Кодирование текстовой информации

保

Защита
(входная дверь,
детская комната)

春

Весна
(предприятия)

美

Красота
(ванная комната)

道

Путь
(решения)

光

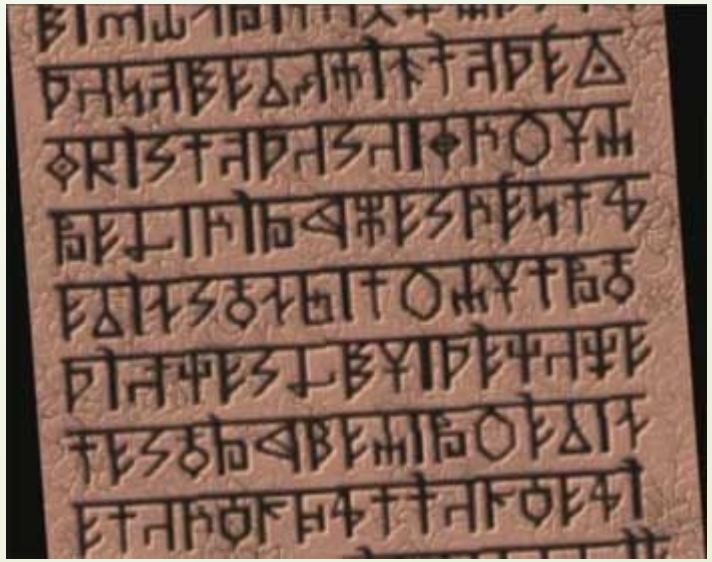
Свет
(депрессии)

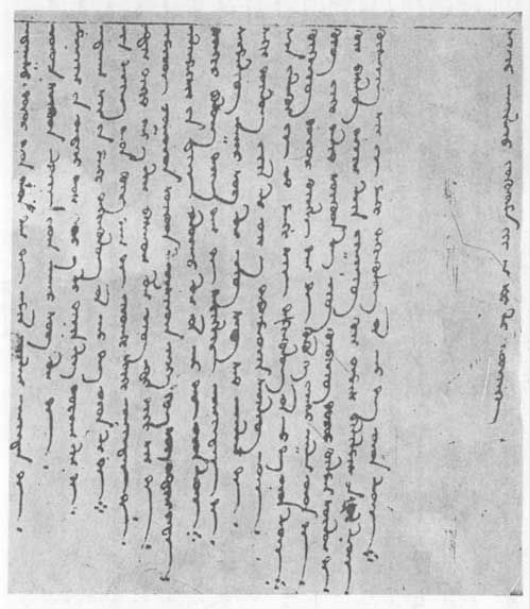
財

Благосостояние
(кошелек, банков-
ские документы)

Ко ов
(по ей
не а,
и

Письменность – способ кодирования устной речи на естественном языке





СПОСОБ КОДИРОВАНИЯ ЗАВИСИТ ОТ НАЗНАЧЕНИЯ КОДА

ПРАВИЛО – КАЖДЫЙ СИМВОЛ АЛФАВИТА ИСХОДНОГО ТЕКСТА ЗАМЕНЯЕТСЯ НА КОМБИНАЦИЮ СИМВОЛОВ АЛФАВИТА КОДИРОВАНИЯ

A

R

00	01	02	03	04	05	06	07
NUL	E 3	LF	A -	SP	S '	I 8	U 7
08	09	0A	0B	0C	0D	0E	0F
CR	D ENQ	R 4	J BEL	N ,	F !	C :	K (
10	11	12	13	14	15	16	17
T 5	Z +	L)	W 2	H £	Y 6	P 0	Q 1
18	19	1A	1B	1C	1D	1E	1F
O 9	B ?	G &	FIGS	M .	X /	V ;	LTRS
Letters		Figures			Control Chars.		

Буквы знаки символы

Режим ввода букв

Телеграфный код ITA2

Кодировка ASCII
(American Standard Code for Information Interchang) – 1963 год – для компьютерной обработки текстовой информации

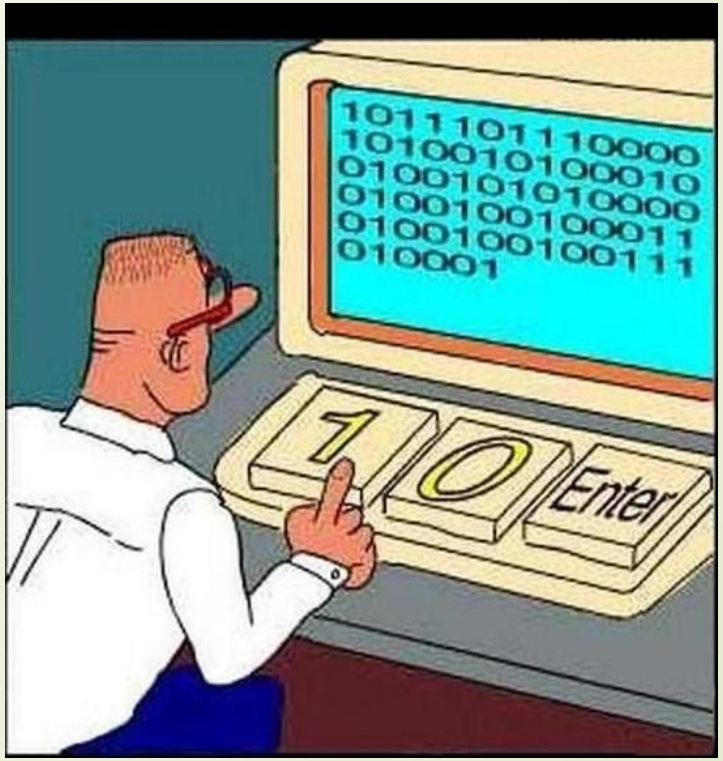
кодирующая первую половину символов с числовыми кодами от 0 до 127

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
0.	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	TAB	LF	VT	FF	CR	SO	SI
1.	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2.		!	"	#	$	%	&	'	(	)	*	+	,	–	.	/
3.	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4.	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5.	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6.	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7.	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

KtoNaNovenkogo.ru

(коды от 0 до 32 отведены не символам, а функциональным клавишам).

- Код символа – порядковый номер
- Первые 32 символа – управляющие. На экране не отражаются, определяют некоторое действие.



Расширение кода ASCII

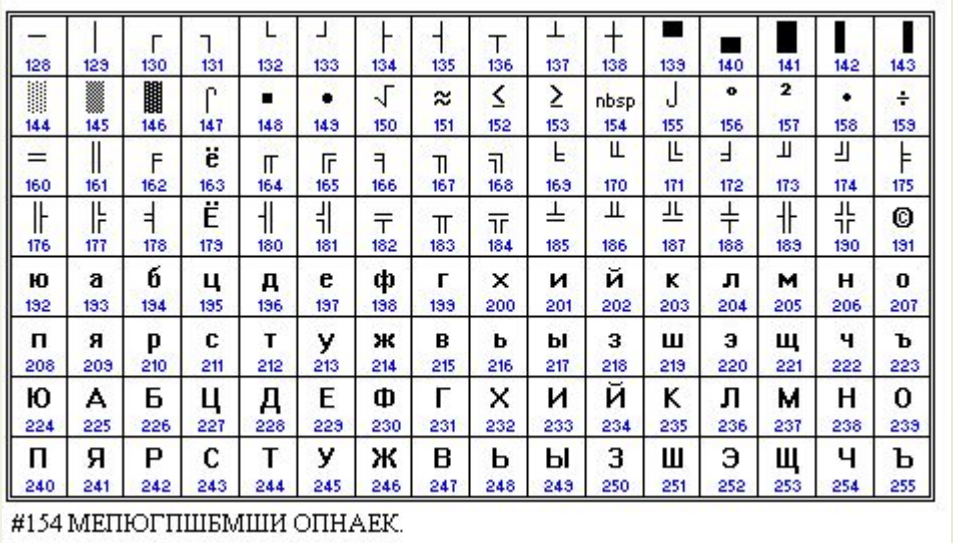
	80	90	A0	B0	C0	D0	E0	F0
0	А	Р	а	▒	└	╨	р	≡
1	Б	С	б	▙	┴	╤	с	±
2	В	Т	в	▛	┬	╥	т	≥
3	Г	У	г	│	├	╙	у	≤
4	Д	Ф	д	┤	─	╘	ф	⌠
5	Е	Х	е	╡	┼	╒	х	⌡
6	Ж	Ц	ж	╢	╞	╓	ц	÷
7	З	Ч	з	╖	╟	╫	ч	≈
8	И	Ш	и	╕	╚	╪	ш	°
9	Й	Щ	й	╣	╔	┘	щ	∙
A	К	Ъ	к	║	╩	┌	ъ	·
B	Л	Ы	л	╗	╦	█	ы	√
C	М	Ь	м	╝	╠	▄	ь	ⁿ
D	Н	Э	н	╜	═	▌	э	²
E	О	Ю	о	╛	╬	▐	ю	■
F	П	Я	п	┐	╧	▀	я	

1 – 127 совпадают с ASCII
128 – 225 – кодовая страница.
Размещаются нелатинские алфавиты, символы псевдографики…

Наиболее распространенной в настоящее время является кодировка Microsoft Windows, обозначаемая сокращением CP1251 ("CP" означает "Code Page", "кодовая

Ѓ 128	ѓ 129	‚ 130	ѐ 131	„ 132	… 133	† 134	‡ 135	€ 136	‰ 137	Ѐ 138	‹ 139	ѝ 140	Ѝ 141	Ѓ 142	Ѓ 143
ѓ 144	‘ 145	’ 146	“ 147	” 148	• 149	– 150	— 151	ё 152	™ 153	ѐ 154	› 155	ѝ 156	ѝ 157	ѓ 158	ѓ 159
nbsp 160	ў 161	Ы 162	Э 163	¤ 164	ы 165	¦ 166	§ 167	Ё 168	© 169	Ю 170	« 171	¬ 172	shy 173	® 174	Я 175
° 176	± 177	ы 178	э 179	´ 180	µ 181	¶ 182	· 183	ё 184	№ 185	ю 186	» 187	э 188	ю 189	я 190	я 191
А 192	Б 193	В 194	Г 195	Д 196	Е 197	Ж 198	З 199	И 200	Й 201	К 202	Л 203	М 204	Н 205	О 206	П 207
Р 208	С 209	Т 210	У 211	Ф 212	Х 213	Ц 214	Ч 215	Ш 216	Щ 217	Ъ 218	Ы 219	Ь 220	Э 221	Ю 222	Я 223
а 224	б 225	в 226	г 227	д 228	е 229	ж 230	з 231	и 232	й 233	к 234	л 235	м 236	н 237	о 238	п 239
р 240	с 241	т 242	у 243	ф 244	х 245	ц 246	ч 247	ш 248	щ 249	ъ 250	ы 251	ь 252	э 253	ю 254	я 255

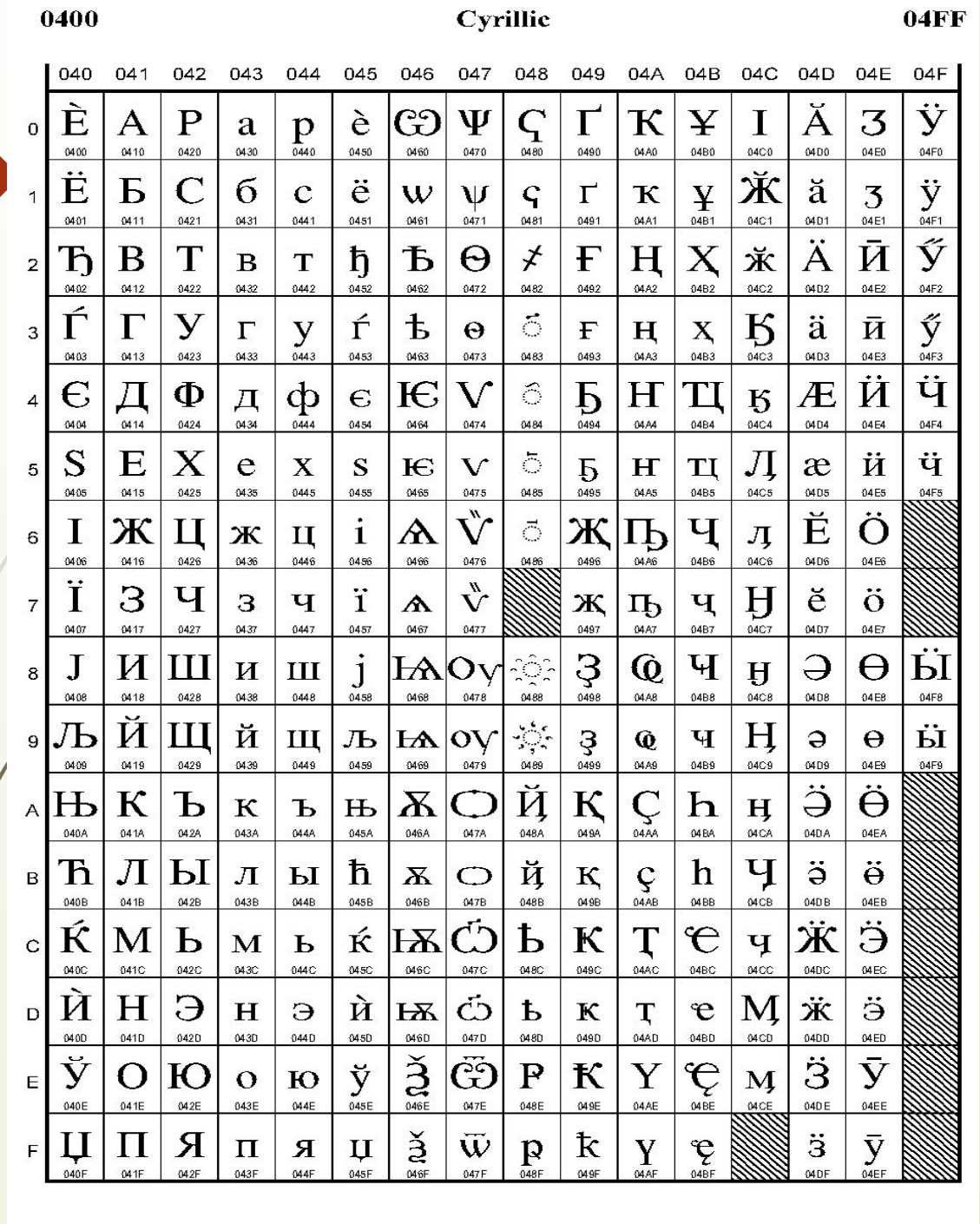
Хронологически одним из первых стандартов кодирования русских букв на компьютерах был КОИ8 ("Код обмена информацией, 8-битный"). Unix



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	─ 128	│ 129	┌ 130	┐ 131	└ 132	┘ 133	├ 134	┤ 135	┬ 136	┴ 137	┼ 138	▀ 139	▄ 140	█ 141	▌ 142	▐ 143
9	░ 144	▒ 145	▓ 146	“ 147	■ 148	• 149	” 150	— 151	№ 152	™ 153	nbsp 154	» 155	® 156	« 157	• 158	¤ 159
A	═ 160	║ 161	╒ 162	ё 163	є 164	╔ 165	і 166	ї 167	╗ 168	╘ 169	╙ 170	╚ 171	╛ 172	ґ 173	ў 174	╞ 175
B	╟ 176	╠ 177	╡ 178	Ё 179	Є 180	╣ 181	І 182	Ї 183	╦ 184	╧ 185	╨ 186	╩ 187	╪ 188	Ґ 189	Ў 190	© 191
C	ю 192	а 193	б 194	ц 195	д 196	е 197	ф 198	г 199	х 200	и 201	й 202	к 203	л 204	м 205	н 206	о 207
D	п 208	я 209	р 210	с 211	т 212	у 213	ж 214	в 215	ь 216	ы 217	з 218	ш 219	э 220	щ 221	ч 222	ъ 223
E	Ю 224	А 225	Б 226	Ц 227	Д 228	Е 229	Ф 230	Г 231	Х 232	И 233	Й 234	К 235	Л 236	М 237	Н 238	О 239
F	П 240	Я 241	Р 242	С 243	Т 244	У 245	Ж 246	В 247	Ь 248	Ы 249	З 250	Ш 251	Э 252	Щ 253	Ч 254	Ъ 255

В конце 90-ых годов появился новый международный стандарт **Unicode,** который ***отводит под один символ не один байт, а два, и поэтому с его помощью можно закодировать не 256, а 65536 различных символов.***

Полная спецификация стандарта Unicode включает в себя все существующие, вымершие и искусственно созданные алфавиты мира, а также множество математических, музыкальных, химических и прочих символов



0400 Cyrillic 04FF

	040	041	042	043	044	045	046	047	048	049	04A	04B	04C	04D	04E	04F
0	Ѐ 0400	А 0410	Р 0420	а 0430	р 0440	ѐ 0450	Ѡ 0460	Ѱ 0470	Ҁ 0480	Ґ 0490	ҚҚ 04A0	ҰҰ 04B0	І 04C0	Ӑ 04D0	Ҙ 04E0	Ӱ 04F0
1	Ё 0401	Б 0411	С 0421	б 0431	с 0441	ё 0451	ѡ 0461	ѱ 0471	ҁ 0481	ґ 0491	ҚҚ 04A1	ҰҰ 04B1	Ӝ 04C1	ӑ 04D1	ҙ 04E1	ӱ 04F1
2	Ђ 0402	В 0412	Т 0422	в 0432	т 0442	ђ 0452	Ѣ 0462	Ѳ 0472	≠ 0482	Ғ 0492	ҢҢ 04A2	ҲҲ 04B2	ӂ 04C2	Ӓ 04D2	Ӣ 04E2	Ӳ 04F2
3	Ѓ 0403	Г 0413	У 0423	г 0433	у 0443	ѓ 0453	ѣ 0463	ѳ 0473	◌҃ 0483	ғ 0493	ҢҢ 04A3	ҲҲ 04B3	Ҋ 04C3	ӓ 04D3	ӣ 04E3	ӳ 04F3
4	Є 0404	Д 0414	Ф 0424	д 0434	ф 0444	є 0454	Ѥ 0464	Ѵ 0474	◌҄ 0484	Ҋ 0494	Ҥ 04A4	ҴҴ 04B4	ҋ 04C4	Ӕ 04D4	Ӥ 04E4	Ӵ 04F4
5	Ѕ 0405	Е 0415	Х 0425	е 0435	х 0445	ѕ 0455	ѥ 0465	ѵ 0475	◌҅ 0485	ҋ 0495	ҥ 04A5	ҵҵ 04B5	Ҍ 04C5	ӕ 04D5	ӥ 04E5	ӵ 04F5
6	І 0406	Ж 0416	Ц 0426	ж 0436	ц 0446	і 0456	Ѧ 0466	Ѷ 0476	◌҆ 0486	Ҍ 0496	ҦҦ 04A6	ҶҶ 04B6	ҍ 04C6	Ӗ 04D6	Ӧ 04E6	
7	Ї 0407	З 0417	Ч 0427	з 0437	ч 0447	ї 0457	ѧ 0467	ѷ 0477		ҍ 0497	ҧҧ 04A7	ҷ 04B7	Ҏ 04C7	ӗ 04D7	ӧ 04E7	
8	Ј 0408	И 0418	Ш 0428	и 0438	ш 0448	ј 0458	Ѩ 0468	Ѹ 0478	◌҇ 0488	Ҏ 0498	Ҩ 04A8	Ҹ 04B8	ҏ 04C8	Ә 04D8	Ө 04E8	Ӹ 04F8
9	Љ 0409	Й 0419	Щ 0429	й 0439	щ 0449	љ 0459	ѩ 0469	ѹ 0479	◌҈ 0489	ҏ 0499	ҩ 04A9	ҹ 04B9	ҐҐ 04C9	ә 04D9	ө 04E9	ӹ 04F9
A	Ѣ 040A	К 041A	Ъ 042A	к 043A	ъ 044A	ѣ 045A	Ѫ 046A	Ѻ 047A	Ґ 048A	Ґ 049A	Ҫ 04AA	ҺҺ 04BA	ґґ 04CA	Ӛ 04DA	Ӧ 04EA	
B	Ѣ 040B	Л 041B	Ы 042B	л 043B	ы 044B	ѣ 045B	ѫ 046B	ѻ 047B	ґ 048B	ґ 049B	ҫ 04AB	ҺҺ 04BB	Ҷ 04CB	ӛ 04DB	ӧ 04EB	
C	Ѓ 040C	М 041C	Ь 042C	м 043C	ь 044C	ѓ 045C	Ѭ 046C	Ѽ 047C	Ѣ 048C	Ґ 049C	Ґґ 04AC	Ҽ 04BC	ҷ 04CC	Ӝ 04DC	Ӝ 04EC	
D	Ѝ 040D	Н 041D	Э 042D	н 043D	э 044D	ѝ 045D	ѭ 046D	ѽ 047D	ѣ 048D	ґ 049D	ґґ 04AD	ҽ 04BD	Ҧ 04CD	ӝ 04DD	ӝ 04ED	
E	Ў 040E	О 041E	Ю 042E	о 043E	ю 044E	ў 045E	Ѯ 046E	Ѿ 047E	Ҕ 048E	Ґ 049E	Ҭ 04AE	Ҿ 04BE	ҧ 04CE	Ӟ 04DE	Ӯ 04EE	
F	Ѿ 040F	П 041F	Я 042F	п 043F	я 044F	ѿ 045F	ѯ 046F	ѿ 047F	ҕ 048F	ґ 049F	ҭ 04AF	ҿ 04BF		ӟ 04DF	ӯ 04EF	

The Unicode Standard 4.0, Copyright © 1991–2003, Unicode, Inc. All rights reserved.

Фрагмент спецификаци и UNICODE 4.0 для кириллицы

Пример 1.

Представьте в форме шестнадцатеричного кода слово «ЭВМ» во всех пяти кодировках. Воспользуйтесь компьютерным калькулятором для перевода чисел из десятичной в

шестнадцатеричную систему счисления

Ответ

- Последовательности десятичных кодов слова «ЭВМ» в различных кодировках составляем на основе кодировочных таблиц:

КОИ8-Р: 252 247 237

CP1251: 221 194 204

CP866: 157 130 140

Mac: 157 130 140

ISO: 205 178 188

- Переводим с помощью калькулятора последовательности кодов из десятичной системы в шестнадцатеричную:

КОИ8-Р: FC F7 ED

CP1251: DD C2 CC

CP866: 9D 82 8C

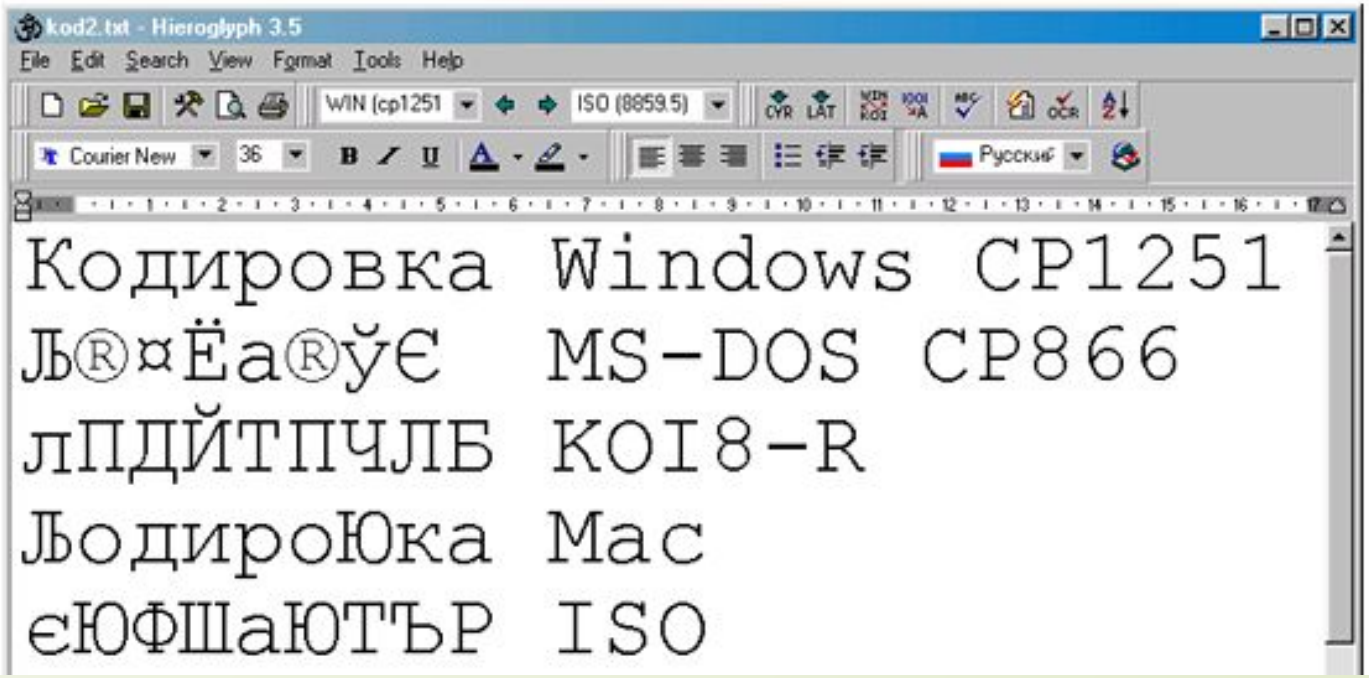
Mac: 9D 82 8C

ISO: CD B2 BC

- Для преобразования русскоязычных текстовых документов из одной кодировки в другую используются специальные программы-конверторы.
- Одной из таких программ является текстовый редактор **Hieroglyph**, который позволяет осуществлять перевод набранного текста из одной кодировки в другую и даже использовать различные кодировки в одном тексте

Пример 2.47. Представить в пяти различных кодировках слово «Кодировка». Выполним это задание с использованием текстового редактора Hieroglyph.

	Перекодирование текста.
1	Запустить текстовый редактор Hieroglyph.
2	В раскрывающемся списке исходных кодировок выбрать кодировку WIN(cp1251) и ввести текст: «Кодировка Windows CP1251».
3	Скопировать текст четыре раза и, выделяя строки, последовательно выбрать в раскрывающемся списке конечные кодировки (DOS, KOI8-R, Mac и ISO), каждый раз нажимая кнопку перекодирования. Для каждой кодировки отредактировать ее название.
4	В результате текст будет состоять из пяти строк, записанных в различных кодировках.



Учимся программировать

Программа Редактирование Вставка Выполнение Инструменты Робот Чертежник Инфо Миры

```
алг Кодовая_таблица
нач
▪ цел а
▪ сим б
▪ нц для а от 32 до 127
▪ ▪ б := символ2(а)
▪ ▪ вывод б,'  '
▪ кц
кон
```

```
а=127
б="▯"
```

Возвращает символ, соответствующий номеру N в таблице Юникода.

```
>> 22:06:03 - Новая программа* - Выполнение начато
  !  "  #  $  %  &  '  (  )  *  +  ,  -  .  /  0  1  2  3  4  5  6  7  8  9  :
 ;  <  =  >  ?  @  A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U
  V  W  X  Y  Z  [  \  ]  ^  _  `  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o
p  q  r  s  t  u  v  w  x  y  z  {  |  }  ~  ⊠
>> 22:06:17 - Новая программа* - Выполнение завершено
```

Возвращает номер символа в таблице КОИ-8r. (стандарт RFC 1489).

Новая программа - Кумир

Программа Редактирование Вставка Выполнение Инструменты Робот Чертежник Инфо Миры

```
1  алг Кодовая_таблица
2  нач
3  ▪ сим а
4  ▪ цел ц
5  ▪ вывод 'введите символ ', нс
6  ▪ ввод а
7  ▪ ц := код(а)
8  ▪ вывод 'код символа',а,'-',ц
9  кон
10
```

а="у"
ц=121

```
>> 22:12:09 - Новая программа* - Выполнение начато
введите символ
у
код символау-121
```

Новая программа - Кумир

Программа Редактирование Вставка Выполнение Инструменты Робот Чертежник Инфо Миры

```
алг Кодовая_таблица
нач
▪ сим a
▪ цел c
▪ вывод 'введите любой символ с клавиатуры ', нс
▪ ввод a
▪ c:=юникод(a)
▪ вывод 'код символа ',a,' в таблице Unicod -',c
кон
```

a="&"
c=38

```
>> 22:24:11 - Новая программа* - Выполнение начато
введите любой символ с клавиатуры
&
код символа & в таблице Unicod -38
>> 22:24:29 - Новая программа* - Выполнение завершено
```