

MASTER DOCUMENTS NUMÉRIQUES MULTILINGUES

Gestion Informatique du Multilinguisme

cours PLURITAL 2006-2007

première session – mardi 23 janvier 2007

3 h – tous documents autorisés

Les questions sont indépendantes et peuvent être traitées dans un ordre quelconque

1. Soit l'écran reproduit par la *figure 1* présentant la texte d'un courriel lu avec Thunderbird. sur un ordinateur dont le système d'exploitation est Windows XP localisé en France. Ce courriel contient le texte d'un article du quotidien tchèque « Lidové Noviny », envoyé depuis le site du journal qui offre la possibilité au lecteur de s'envoyer par courriel l'article qu'il est en train de consulter.
A première vue, le courriel est pratiquement illisible tant le rendu des caractères diacrités du tchèque est mauvais comme vous pouvez le constater en le comparant à celui reproduit par la *figure 2* qui lui est tout à fait correct.
 - a) Comment pouvez-vous expliquer la cause de l'affichage reproduit par la *figure 1*? Vous devez argumenter et justifier votre réponse.
 - b) Comment pouvez-vous remédier à ce problème et retrouver un rendu correct tel qu'il est reproduit par la *figure 2*?
2. La *figure 3* représente le « dump » d'un fichier texte effectué avec un éditeur hexadécimal. En observant les codes d'une part et le rendu des caractères d'autre part, dites quel est le format d'encodage utilisé: 8 bits ou multi-octets et dans ce dernier vous préciserez exactement lequel. Vous devez expliquer les raisons de votre choix.
3. Soit un programme C qui compte le nombre de caractères <œ> contenu dans un texte. Le code source de ce programme est présenté p.5 de ce document. Il a été entré à l'aide d'un éditeur de programme qui est une application Windows. Lorsque le programme est exécuté sur un texte saisi à l'aide d'une application Linux, il affiche que le fichier contient 0 caractère <œ> même s'il en contient plusieurs. L'algorithme est cependant tout à fait correct. Pouvez-vous expliquer ce résultat?

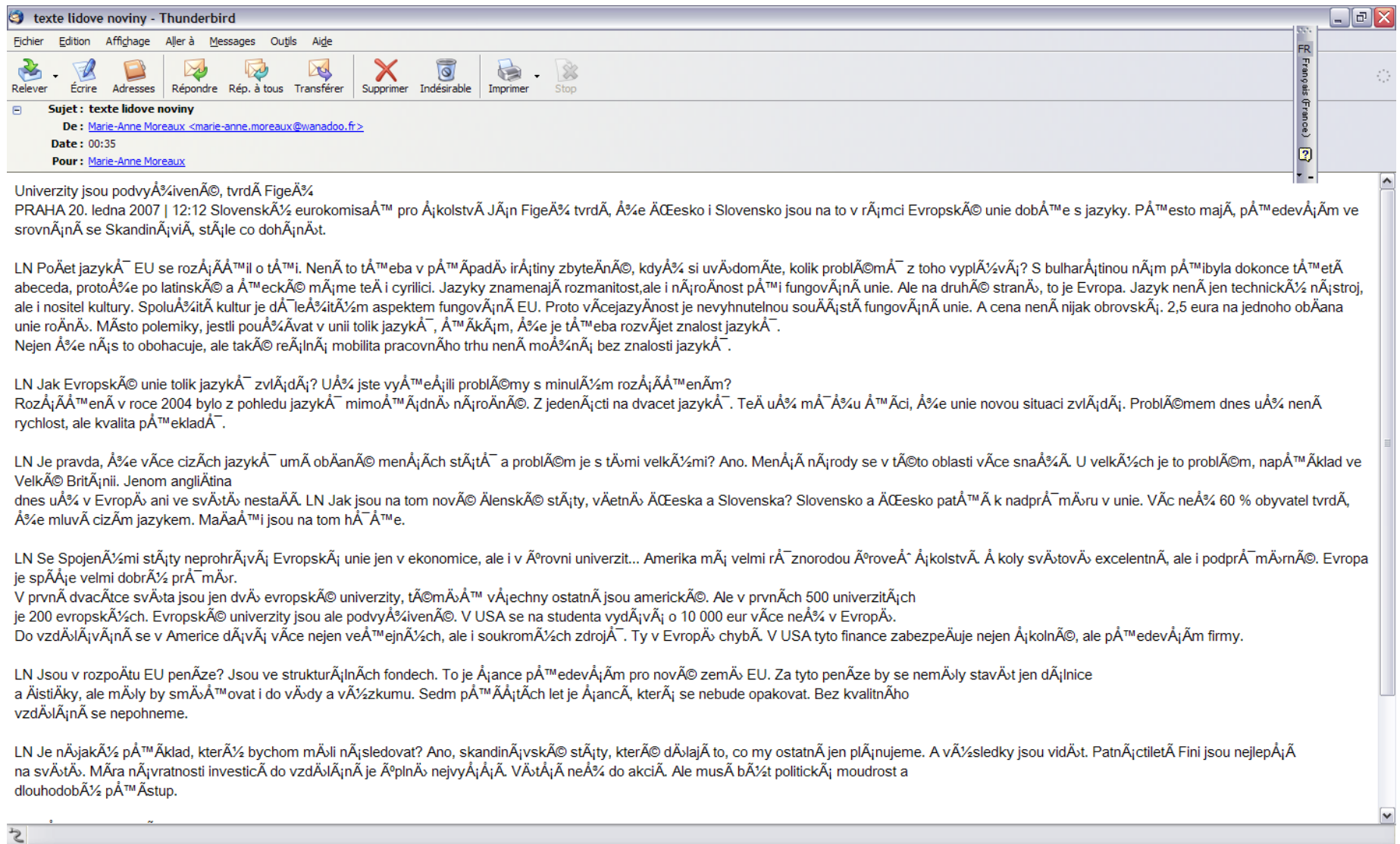


Figure 1

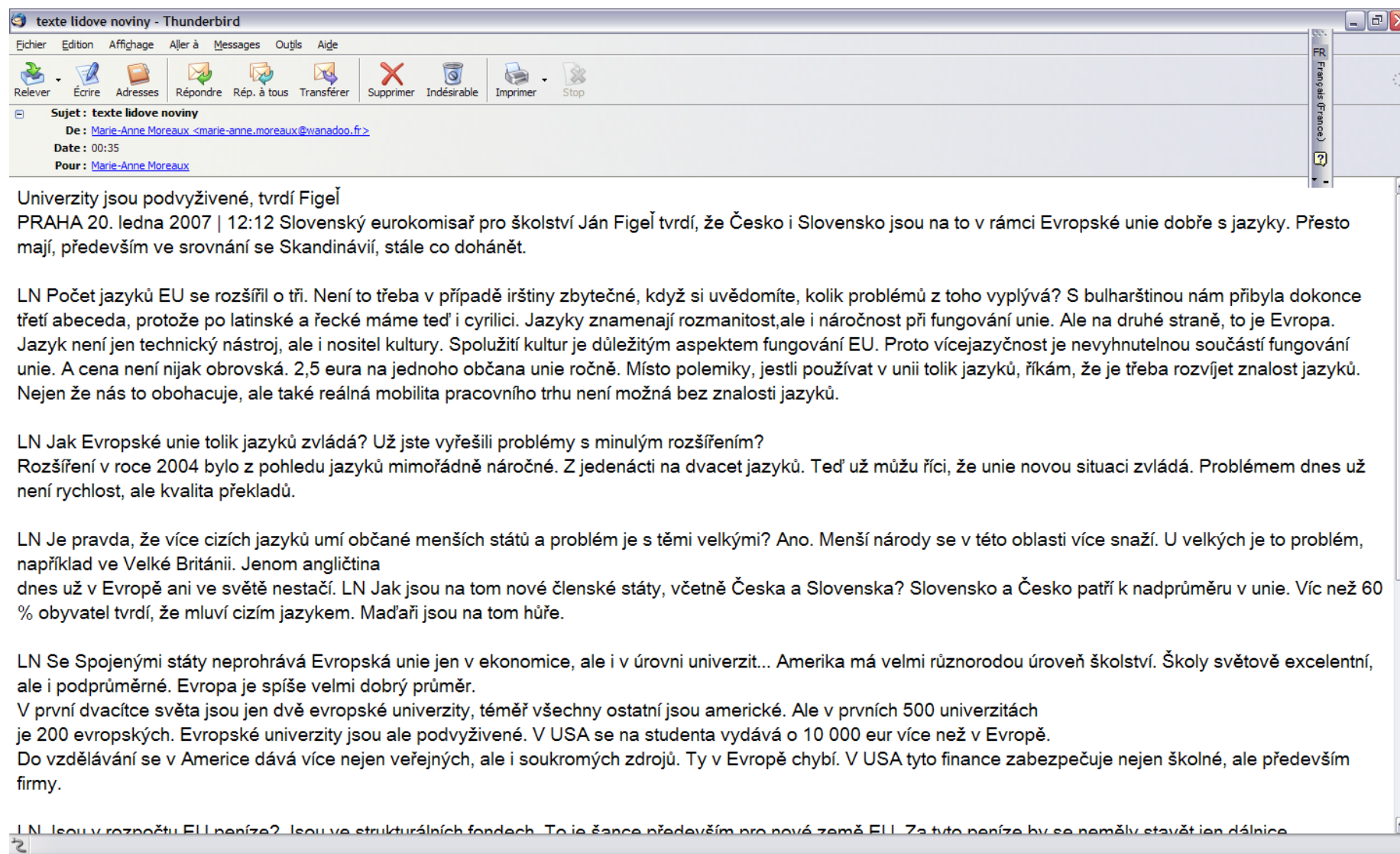


Figure 2

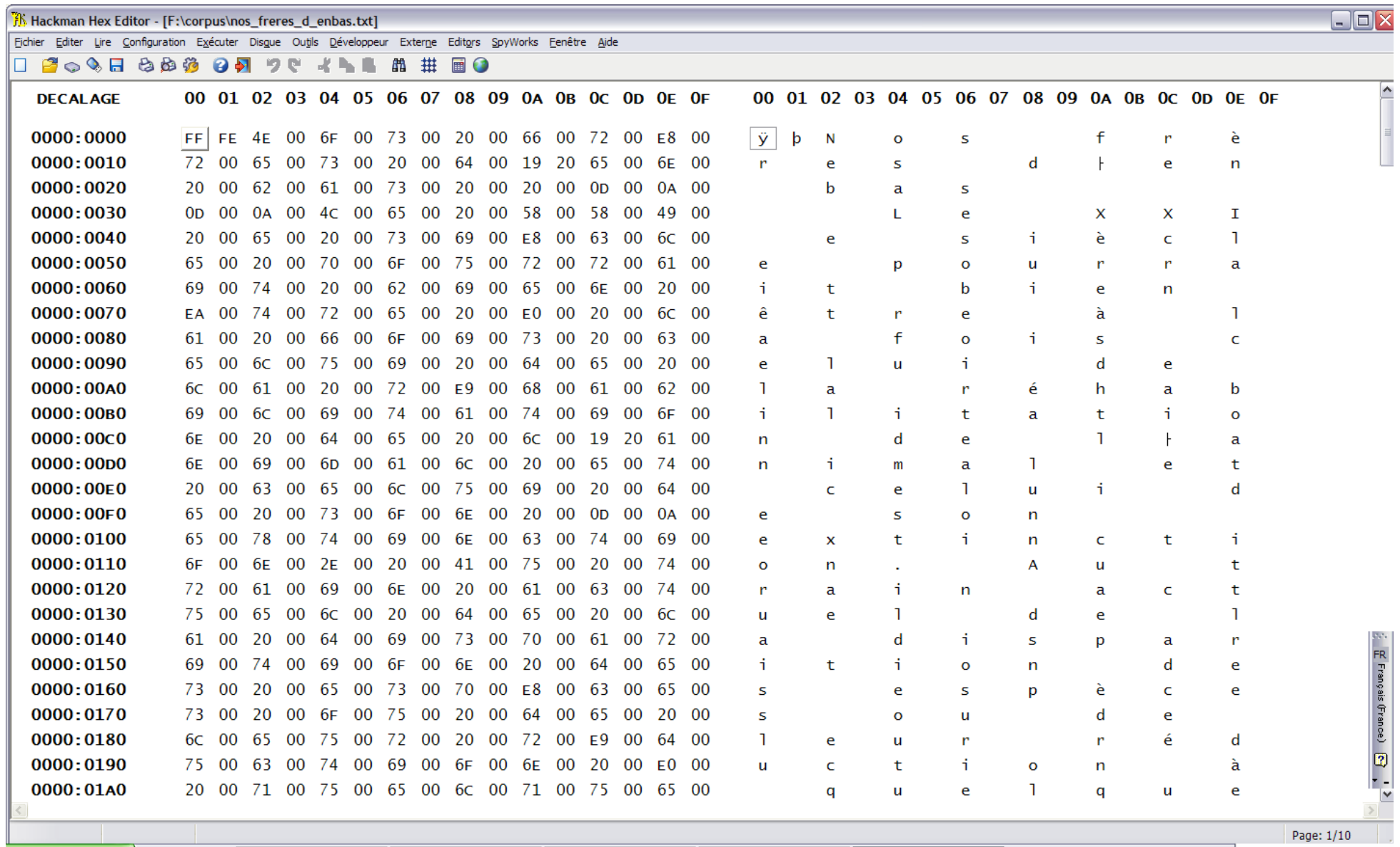


Figure 3

```

/* Ce programme compte le nombre d'occurrences du caractère < œ > figurant dans *
* un fichier texte dont le nom est passé en argument à la ligne de commande      *
*/
#include <stdio.h>

int main(int argc, char *argv[])
{
    FILE * f_in;

    if (argc != 2)
    {
        printf("Utilisation: %s nom_fichier_traité\n", argv[0]);
        return(10);
    }

    f_in = fopen(argv[1], « r »);
    if (!f_in)
    {
        printf("Erreur – impossible d'ouvrir le fichier %s\n", argv[1]);
        return(12);
    }

    char c;
    unsigned long nb_occ = 0;
    while(true)
    {
        c=fgetc(f_in);
        if ( c == EOF)
            break;
        if (c == "œ" || c == "Œ")
            nb_occ++;
    }

    printf("Le fichier %s contient %lu occurrence(s) du caractère < œ >\n", argv[1], nb_occ);
    fclose(f_in);
    return(0);
}

```

Annexe 1 – Tables 8 bits

20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
,	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	
p	q	r	s	t	u	v	w	x	y	z	{		}	~	

iso646 (ascii)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	ı	ϕ	£	⌘	¥	ı	š	..	©	≡	«	¬	-	®	-
B0	°	±	²	³	ˆ	μ	¶	.	ˆ	ı	º	»	¼	½	¾
C0	˜A	˜A	˜A	˜A	˜A	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D0	Ð	Ñ	Õ	Õ	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
F0	ð	ñ	õ	õ	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

iso8859-1 (latin 1)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	ı	ϕ	£	€	¥	Š	š	Š	©	≡	«	¬	-	®	-
B0	°	±	²	³	Ž	μ	¶	.	Ž	ı	º	»	Œ	œ	ÿ
C0	˜A	˜A	˜A	˜A	˜A	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D0	Ð	Ñ	Õ	Õ	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
F0	ð	ñ	õ	õ	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

iso8859-15 (latin 9)

80	€															
	91	92	93	94	95	96	97	98	99	9A	9B	9C		9E	9F	
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF	
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF	
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF	
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF	
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF	

Win1252

Annexe 2 – Unicode, sous-ensembles du script latin

0000

C0 Controls and Basic Latin

007F

	000	001	002	003	004	005	006	007
0	NUL 0000	DLE 0010	SP 0020	0 0030	@ 0040	P 0050	` 0060	p 0070
1	SOH 0001	DC1 0011	! 0021	1 0031	A 0041	Q 0051	a 0061	q 0071
2	STX 0002	DC2 0012	" 0022	2 0032	B 0042	R 0052	b 0062	r 0072
3	ETX 0003	DC3 0013	# 0023	3 0033	C 0043	S 0053	c 0063	s 0073
4	EOT 0004	DC4 0014	\$ 0024	4 0034	D 0044	T 0054	d 0064	t 0074
5	ENQ 0005	NAK 0015	% 0025	5 0035	E 0045	U 0055	e 0065	u 0075
6	ACK 0006	SYN 0016	& 0026	6 0036	F 0046	V 0056	f 0066	v 0076
7	BEL 0007	ETB 0017	' 0027	7 0037	G 0047	W 0057	g 0067	w 0077
8	BS 0008	CAN 0018	(0028	8 0038	H 0048	X 0058	h 0068	x 0078
9	HT 0009	EM 0019) 0029	9 0039	I 0049	Y 0059	i 0069	y 0079
A	LF 000A	SUB 001A	* 002A	: 003A	J 004A	Z 005A	j 006A	z 007A
B	VT 000B	ESC 001B	+ 002B	; 003B	K 004B	[005B	k 006B	{ 007B
C	FF 000C	FS 001C	, 002C	< 003C	L 004C	\ 005C	l 006C	 007C
D	CR 000D	GS 001D	- 002D	= 003D	M 004D] 005D	m 006D	} 007D
E	SO 000E	RS 001E	. 002E	> 003E	N 004E	^ 005E	n 006E	~ 007E
F	SI 000F	US 001F	/ 002F	? 003F	O 004F	_ 005F	o 006F	DEL 007F

	008	009	00A	00B	00C	00D	00E	00F
0	XXX 0080	DCS 0090	NB SP 00A0	◊ 00B0	À 00C0	Ð 00D0	à 00E0	ð 00F0
1	XXX 0081	PU1 0091	¡ 00A1	± 00B1	Á 00C1	Ñ 00D1	á 00E1	ñ 00F1
2	BPH 0082	PU2 0092	¢ 00A2	² 00B2	Â 00C2	Ò 00D2	â 00E2	ò 00F2
3	NBH 0083	STS 0093	£ 00A3	³ 00B3	Ã 00C3	Ó 00D3	ã 00E3	ó 00F3
4	IND 0084	CCH 0094	¤ 00A4	´ 00B4	Ä 00C4	Ô 00D4	ä 00E4	ô 00F4
5	NEL 0085	MW 0095	¥ 00A5	µ 00B5	Å 00C5	Õ 00D5	å 00E5	õ 00F5
6	SSA 0086	SPA 0096	¦ 00A6	¶ 00B6	Æ 00C6	Ö 00D6	æ 00E6	ö 00F6
7	ESA 0087	EPA 0097	§ 00A7	• 00B7	Ç 00C7	× 00D7	ç 00E7	÷ 00F7
8	HTS 0088	SOS 0098	¨ 00A8	¸ 00B8	È 00C8	Ø 00D8	è 00E8	ø 00F8
9	HTJ 0089	XXX 0099	© 00A9	¹ 00B9	É 00C9	Ù 00D9	é 00E9	ù 00F9
A	VTS 008A	SCI 009A	ª 00AA	º 00BA	Ê 00CA	Ú 00DA	ê 00EA	ú 00FA
B	PLD 008B	CSI 009B	« 00AB	» 00BB	Ë 00CB	Û 00DB	ë 00EB	û 00FB
C	PLU 008C	ST 009C	¬ 00AC	¼ 00BC	Ì 00CC	Ü 00DC	ì 00EC	ü 00FC
D	RI 008D	OSC 009D	SHY 00AD	½ 00BD	Í 00CD	Ý 00DD	í 00ED	ý 00FD
E	SS2 008E	PM 009E	® 00AE	¾ 00BE	Î 00CE	Þ 00DE	î 00EE	þ 00FE
F	SS3 008F	APC 009F	¯ 00AF	¿ 00BF	Ï 00CF	ß 00DF	ï 00EF	ÿ 00FF

	010	011	012	013	014	015	016	017
0	Ā	Ð	Ġ	İ	Ĭ	Ŏ	Š	Ů
	0100	0110	0120	0130	0140	0150	0160	0170
1	ā	đ	ġ	ı	ł	ő	š	ů
	0101	0111	0121	0131	0141	0151	0161	0171
2	Ǻ	Ē	Ģ	Ĳ	ł	Œ	Ŧ	Ū
	0102	0112	0122	0132	0142	0152	0162	0172
3	ǻ	ē	ģ	ıĵ	Ń	œ	ţ	ų
	0103	0113	0123	0133	0143	0153	0163	0173
4	Ą	Ĕ	Ĥ	Ĵ	ń	Ŕ	Ţ	Ŵ
	0104	0114	0124	0134	0144	0154	0164	0174
5	ą	ĕ	ĥ	ĵ	Ņ	ŕ	ţ	ŵ
	0105	0115	0125	0135	0145	0155	0165	0175
6	Ć	Ė	Ħ	Ķ	ņ	Ŗ	Ŧ	Ŷ
	0106	0116	0126	0136	0146	0156	0166	0176
7	ć	ė	ħ	ķ	ņ	ŗ	ţ	ŷ
	0107	0117	0127	0137	0147	0157	0167	0177
8	Ĉ	Ė	Ĩ	κ	ň	Ř	Ů	ÿ
	0108	0118	0128	0138	0148	0158	0168	0178
9	ĉ	ę	ĩ	ł	ń	ř	ů	ž
	0109	0119	0129	0139	0149	0159	0169	0179
A	Ċ	Ĕ	Ī	Í	Ņ	Ś	Ū	ž
	010A	011A	012A	013A	014A	015A	016A	017A
B	ċ	ĕ	ī	ı	ŋ	ś	ū	ž
	010B	011B	012B	013B	014B	015B	016B	017B
C	Č	Ĝ	Ĭ	ı	Ŏ	Ŝ	Ů	ž
	010C	011C	012C	013C	014C	015C	016C	017C
D	č	ĝ	ı	ł	ő	ŝ	ů	ž
	010D	011D	012D	013D	014D	015D	016D	017D
E	Ď	Ģ	ı	ł	Ŏ	Ş	Ů	ž
	010E	011E	012E	013E	014E	015E	016E	017E
F	ď	ģ	ı	ł	ő	ş	ů	ř
	010F	011F	012F	013F	014F	015F	016F	017F