INALCO
Département Russie

RUS2A 05f
Traitement informatique du russe

Informatique avec les langues
à alphabet non latin (cyrillique)

CODAGE de L'INFORMATION

CODAGES de CARACTÈRES

POLICES

NOTION DE CODAGE D'INFORMATION

Toute information, quelle que soit sa nature, traitée sur un ordinateur est nécessairement représentée par une suite finie de positions binaires, à la fois sur les supports matériels de mémoire (registres du processeur, mémoire centrales, mémoires auxiliaires) et sur les supports matériels de transmission entre les différents composants.

♦ Définitions

1. Chiffres binaires

Toute valeur binaire est constituée à partir de deux symboles notés 0 et 1 que l'on appelle les chiffres binaires.

2. Chaînes binaires

Une suite binaire est une suite finie de chiffres binaires. Le nombre de chiffres d'une suite binaire est appelé sa longueur.

3. Bit

Le mot « bit » est l'acronyme de « binary digit » qui signifie « chiffre binaire ».

C'est l'unité élémentaire permettant de formuler une valeur binaire symbolisant une information soit pour la stocker, soit la transmettre.

Sur un support de stockage (mémoire centrale ou mémoire externe), un bit est un emplacement permettant de représenter (matérialiser) un chiffre binaire.

Lors d'une transmission entre différents composants (d'un même ou de différents ordinateurs), un bit correspond à une tranche de temps d'un signal émis sur une voie de connexion.

4. Octet

Un octet est une suite ordonnée de 8 bits soit sur un support matériel, soit lors d'un transfert.

C'est l'unité de mesure élémentaire des supports d'informations : espace mémoire, taille d'un document.

Les bits d'un octet sont ordonnés : on peut p.ex. les numéroter de droite à gauche. Le bit le plus à droite porte alors le numéro 0, le plus à gauche le numéro 7. On appelle le numéro d'ordre d'un bit « poids du bit » : le bit 7 a le plus fort poids (il est dit bit de poids fort), le bit 0 le plus faible (il est dit bit de poids faible).

7	6	5	4	3	2	1	0

Les noms des unités supérieures à l'octet se construisent avec les préfixes : *kilo*, *méga* et *giga*. Ces préfixes ont cependant une signification un peu différente de la signification habituelle, puisqu'ils ne désignent pas des quantités correspondant à des puissances de 1à, mais de 2 :

1 Kilo-octet = 2^{10} octets soit 1024 octets
1 Méga-octet = 2^{20} octets soit 1024 x 1024 octets = 1 048 576 octets
1 Giga-octet = 2^{30} octets soit 1024 x 1024 x 1024 octets = 1 073 741 824 octets

BINAIRE DECIMAL HEXADECIMAL

■Bit binary digit = chiffre binaire 0 ou 1
■Chaîne binaire = 0100101111010111
■ Octet = suite ordonnée de 8 bits sur support matériel, ou lors d'un transfert.

7	6	5	4	3	2	1	0
1	1	0	1	1	0	1	0

1	**1**	**1**
10	**2^1=2**	**2**
100	**2^2 =4**	**4**
101	**2^2 +1 =5**	**5**
1000	**2^3= 8**	**8**
1010	**2^3+2=10**	**A**
1011	**2^3+2+1=11**	**B**
1111	**2^{3+} 2^2+2+1**	**F**
10000	**$2^{4=}$ 16**	**10**
11111	**31**	**1F**
100000	**$2^{5=}$ 32**	**20**
1010101	**170**	**AA**
1111111	**255**	**FF**

Systèmes de codage des caractères

♦ Définitions

1. répertoire de caractères

Un répertoire de caractères est un ensemble convenu, fini et non-ordonné de caractères que l'on considère comme étant complet pour une utilisation donnée.

On ne suppose aucune représentation pour le stockage dans la mémoire d'un ordinateur ou pour le transfert d'information. Un répertoire ne définit aucun ordre sur les caractères, p.ex. pour trier les informations ; il doit être défini séparément.

Habituellement un répertoire est défini par la spécification du nom de chaque caractère, accompagné d'une forme de présentation servant de modèle pour visualiser le caractère.

Un répertoire de caractères peut contenir des caractères qui semblent identiques par la forme de présentation, mais qui logiquement sont des caractères distincts, comme p.ex. Latin uppercase B, Cyrillic uppercase B (vé), Greek uppercase B (bêta).

2. jeux de caractères codés

Un jeu de caractères codés est une application établissant une relation entre les éléments d'un répertoire de caractères et un ensemble d'entiers positifs : on assigne donc ainsi à chaque élément du répertoire un code numérique unique, sa position de codage (code numérique, élément de code, code, *code point*).

L'ensemble des positions de codage définit un espace de codage. Un caractère associé à une position de codage est dit **caractère codé**.

Les jeux de caractères codés sont en général présentés sous la forme de tables (une ou plusieurs) que l'on appelle **tables de caractères**.

Une table de codage de caractères est donc une liste de couples : chaque élément de la table correspond à l=association d=une donnée numérique et d=un symbole permettant de coder un texte:

$Elle possède un nom afin qu'un système d'encodage et de décodage puisse y faire référence

$Elle définit un format de codage: le nombre de positions binaires utilisé pour composer les codes représentant les symboles. (Cela permet également de connaître le nombre d'éléments que la table comprend).

$Les positions de codage peuvent être considérées comme des indices.

$Elle doit être opératoire et être définie de telle sorte que l'on puisse simuler p.ex. l'ordre alphabétique ou l'ordre d'énumération des chiffres, cependant les codes numériques ne correspondent pas obligatoirement à des entiers consécutifs. Une table de codage peut comporter des « trous » : un certain nombre de position de codage réservées pour des opérations de contrôle ou devant être définies par la suite.

Ex : le caractère <a> dans la table UNICODE

U + 0061 LATIN SMALL LETTER A

UNICODE représentation en hexadécimal sur 4 positions

Un jeu de caractères est ce qui assure la traduction

- en entrée: caractère → octet
- en sortie: octet→ caractère

3. formes d'encodage (Character Encoding Form – CEF)

Une forme d'encodage de caractères est une méthode (un algorithme) permettant de représenter les caractères d'un jeu de caractères codés en transformant leur code numérique en une séquence d'octets.

- une forme d'encodage fixe utilise des séquences de même longueur pour tous les caractères d'une table
- une forme d'encodage variable utilise des séquences de longueur variable en fonction du caractère à coder : une unité ou un multiple de cette unité

Dans le cas le plus simple, chaque caractère, par référence à une table de caractères, est mis en relation avec un entier compris entre 0 et 255 et cet entier est utilisé tel quel en représentation binaire sur un format d'un octet. Cela n'est possible que dans le cas d'un répertoire restreint, comportant au maximum 256 éléments.

Dans le passé, et cela est encore fréquent, on ne faisait pas de distinction entre code caractères et forme d'encodage et on spécifiait simplement un encodage en terme de code caractères et du répertoire de caractères qu'il implique. Cela ne posait pas trop de problèmes car le répertoire était restreint et les codes numériques étaient uniquement des entiers compris entre 0 et 127 ou 0 et 255 (cf. tables 8 bits).

Une autre confusion fréquente (elle est faite, p.ex. par la plupart des navigateurs internet) : la possibilité de choisir un jeu de caractères, un code caractères ou une forme d'encodage est présentée comme la possibilité de choisir une langue.

L'organisation de données multi-octets en mémoire est déterminée par l'architecture du processeur :

- organisation grand boutiste [big endian] : l'octet de poids fort est implanté à l'adresse la plus petite, puis les octets suivants aux adresses suivantes (on remonte la mémoire de l'adresse la plus petite à la plus grande). Les processeurs Sun et Motorola (MacIntosh) fonctionnent selon cette architecture.

- organisation petit boutiste (little endian] : l'octet de poids fort est implanté à l'adresse la plus grande, puis les octets suivants aux adresses suivantes (on redescend la mémoire de l'adresse la plus grande à la plus petite). Les processeurs Intel fonctionnent selon cette architecture.

notation	octet de poids fort	octet de poids faible	petit boutiste	gros boutiste
binaire	01100101	01011111	0101111101100101	0110010101011111
hexadécimale	65	5F	5F65	655F

Ce problème se pose pour les encodages UTF-32 et UTF-16.

4. caractères

Définition Unicode :

> *The smallest component of written language that has semantic value, refers to the abstract meaning and/or shape, rather than a specific shape (see also glyph), though in code tables some form of visual representation is essential for the reader's understanding.* [1]

Le terme « caractère » désigne une notion abstraite : c'est une unité d'information qui permet d'organiser, de contrôler ou de représenter des données textuelles. Lorsqu'il permet de représenter un caractère est une classe de formes faisant référence à la même signification ou ayant une apparence similaire.

Le caractère <lettre b minuscule> (pas de signification)

b **b** *b* b b b **b** b **b** b b

b b b b

Le caractère <dollar> (une signification, c'est un signe)

$ $ $ $

5. glyphes

Les différentes représentations graphiques d'un caractère : on pourrait dire qu'un glyphe est une instance de caractère. Cependant l'utilisation du terme est souvent ambiguë

1)forme abstraite représentant une ou plusieurs formes (typo)graphiques

2)synonyme d'image (typo)graphique[2]

Il n'est pas toujours facile de faire la différence entre caractère et glyphe.

Une police est une collection de glyphes utilisée pour décrire visuellement des données caractères. Elle est associée à un ensemble de paramètres : taille, position, graisse… permettant lorsque des valeurs particulières leur sont assignés de générer une collections de formes graphiques.

Les glyphes d'une police peuvent ou non correspondre aux éléments d'une table de caractères connue (norme internationale ou norme industrielle). Sous Windows, p.ex., la plupart des polices sont associées à la page des codes CP 1252 qui est un sur-ensemble de la table ISO 8859-1.

---

[1] L'unité élémentaire du langage écrit ayant une valeur sémantique, faisant référence à la signification abstraite et/ou à la forme, et non pas une forme spécifique (cf. glyphe) bien que pour la bonne compréhension du lecteur, une forme de visualisation soit nécessaire dans une table de codage.

[2] Partie du caractère comprenant le dessin de la lettre formant relief , et qui s'imprime sur le papier (P.R)

Tableau récapitulatif des définitions

ASCII (American Standard Code for Information Interchange)	code standard américain pour l'échange d'information. Utilisé sur tous les types d'ordinateurs (PC et Mac) Codage sur 7 bits qui permet l'utilisation simultanée de 128 combinaisons différentes (31 pour les codes de fonctionnement et 97 pour les codes des caractères affichables.
Extended ASCII	ASCII étendu : Codage sur 8 bits qui permet l'utilisation simultanée de 256 combinaisons différentes (31 pour les codes de fonctionnement et 225 pour les codes des caractères affichables (permet l'utilisation de deux «alphabets»)
ANSI	Institut national américain des standards Page de code utilisée sous Widnows pour coder 256 caractères. les 128 premiers sont les mêmes que ceux du code ASCII.
bit	◊ élément de mémoire pour conserver les valeurs numérique 0 ou 1.
octet (byte)	1. ◊ plus petite unité de mémoire adressable (généralement 8 bits). Dans les systèmes 8 bits, chaque octet représente un caractère. (Dans les systèmes de codage 16 bits chaque caractère est codé sur deux octets. 2. ◊ unité de mesure de mémoire. kilooctet= 2^{10} octets=1024 ;mégaoctet=2^{20} =1.048.576 ; gigaoctet = 2^{30} = 1.073.741.824
glyphe	représentation graphique d'un caractère **D** $\mathcal{D}$ D $\mathscr{D}$ $\boldsymbol{D}$ sont différents glyphes pour le caractère ASCII N° 68
ISO	Instituteof Standard Organisation a établi des standards différents de ceux de Microsoft pour les codages des caractères ISO 8859-1= occidental ; ISO 8859-5 = Cyrillique
page de code	ensemble de caractères (ou symboles) regroupés. Chaque page de code correspond à un groupe de langue « sœurs »
fichier (file)	unité ordonnée de données, possédant un nom et une extension précisant son type. (programme, texte, image, document Word, son, fichier système etc…)
police (font)	ensemble (complet ?) de caractères regroupés dans un fichier, obéissant généralement à un principe de cohérence typographique (on a ainsi : Arial, Times, Helvetica, Courier qui définissent une typographie spécifique) et/ou linguistique, chaque police correspondant généralement à une page de code. Pour afficher correctement un texte avec des caractères différents, il faut une police correspondant à la page de code du texte.
transcodeur	logiciel qui permet le transcodage de textes bruts non formatés.
Unicode	Consortium regroupant les principales firmes informatiques qui a mis au point un codage 16 bits permettant 2^{16} soit 65536 caractères, autorisant l'affichage simultané de toutes les langues du monde (y compris arabe, chinois, coréen, japonais etc…) dans une même police et dans un même document.

Краткий словарь

ASCII (American Standard Code for Information Interchange)	американский стандартный код для обмена информацией, широко используется во многих машинах (во всех PC и Macintosh). Это семиразрядный код, к-рый обеспечивает 128 различных битовых комбинаций, включая 31 управляющих комбинаций.
Extended ASCII	расширенный код ASCII: это восьмиразрядный код, к-рый обеспечивает 256 различных битовых комбинаций, включая 31 управляющих комбинаций.
ANSI (АНИС)	Американский национальный институт стандартов, который разрабатывал набор символов ANSI. Это кодовая страница, используемая в Windows для представления 256 символов. Первая половина этого набора совпадает с с кодом ASCII
бит (bit)	1. ◊ элемент памяти для хранения цифрового значения 0 или 1 ; 2. один разряд ◊ Binary Information transfert = передача двоичных данных
байт (byte)	1. ◊ наименьшая адресуемая единица памяти (обычно содержит 8 двоичных разрядов) ◊ последовательность битов. Обычно используется восьмибитовый байт; в восьмибитовой системе каждый байт соответсвует одному символу, знаку, букву, знаку препинания… 2.единица измерения памяти. килобайт= 2^{10} байт=1024 ;мегабайт=2^{20} =1048576
глиф (glyph)	образ для символа (знака) в битовой карте отображения информации
кодовая страница	набор символов. Каждая кодовая страница имеет собственный номер и определяет кодировку для отдельного национального языка или группы "родственных языков"
перекодировщик	ПО для перекодировки простых неоформленных текстов
файл (file)	упорядоченный набор записей, имеющий имя и расширение, указывающее на тип файла (программа, текст, картинка, документ Word, звуковой файл, системный файл)
шрифт (font)	(полный) набор символов заданного начертания. Семейство шрифтов образует гарнитуру: шрифты объединены общим дизайном (Arial, Times, Helvetica, Courier) . В принципе шрифт соответствует одной кодовой странице.

♦ Jeux de caractères codés

1. normes internationales

Nom de la table	Format de codage	Forme d'encodage
ISO 646 IRV	7 bits	Iso 646 (1 octet)
ISO8859-n pour n = [1;16] donc 16 tables	8 bits	Iso 8859 (1 octet)
ISO 10646 (UCS)	32 bits	UCS-4 (4 octets) UCS-2 (2 octets) UTF-16(2 octets ou 2 x 2 octets) UTF-8 (1 à 6 octets) UTF-7 (1 à 4 octets)

2. normes industrielles

Nom de la table	Format de codage	Forme d'encodage
EBCDIC (IBM)	8 bits	1 octet
Pages de codes de DOS 437, 850... (Microsoft)	8 bits	1 octet
Page de codes Windows 1250, 1251, 1252 (Microsoft) [Windows 1252 dite ANSI]	8 bits	1 octet
UNICODE (Consortium Unicode) Versions 1.*x* à 3.*x* Version 4.*x* à 5	 16 bits 20 bits	UCS-2 (2 octets) UTF-16(2 octets ou 2 x 2 octets) UTF-8 (1 à 6 octets) UTF-7 (1 à 4 octets

Unicode

Le but d'UNICODE est de pouvoir fournir un codage non-ambigu sur 16 bits jusqu'à la version 3.2, sur 20 bits depuis la version 4, qui n'a pas besoin de séquences de contrôle. Il permet l'échange, le traitement et la visualisation des caractères utilisés par la plupart des langues vivantes: scripts latin, grec, cyrillic, arménien, hébreu, arabe, devanagari, bengali, gurmukhi, gujarati, oriya, tamul, télugu, kannada, malaysien, siamois, lao, géorgien, tibétain, kana, hangul, CJK (ensemble unifié des caractères idéographiques chinois, japonais, coréens).

Unicode définit un caractère (élément de codage d'un texte) en terme de 1 code + un nom mais ne définit aucun glyphe, c'est le dispositif qui utilise la table qui doit prendre en charge l'apparence du caractère. Actuellement, la table comprend environ 96 447 caractères (associations code-nom).

Les caractères sont regroupés en «scripts» dans des bloques de codes. Un script est un système de caractères ayant des propriétés communes. S'il y a un ordre habituel sur ces caractères, p.ex. ordre alphabétique, Unicode ordonne les caractère de telle sorte que cet ordre soit maintenu.

Le projet UNICODE ne se contente pas de référencer, d'organiser et de classer les différents symboles des écritures. Il cherche à rationaliser leur utilisation et à établir des règles concernant leur manipulation. Il donne des recommandations et définit :

- les caractères combinés : symboles complexes formés à partir de plusieurs symboles. UNICODE recense ces combinaisons et autorise leur définition par concaténation des caractères élémentaires,

voire comme caractère unique à des fins de compatibilité avec les standard antérieur (c'est le cas des lettres diacritées du français). [3]

·la normalisation des caractères afin d'établir des correspondances entre caractères de code points différents mais ayant la même interprétation ou la même fonction, entre caractères de casses (minuscule, majuscule et tittle-case) différentes pour rationaliser les conversions (p.ex. latin → cyrillic) et faciliter les comparaison et les tris.

·l'encodage des caractères (cf. tableau p. précédente)

ISO 10646

Le standard international ISO 10646 définit le jeu de caractères international, Universal Character Set (UCS). Ce jeu de caractère est un super-ensemble de tous les autres jeux standard. Il garantit une compatibilité réversible avec tous les autres jeux: il n'y a aucune perte d'information si un texte est converti en UCS puis reconverti dans code d'origine

Il définit un jeu de caractères codés sur 31 bits.

Le sous-ensemble sur 16 bits de UCS s'appelle le BMP (Basic Multilingual Plan). La norme le définissant à été publiée en 1993 sous le nom de ISO 10646-1.

UCS assigne à chaque caractère un code et un nom. Le code est un nombre en représentation hexadécimale. On a l'habitude lorsque l'on donne un code UCS (et Unicode) de le faire précéder de la lettre. Le nom est un nom standardisé.

ex. U+0041 Latin capital letter A

Les caractères de U+0000 à U+007F sont identiques au jeu ASCII; de U+0000 à U+00FF à ISO 8859-1.

UNICODE et ISO-10646 se développent actuellement conjointement.

LE BMP (BASIC MULTILINGUAL PLAN)

Sa structure est la suivante: il est réparti en 4 zones

- zone A: alphabets arabe, arménien, cyrillique, grec, hangul, hébreu, indiens, kana, tha , ..., symboles diacritiques, symboles divers, éléments graphiques...

  Les 256 premiers caractères correspondent aux caractères définis par ISO 8859-1 (ISO LATIN 1).
- zone I: idéogrammes (caractères chinois unifiés)

  Elle comporte environ 21000 caractères chinois unifiés de Chine, Corée et Japon. Ils ont été choisis dans les jeux de caractères définis par les normes GB2312 pour la Chine, Big-5 pour Taïwan, Jis X 0208 et Jis X 0212 pour le Japon.
- zone O: ouverte (réservée pour extension, mais une partie est utilisée pour les hangul sous forme complète)
- zone R: réservée (pour usage privée et pour permettre les conversions de code).

[3] Cela implique qu'avant toute comparaison deux chaînes constituées de caractères encodés selon le standard UNICODE doivent être normalisées (ramenées au même format).

♦ Dump de fichier texte

On peut ainsi examiner les octets du fichiers et voir les codes des caractères, y compris les caractères de contrôle

Notion de codage
Codage des caractères.
Le jeu de caractères de la machine est le jeu ASCII étendu
(DOS: version anglais).
chiffres: 0 1 2 3 4 5 6 7 8 9
minuscules: abcdefg...
majuscules: ABCDEFG...
caractères de contrôles:^G, ^K

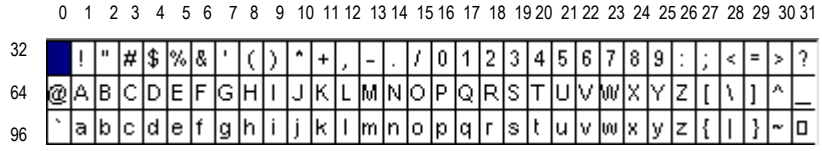
```
17B4:0100  20 20 20 20 20 4E 6F 74_69 6F 6E 20 64 65 20 63        Notion de c
17B4:0110  6F 64 61 67 65 2E 0D 0A_09 43 6F 64 61 67 65 20     odage....Codage
17B4:0120  64 65 73 20 63 61 72 61_63 74 8A 72 65 73 2E 0D     des caract.res..
17B4:0130  0A 4C 65 20 6A 65 75 20_64 65 20 63 61 72 61 63   . Le jeu de carac
17B4:0140  74 8A 72 65 73 20 64 65_20 6C 61 20 6D 61 63 68     t.res de la mach
17B4:0150  69 6E 65 20 65 73 74 20_6C 65 20 6A 65 75 20 41     ine est le jeu A
17B4:0160  53 43 49 49 20 82 74 65_6E 64 75 0D 0A 28 44 4F     SCII .tendu..                ...
```

Quelques références

http://www.unicode.org
http://hapax.iquebec.com/
http://www.alanwood.net/unicode

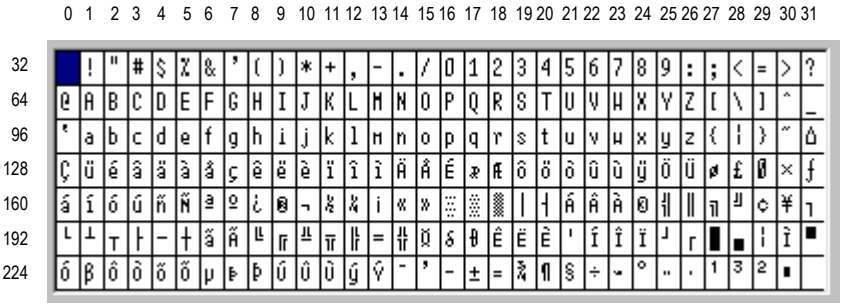
Table des caractères

♦ **Table *1a* ASCII 7 bits** anglais
96 caractères: codes 32-127



♦ **Table *1b* KOI -7 russe**

♦ **Table *2a* : ASCII 8 bits : Page de codesDOS № 850 (Latin I)**
224 caractères : codes 32-255 anglais +langues occidentales

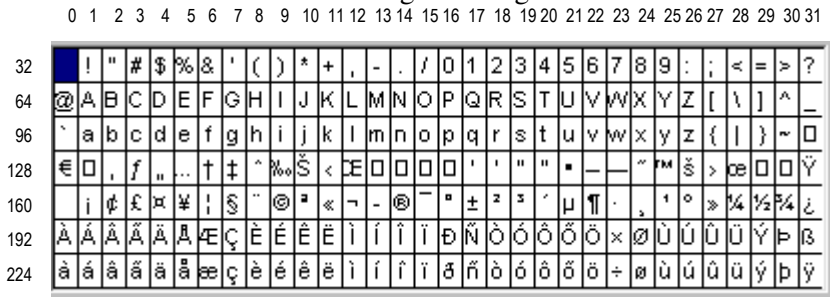


♦ **Table *2b* : Page de codesDOS ALT № 866 (Russe)**
224 caractères : codes 32-255 anglais + russe

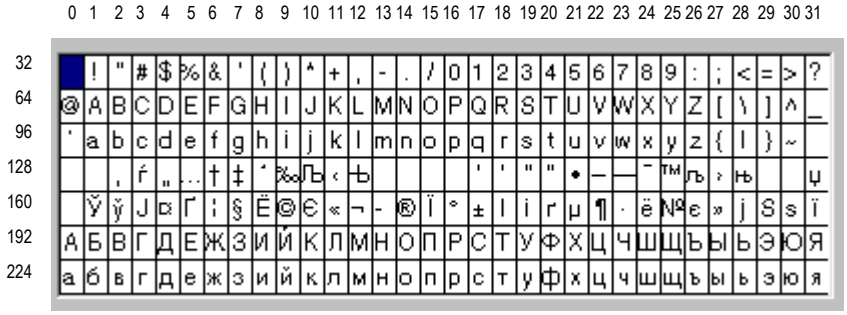
♦ **Il existe également la page de codesDOS № 852 (Slave latin)**
224 caractères : codes 32-255 anglais + langues slaves à alphabet latin
Les caractères accentués propres au langues romanes sont remplacés par les caractères des langues slaves: **ć č ě ł ń ř ś š ť ů ź ż ž** etc**.**

tables ANSI

- **Table *3a* : WINDOWS 1252 Occidental:**
  224 caractères : codes 32-255 anglais +langues occidentales

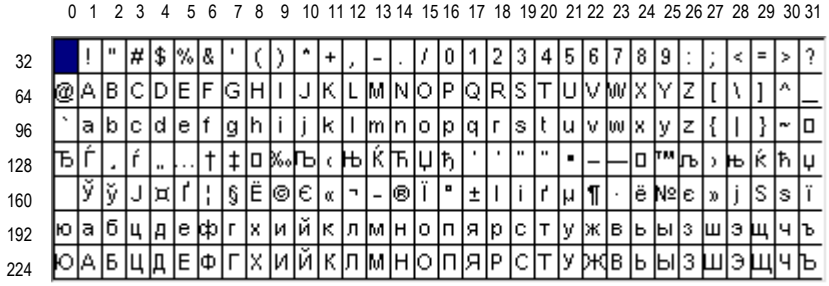


- **Table *3b* : WINDOWS 1251 Cyrillique**
  224 caractères : codes 32-255 anglais + cyrillique (russe+..)



- **Table 3c : KOI-8 U**

anglais + cyrillique (russe + ukrainien, biélorrusse, bulgare, macédonien, serbe)



- **Autres tables: Windows Europe centrale**

anglais + slave latin(croate, slovène, tchèque, polonais) + roumain, hongrois …
Les codes entre 128 et 255 sont remplacés par les lettres propres à ces langues:

Ć Č Ě Ł Ň Ń Ř Ś Š Ť Ű Ů Ź Ż Ž
ć č ě ł ň ń ř ś š ť ű ů ź ż ž

- **Autres tables: Windows Turkish, Windows Baltic , Widnows Arabic, Windows Hebrew** etc**.**

Table du codage Unicode pour Windows 95/98/2000/XP/NT et les systèmes Unicodés

LATIN

	000	001	002	003	004	005	006	007
0	NUL 0000	DLE 0010	SP 0020	0 0030	@ 0040	P 0050	` 0060	p 0070
1	STX 0001	DC1 0011	! 0021	1 0031	A 0041	Q 0051	a 0061	q 0071
2	SOT 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

LATIN OCCIDENTAL ETENDU

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

LATIN EUROPE CENTRALE

	010	011	012	013	014	015	016	017
0	Ā 0100	Đ 0110	Ġ 0120	İ 0130	ŀ 0140	Ő 0150	Š 0160	Ű 0170
1	ā 0101	đ 0111	ġ 0121	ı 0131	Ł 0141	ő 0151	š 0161	ű 0171
2	Ă 0102	Ē 0112	Ģ 0122	IJ 0132	ł 0142	Œ 0152	Ţ 0162	Ų 0172
3	ă 0103	ē 0113	ģ 0123	ij 0133	Ń 0143	œ 0153	ţ 0163	ų 0173
4	Ą 0104	Ĕ 0114	Ĥ 0124	Ĵ 0134	ń 0144	Ŕ 0154	Ť 0164	Ŵ 0174
5	ą 0105	ĕ 0115	ĥ 0125	ĵ 0135	Ņ 0145	ŕ 0155	ť 0165	ŵ 0175
6	Ć 0106	Ė 0116	Ħ 0126	Ķ 0136	ņ 0146	Ŗ 0156	Ŧ 0166	Ŷ 0176
7	ć 0107	ė 0117	ħ 0127	ķ 0137	Ň 0147	ŗ 0157	ŧ 0167	ŷ 0177
8	Ĉ 0108	Ę 0118	Ĩ 0128	ĸ 0138	ň 0148	Ř 0158	Ũ 0168	Ÿ 0178
9	ĉ 0109	ę 0119	ĩ 0129	Ĺ 0139	ŉ 0149	ř 0159	ũ 0169	Ź 0179
A	Ċ 010A	Ě 011A	Ī 012A	Í 013A	Ŋ 014A	Ś 015A	Ū 016A	ź 017A
B	ċ 010B	ě 011B	ī 012B	Ļ 013B	ŋ 014B	ś 015B	ū 016B	Ż 017B
C	Č 010C	Ĝ 011C	Ĭ 012C	ļ 013C	Ō 014C	Ŝ 015C	Ŭ 016C	ż 017C
D	č 010D	ĝ 011D	ĭ 012D	Ľ 013D	ō 014D	ŝ 015D	ŭ 016D	Ž 017D
E	Ď 010E	Ğ 011E	Į 012E	ľ 013E	Ŏ 014E	Ş 015E	Ů 016E	ž 017E
F	ď 010F	ğ 011F	į 012F	Ŀ 013F	ŏ 014F	ş 015F	ů 016F	ſ 017F

CYRILLIQUE

	040	041	042	043	044	045	046	047
0		А 0410	Р 0420	а 0430	р 0440		Ѡ 0460	Ψ 0470
1	Ё 0401	Б 0411	С 0421	б 0431	с 0441	ё 0451	ѡ 0461	ψ 0471
2	Ђ 0402	В 0412	Т 0422	в 0432	т 0442	ђ 0452	Ѣ 0462	Ѳ 0472
3	Ѓ 0403	Г 0413	У 0423	г 0433	у 0443	ѓ 0453	ѣ 0463	ѳ 0473
4	Є 0404	Д 0414	Ф 0424	д 0434	ф 0444	є 0454	Ѥ 0464	Ѵ 0474
5	S 0405	Е 0415	Х 0425	е 0435	х 0445	s 0455	ѥ 0465	ѵ 0475
6	І 0406	Ж 0416	Ц 0426	ж 0436	ц 0446	і 0456	Ѧ 0466	Ѷ 0476
7	Ї 0407	З 0417	Ч 0427	з 0437	ч 0447	ї 0457	ѧ 0467	ѷ 0477
8	J 0408	И 0418	Ш 0428	и 0438	ш 0448	j 0458	Ѩ 0468	Ѹ 0478
9	Љ 0409	Й 0419	Щ 0429	й 0439	щ 0449	љ 0459	ѩ 0469	ѹ 0479
A	Њ 040A	К 041A	Ъ 042A	к 043A	ъ 044A	њ 045A	Ѫ 046A	Ѻ 047A
B	Ћ 040B	Л 041B	Ы 042B	л 043B	ы 044B	ћ 045B	ѫ 046B	ѻ 047B
C	Ќ 040C	М 041C	Ь 042C	м 043C	ь 044C	ќ 045C	Ѭ 046C	Ѽ 047C
D		Н 041D	Э 042D	н 043D	э 044D		ѭ 046D	ѽ 047D
E	Ў 040E	О 041E	Ю 042E	о 043E	ю 044E	ў 045E	Ѯ 046E	Ѿ 047E
F	Џ 040F	П 041F	Я 042F	п 043F	я 044F	џ 045F	ѯ 046F	ѿ 047F

Ambiguïtés

é → 7B = 123 *ASCII fr*
é → 82 = 130 *Dos 437, Dos 850,*
é → E9 = 233 *Win 1250, 1252*

123 → **é** *ASCII fr*
123 → **{** *Win 1252*
233 → **é** *Win 1252*
233 → **й** *Win 1251*

Principaux problèmes d'affichage du cyrillique

« Я говорю по-русски »

tapé dans différents codages (source) et vu avec un codage impropre (visualisation).

	codage source	codage visualisation	apparence du texte	solution
1	**КОИ 7**	**- ASCII**	q GOVOR@ PO RUSSKI	*transcodeur*
2	**DOS 866** **DOS 866** **DOS 866**	**- DOS 850** **- Win 1252** **- Win 1251**	ł NYЛY□□ ЗY□□□□□ʊR Ÿ £®¢®àî ¯®-àãáá¨ ц J®ў®ao Ï®-агббЄЁ	*cyrillisateur DOS* *Shtirlitz+police 8 bits* *transcodeur*
3	**Win 1251** **Win 1251**	**- Win1252** **- Unicode**	ß ãîâîðþ ïî-ðóññêè ß ãîâîðþ ïî-ðóññêè	*police cyrillique 8 bits (Arial Cyr)* *transcodeurs + police 16 bits (Arial)*
4	**KOI 8** **КОИ 8**	**- Win 1252** **- Win 1251**	ñ ÇÏ×ÏÒÀ ÐÏ-ÒÕÓÓËÉ с ЗПЧПТА РП-ТХУУЛЙ	*police KOI 8 et/ou transcodeur*
5	**Mac Cyrillic -** **Mac Cyrillic -**	**Win 1251** **Win 1252**	ц говорю по-русски Ÿ ãîâîðþ ïî-ðóññêè	*transcodeur* *transcodeur+ police*
6a	**Chiwriter - ANSI**	**ASCII** /	W sjdjh : gj-heccrb	*macro complexe ou police spécifique*
6b	**Mac "Тула" -**	**ANSI**	® govor¼ po-russki	*macro complexe ou police spécifique*
7a	**Unicode**	**– RTF**	{\f5\fswiss\fcharset204{*\fname Arial ;}Arial Cyr ;}{\f6\fswiss Arial ;}{\f7\fswiss\fprq2 System ;}}{\colortbl\red0\green0 \blue0 ;}\deflang1036\pard\plain \lang1049\f6\fs22 \'df \'e3\'ee \'e2 \'ee\ 'f0\'fe \'ef\'ee-\'f0\'f3 \'f1\'f1\'ea\'e8\plain\f6\fs22 \par }	*Word ou rien à faire*
7b	**Unicode**	**- ANSI**	? ? ? ? ? ? ? ? ?– ? ? ? ? ? ?	*texte définitivement perdu*
7c	**Unicode**	**- ANSI**	_ _____ __-_____	*système russe, sinon fichier inutilisable*
7d	**Unicode**	**- ANSI**	Я говорю по-русски	*police 8 bits*

7e	**UNICODE - source text**	Ð¯ Ð³Ð¾Ð²Ð¾Ñ€ÑŽ Ð¿Ð¾-Ñ€ÑƒÑ Ñ Ð°Ð	*Internet explorer*
7d	**UNICODE - text HTML Word**	&#1071 ; &#1075 ;&#1086 ; &#1074 ;&#1086 ;&#1088 ;&#1102 ; &#1087 ;&#1086 ;-&#1088 ;&#1091 ;&#1089 ;&#1089 ;&#1082 ;&#1080 ;	*Internet Explorer ou macro Word*
8	**menu de programme en russe sans installation de fontes systèmes cyrilliques**	Ôàéë Ïðàâêà Ïåðåêîäèðîâàòü Ïàðàìåòðû Ñîçäàòü Îòêðûòü ... Ñîõðàíèòü ... Ñîõðàíèòü êàê ... Ïå÷àòü Íàñòðîéêà ïðèíòåðà ... Âûõîä	*système russe ou configuration écran avec polices 8 bits cyr sous Windows 95-98*
9	**menu d'un programme en français lors d'une russification totale de l'ordinateur**	✔ Barre d'outils Rиgle ✔ Barre d'иtat Unitиs de mesure ▸ Fenкtre ? Nouvelle fenкtre Rиorganiser tout Fractionner	*revenir en système français*

Les polices de caractères (fontes, alphabet)

Windows fonctionne principalement avec les polices de caractères True Type (.TTF). Elles sont installées dans le répertoire Windows\Fonts.

Polices Unicode et Windows 1251 (www.paratype.ru)

EXEMPLES DE POLICES (CARACTERES & TAILLES DIFFERENTES):

Police (Fontes) Latines:

Times New Roman 10: Institut National des Langues et Civilisations Orientales

Times New Roman 12: Institut National des Langues et Civilisations Orientales

Arial 10 : Institut National des Langues et Civilisations Orientales

Courrier New 8 : Institut National des Langues et Civilisations Orientales

Impact 12 : Institut National des Langues et Civilisations Orientales

Police (Fontes) Cyrilliques:

Times New Roman Cyr 10 : Санкт-Петербургский Государственный Университет

Times new Roman Cyr 12 (курсив): Санкт-Петербургский Государственный Университет

Aria Cyr 10: Санкт-Петербургский Государственный Университет

Courier Cyr 8: Санкт-Петербургский Государственный Университет

Courier Cyr 10 (курсив): Санкт-Петербургский Государственный Университет

Courier Cyr 12: Санкт-Петербургский Государственный Университет

Сказка 16: Санкт-Петербурский Государственный Университет

Cyrilliser son ordinateur PC sous Windows XP

Vous devez posséder le CD-ROM de Windows 2000 ou XP
Vérifiez votre version de Windows : clic sur Démarrer, aller sur Paramètres; clic sur Panneau de configuration; clic sur Système . La version de Windows doit être indiquée clairement Windows 2000 ou XP

Activation du cyrillique sur l'ordinateur Windows 2000

- clic sur **Démarrer**
  clic sur **Paramètres**
  clic sur **Panneau de configuration**
  clic sur **Options régionales**
  clic sur l'onglet **Paramètres régionaux d'entrée**
  clic sur **Ajouter**
  clic sur **Français** qui apparaît sélectionné (surligné en bleu) faire défiler l'ascenseur vers la bas et sélectionner **Russe** par un clic
  vérifier que vous avez clic Russe en paramètres régionaux d'entrée et en configuration clavier
  clic sur **OK**
  vérifier la combinaison des touches qui permet de passer d'une langue à l'autre (en principe Alt Maj ou Control Maj
  cocher la case activer l'indicateur sur la barre des tâches
  clic sur **OK**

Activation du cyrillique sur l'ordinateur Windows XP

- clic sur **Démarrer**
  clic sur **Paramètres**
  clic sur **Panneau de configuration**
  clic sur **Options régionales et linguistiques**
  clic sur l'onglet **Langues** puis **Détails**
  clic sur **Ajouter**
  clic sur **Français** qui apparaît sélectionné (surligné en bleu) faire défiler l'ascenseur vers la bas et sélectionner **Russe** par un clic
  vérifier que vous avez clic Russe en paramètres régionaux d'entrée et en configuration clavier
  clic sur **OK**
  vérifier la combinaison des touches qui permet de passer d'une langue à l'autre (en principe Alt Maj ou Control Maj
  cocher la case activer l'indicateur sur la barre des tâches
  clic sur **OK**

Si le message" Insérez le disque d'installation de Windows" apparaît, cela signifie que Windows n'a pas été complètement installé sur votre disque dur et il faut mettre le CD-ROM original d'installation de Windows
attendre que Windows ait fini de copier les fichiers nécessaires
Si le message " Vous devez redémarrer votre ordinateur pour que les changement prennent effet" s'affiche, vous devez éteindre et relancer l'ordinateur.

Saisir le cyrillique

Installation du clavier russe / serbe, bulgare etc. (mise au point de clavier spécifique translittéré)
Utilisation du clavier visuel

Transcodage

Principe du transcodage

Transcoder, c' est changer le code des lettres, de manière à avoir le texte russe dans le codage que l'on possède.
Le transcodage se fait soit automatiquement (copier coller entre deux applications MS IE6 → Word, Txt → DOC), soit manuellement (IE6, Outlook Express, Word), par macro ou programme spécifique (Shtirlitz).
C'est une opération assez simple, mais très fastidieuse.

Paramétrage de base de Word (confirmation des conversion lors de l'ouverture)

Quelques exemples du même texte visualisé dans différents codages:
On voit tout de suite qu'il est très difficile de garder simultanément le russe et les lettres accentuées françaises.

Texte saisi en standard russe 866 (pas de conversion: module texte seulement)
Ce texte est le d,but de la nouvelle de Tch,kov " La dame au petit chien ":
ƒ®¢®à¨«¨, çâ® ¡¥à¥¦®© ¯®ï¢¨«®áì ®¢®¥ «¨æ®: ¤ ¬ á á®¡ ç°®©.

Texte saisi sous Windows 1251 (mais on ne possède pas les polices russes)
Ce texte est le début de la nouvelle de Tchékhov "La dame au petit chien"
Ãîâîðèëè, ÷òî íà íàáåæíé ïîÿâèëîñü íîâîå ëèöî: äàìà ñ ñîáà÷êîé

Accents, lettres diacritées

Problème de la notation de l'accent en russe

Codage spécifique de caractère non prévu par Unicode : Macro, Police spécifique

Russification de l'ordinateur
Programme non Unicode

Problème du fonctionnement des logiciels « made in Russia »

Codage spécifique de caractère non prévu par Unicode : Macro, Police spécifique

Paramétrage de l'ordinateur pour un fonctionnment optimal 'bilingue'