La saisie multilingue

Estelle Debouy

Un texte dans une langue quelconque peut se décomposer en une suite de caractères. En français, les caractères sont les lettres, les chiffres, la ponctuation, les espaces. Le œ est un caractère à part entière, parce qu'il a une existence propre dans l'orthographe de la langue.

Un fichier sur un ordinateur n'est ni plus ni moins qu'une suite d'octets, c'est-à-dire en quelque sorte des nombres entre 0 et 255. Toute la difficulté réside dans le fait de représenter une suite de caractères par une suite d'octets. Nous allons voir que ce n'est pas si simple.

1 Les premières solutions

Tout commence par une constatation très simple : les premiers informaticiens parlaient anglais. Et l'anglais s'écrit avec peu de chose : deux fois 26 lettres, 10 chiffres, une trentaine de signes de ponctuation, de signes mathématiques, sans oublier le symbole dollar : avec 95 caractères au total on peut se débrouiller. À l'époque dont je parle, on ne pouvait utiliser que la moitié des octets, soit 128 valeurs. On en a pris 33 comme caractères de « contrôle » (comme le retour à la ligne par exemple), plus les 95 dont on avait besoin pour écrire l'anglais. On a attribué des numéros à toutes ces valeurs : le code ASCII est né[1].

1. Soit l'American Standard Code for Information Interchange (en français, le code américain normalisé pour l'échange d'informations).

ASCII TABLE

Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]
1	1	1	1	[START OF HEADING]
2	2	10	2	[START OF TEXT]
3	3	11	3	[END OF TEXT]
4	4	100	4	[END OF TRANSMISSION]
5	5	101	5	[ENQUIRY]
6	6	110	6	[ACKNOWLEDGE]
7	7	111	7	[BELL]
8	8	1000	10	[BACKSPACE]
9	9	1001	11	[HORIZONTAL TAB]
10	A	1010	12	[LINE FEED]
11	B	1011	13	[VERTICAL TAB]
12	C	1100	14	[FORM FEED]
13	D	1101	15	[CARRIAGE RETURN]
14	E	1110	16	[SHIFT OUT]
15	F	1111	17	[SHIFT IN]
16	10	10000	20	[DATA LINK ESCAPE]
17	11	10001	21	[DEVICE CONTROL 1]
18	12	10010	22	[DEVICE CONTROL 2]
19	13	10011	23	[DEVICE CONTROL 3]
20	14	10100	24	[DEVICE CONTROL 4]
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]
22	16	10110	26	[SYNCHRONOUS IDLE]
23	17	10111	27	[ENG OF TRANS. BLOCK]
24	18	11000	30	[CANCEL]
25	19	11001	31	[END OF MEDIUM]
26	1A	11010	32	[SUBSTITUTE]
27	1B	11011	33	[ESCAPE]
28	1C	11100	34	[FILE SEPARATOR]
29	1D	11101	35	[GROUP SEPARATOR]
30	1E	11110	36	[RECORD SEPARATOR]
31	1F	11111	37	[UNIT SEPARATOR]
32	20	100000	40	[SPACE]
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	0
49	31	110001	61	1
50	32	110010	62	2
51	33	110011	63	3
52	34	110100	64	4
53	35	110101	65	5
54	36	110110	66	6
55	37	110111	67	7
56	38	111000	70	8
57	39	111001	71	9
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	A
66	42	1000010	102	B
67	43	1000011	103	C
68	44	1000100	104	D
69	45	1000101	105	E
70	46	1000110	106	F
71	47	1000111	107	G
72	48	1001000	110	H
73	49	1001001	111	I
74	4A	1001010	112	J
75	4B	1001011	113	K
76	4C	1001100	114	L
77	4D	1001101	115	M
78	4E	1001110	116	N
79	4F	1001111	117	O
80	50	1010000	120	P
81	51	1010001	121	Q
82	52	1010010	122	R
83	53	1010011	123	S
84	54	1010100	124	T
85	55	1010101	125	U
86	56	1010110	126	V
87	57	1010111	127	W
88	58	1011000	130	X
89	59	1011001	131	Y
90	5A	1011010	132	Z
91	5B	1011011	133	[
92	5C	1011100	134	\
93	5D	1011101	135	]
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	`
97	61	1100001	141	a
98	62	1100010	142	b
99	63	1100011	143	c
100	64	1100100	144	d
101	65	1100101	145	e
102	66	1100110	146	f
103	67	1100111	147	g
104	68	1101000	150	h
105	69	1101001	151	i
106	6A	1101010	152	j
107	6B	1101011	153	k
108	6C	1101100	154	l
109	6D	1101101	155	m
110	6E	1101110	156	n
111	6F	1101111	157	o
112	70	1110000	160	p
113	71	1110001	161	q
114	72	1110010	162	r
115	73	1110011	163	s
116	74	1110100	164	t
117	75	1110101	165	u
118	76	1110110	166	v
119	77	1110111	167	w
120	78	1111000	170	x
121	79	1111001	171	y
122	7A	1111010	172	z
123	7B	1111011	173	{
124	7C	1111100	174	|
125	7D	1111101	175	}
126	7E	1111110	176	~
127	7F	1111111	177	[DEL]

Mais très vite on a aussi voulu saisir du français ou de l'allemand sur son ordinateur. Heureusement, entre temps, il était devenu possible d'utiliser les valeurs laissées de côté par l'ASCII. Dans cette place, il a été possible de caser les caractères accentués et divers autres symboles utilisés par les langues d'Europe de l'ouest. Dans ces 128 valeurs, il n'y a hélas pas la place de caser les caractères pour les langues occidentales et l'alphabet cyrillique et l'alphabet grec et l'alphabet hébreu.

Pour pouvoir taper plusieurs langues sur un même ordinateur et pour que les ordinateurs puissent communiquer entre eux, des organismes de standardisation ont créé des tables de correspondance, comme l'ISO-8859-1, qui propose un jeu de caractères pour les langues occidentales, l'ISO-8859-5 qui offre du cyrillique, l'ISO-8859-7, qui propose du grec, etc. Mais, malgré tout, il n'a pas été possible de faire rentrer les 1945 idéogrammes du japonais officiel dans un octet, ni les 11 172 syllabes coréennes, ni les dizaines de milliers d'idéogrammes chinois qu'on arrive à recenser…

2 L'unicode

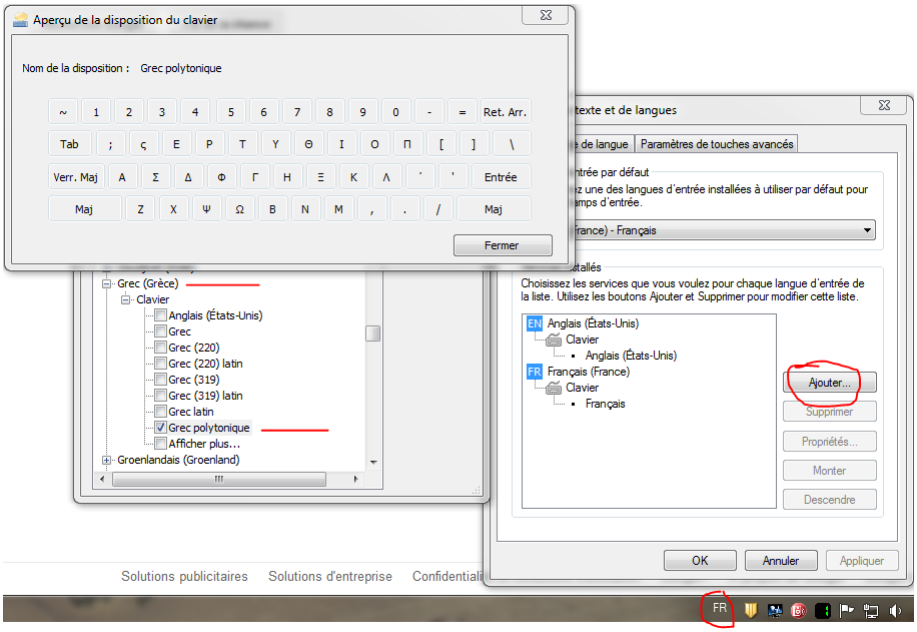
Pour résoudre durablement tous ces problèmes de langues, il s'est formé un consortium regroupant des grands noms de l'informatique et de la linguistique : le consortium Unicode. Sa tâche : recenser et numéroter tous les caractères existant dans toutes les langues du monde. Est donc né un jeu universel de caractères, acceptant plusieurs encodages [2], l'unicode. En 2007, le standard publié comportait environ 60 000 caractères. Avec l'unicode, un texte dans n'importe quelle langue peut se représenter comme une suite de nombres. Quelle simplification ! L'un des encodages les plus utilisés est l'UTF-8 car il présente l'avantage d'être compatible avec l'ASCII, de sorte que les parties écrites avec l'alphabet latin de base d'un texte codé en UTF-8 seront à peu près lisibles même avec un logiciel qui ne comprend pas ce codage.

Prenons, par exemple, le sigma majuscule : il a été encodé avec le point de code U+03A3 :

Cela dit, comment faire pour saisir une citation en espagnol, chinois, arabe ou grec ancien au milieu d'un texte en français ? Il faut non seulement disposer d'une police unicode (comme Gentium), mais encore d'un clavier virtuel qui vous permet de savoir où se trouvent les caractères.

Ainsi, pour être en mesure de saisir du texte dans une langue autre que le français, vous devez attribuer à votre clavier la langue de saisie souhaitée. Par exemple, pour taper *οὐκ ἔλαβον πόλιν*, vous devez configurer votre clavier de façon à saisir π quand vous tapez sur la touche P. Pour ce faire, il suffit de cliquer du droit sur l'icône FR (qui apparaît en bas de votre écran sur votre bureau), puis de choisir « Paramètres » et « Ajouter ». Si vous ne voyez pas l'icône en question, allez dans le panneau de configuration et choisissez « Horloge, langue et région ». Vous aurez alors la possibilité d'ajouter une langue.

2. Unicode est basiquement un jeu de caractères (un ensemble de caractères auxquels on attribue à chacun un point de code unique) et non un encodage (façon de représenter ce point de code en mémoire). C'est ici que la distinction prend tout son sens. Auparavant, les deux se confondaient, puisque tous les jeux de caractères étaient associés à un encodage simple.



Il vous est aussi possible d'utiliser des claviers virtuels en ligne, comme celui disponible à l'adresse suivante : http://www.lexilogos.com/clavier/multilingue.htm