è	170	AA	٦.		202	CA	Ŧ	234	EA	Ω
139	8B	î	171	AB	1/2		203	СВ	Ŧ	235	EB	8
140	8C	î	172	AC	1/4		204	cc	ŀ	236	EC	∞
141	8D		173	AD	i		205	CD	=	237	ED	φ
142	8E	Ä	174	AE	<b>«</b>		206	CE	1	238	EE	Ε
143	8F	Å	175	AF	»		207	CF		239	EF	N
144	90	É	176	B0	1 1		208	D0	1	240	F0	≡
145	91	æ	177	B1	111		209	D1	₹	241	F1	<u> </u>
146	92	Æ	178	B2			210	D2	I	242	F2	}
147	93	ô	179	В3	ΙT		211	D3		243	F3	<u>*</u> \\
148	94	ö	180	B4	-		212	D4	L	244	F4	
149	95	ò	181	B5	Н		213	D5	F	245	F5	J
150	96	û	182	В6	-		214	D6	ŗ	246	F6	÷
151	97	ù	183	В7	Ī		215	D7	+	247	F7	≈
152	98	ÿ	184	B8	7		216	D8	1	248	F8	0
153	99	0	185	В9	-		217	D9		249	F9	•
154	9A	Ü	186	BA			218	DA	1	250	FA	:
155	9B	¢	187	вв	٦		219	DB		251	FB	1 1
156	9C	£	188	BC	1		220	DC		252	FC	n
157	9D	¥	189	BD	ı,		221	DD	L.	253	FD	2
158	9E	Pt	190	BE	4		222	DE	I I	254	FE	•
159	9F	f	191	BF	٦		223	DF	•	255	FF	

#### La table

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
128	80	С	160	A0	á	192	СО	L	224	EO	α
129	81	Ç:u'e â:a	161	A1	í	193	C1	1	225	E1	В
130	82	é	162	A2	ó	194	C2	Т	226	E2	Г
131	83	â	163	А3	ú	195	C3	<del> </del>	227	E3	Σ
132	84	ä	164	A4	ñ	196	C4	<u>-</u>	228	E4	Σ
133	85	à	165	A5	Ñ	197	C5	+	229	E5	σ
134	86	å	166	A6	0	198	C6	F	230	E6	μ
135	87	Ç	167	A7	ō	199	C7	j-	231	E7	ľ
136	88	ê	168	A8	i l	200	C8	Ŀ	232	E8	φ
137	89	ë	169	A9	-	201	C9	F	233	E9	θ
138	8A	ê:e e:iî;î	170	AA	-	202	CA	T	234	EA	
139	8B	ï	171	AB	½ ¼	203	СВ	T	235	EB	δ
140	8C	î	172	AC	1/4	204	cc	ŀ	236	EC	
141	8D	1	173	AD	i	205	CD	=	237	ED	ф
142	8E	Ä	174	AE	(	206	CE	1	238	EE	E
143	8F	Å	175	AF	»	207	CF		239	EF	N .
144	90	É	176	B0	- 6	208	D0	Т	240	F0	≣
145	91	æ	177	В1		209	D1	₹	241	F1	ŧ
146	92	Æ	178	B2		210	D2	Ţ	242	F2	<u>*</u> <u>\$</u>
147	93	ô	179	В3	ΙĪΙ	211	D3		243	F3	Š
148	94	0	180	B4	-	212	D4	L	244	F4	Ī
149	95	ò	181	B5		213	D5	F	245	F5	J
150	96	û	182	В6	H 1	214	D6	r	246	F6	÷
151	97	ù	183	В7	T	215	D7	+	247	F7	≈
152	98	ÿ	184	B8	7	216	D8	<del>†</del>	248	F8	•
153	99	0	185	В9	-	217	D9		249	F9	•
154	9A	Ü	186	BA		218	DA	<u>T</u>	250	FA	1:
155	9B	£	187	BB	🤊	219	DB		251	FB	1
156	9C	Ł	188	BC	4	220	DC		252	FC	n .
157	9D	¥	189	BD	J.	221	DD	L I	253	FD	2
158	9E	Pt	190	BE	J	222	DE	<u> </u>	254	FE	٠
159	9F	f	191	BF	٦	223	DF	•	255	FF	

Problème d'encodage : @ ë □ . . .

#### Le codage Unicode

#### Problèmes:

- Comment encoder les alphabets non-occidentaux avec de nombreux caractères?
- Comment permettre de rajouter des caractères (extensibilité)?
- Comment mettre tout le monde d'accord?
- Comment être compatible avec les anciens formats, de type ASCII?

### Le codage Unicode

#### Problèmes:

- Comment encoder les alphabets non-occidentaux avec de nombreux caractères?
- Comment permettre de rajouter des caractères (extensibilité)?
- ► Comment mettre tout le monde d'accord?
- Comment être compatible avec les anciens formats, de type ASCII?

#### La solution : le standard Unicode.

- On utilise jusqu'à 4 octets : UTF8, UTF16, UTF32.
- On utilise un encodage de longueur variable (les 1 du début du premier octet donnent cette longueur).
- Une base de donnée officielle contenant le nom et le numéro de chaque symbole.
- ► Les caractères ASCII gardent le même code sur un octet en UTF8, pour la compatibilité.

#### Texte enrichi

Un texte n'est pas constitué uniquement de caractères!

Il y a aussi une structure :

- placement du texte
- ▶ liste
- ► lien
- ▶ image . . .

#### Texte enrichi

Un texte n'est pas constitué uniquement de caractères!

Il y a aussi une structure :

- placement du texte
- ► liste
- ► lien
- ▶ image . . .

On utilise des langages descriptifs de document comme le <a href="http://https