Кодирование символов:

ASCII, KOI8, UNICODE

Все, что мы видим на экране монитора — это символы. Для вывода каждого символа нужен машинный код, который будет соответствовать только этому символу, или же правило, организующее корректный вывод каждого символа на дисплей.

Попробуем прикинуть, сколько же нужно всего символов пользователю: для начала, 26 букв английского алфавита (строчных), во-вторых, 26 прописных, пробел, 10 цифр, 9 знаков препинания (. , : ! " ; ? ( )), 5 арифметических действий (+, — ,*, /, ^) и спецсимволы (№ % _ # $, ^, &, >, <, |, \). В итоге, получаем немногим больше 100. Такой базовый набор символов легко закодировать в двоичной системе счисления от 0 до 127 (всего 128 позиций), что и было сделано.

ASCII

Для отображения всех этих символов была создана таблитца ASCII (*англ. American Standard Code for Information Interchange)* — американский стандартный код для обмена информацией; произносится [э́ски].

Изначально разработана как 7-битная, потом ASCII стала восприниматься как 8-битная. Так выглядят таблицы ASCII-кодов с печатаемыми и непечатаемыми символами (для удобства в таблицах приведены коды в шестнадцатеричной системе счисления).

ASCII-кодировка: печатаемые символы

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

Но скоро набора кодов стало не хватать. Возникла новая таблица кодировок, названная «*расширенная таблица ASCII*», число знакомест в которой возросло до 256. Таблица имела полностью восьми битный код — *Latin-1*.

Дальнейшее развитие привело к появлению понятия «*кодовая страница*», т.е. набор из 256 символов для определения группы языков (например, некоторые славянские языки с латинским алфавитом, турецкий, мальтийский, эсперанто и т.д.), но она не позволяла смешивать языки, и к тому же, не могла создать кодовые страницы японского и китайского языков.

КОИ-8

KOI8 — восьмибитовая **ASCII**-совместимая кодовая страница, созданная для кодирования букв кириллических алфавитов.

В **КОИ-8** символы русского алфавита поместили в верхнюю часть кодовой таблицы так, что позиции кириллических символов соответствуют их фонетическим аналогам в английском алфавите в нижней части таблицы. Это значит, что убрав в тексте, написанном в **КОИ-8**, восьмой бит каждого символа, то получится текст, написанный латинскими символами. Например, слова «*Кодировка*» превратились бы в «*kODIROVKA*».

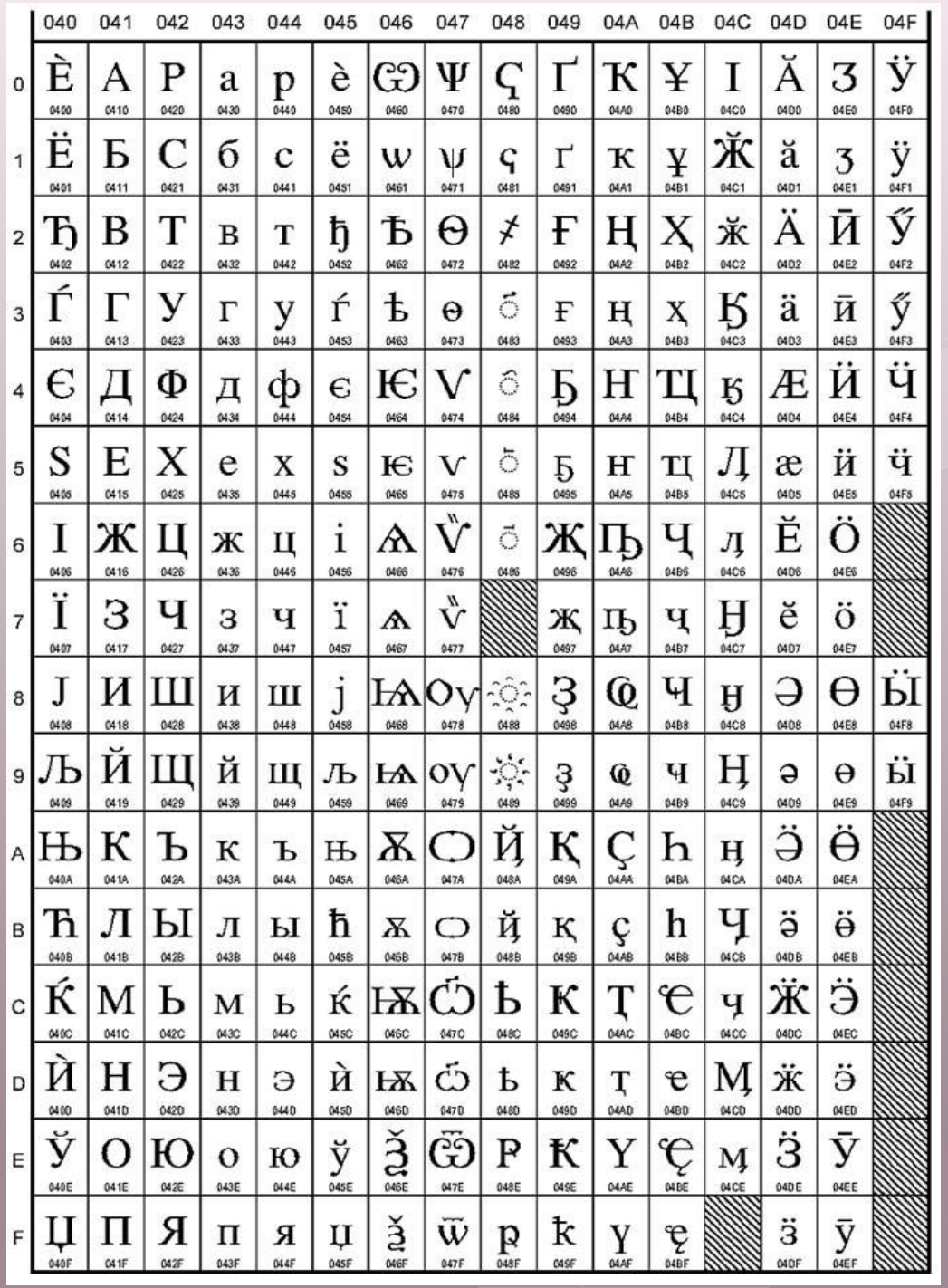
ASCII-кодировка: непечатаемые символы

Число	Команда	Значение
0	NUL	NULL
1	SOH	Start of Heading
2	STX	Start of Text
3	ETX	End of TeXt
4	EOT	End Of Transmission
5	ENQ	ENQurity
6	ACK	ACKnolidgement
7	BEL	BELl
8	BS	Back Space
9	HT	Horizontal Tab
A	LF	Line Feed
B	VT	Vertical Tab
C	FF	From Feed
D	CR	Carriage Return
E	SO	Shift Out
F	Si	Shift In
10	DLE	Data Link Escape
11	DC1	Device Control 1
12	DC2	Device Control 2
13	DC3	Device Control 3
14	DC4	Device Control 4
15	NAK	Negative ACKnolidgement
16	SYN	SYNcronous idle
17	ETB	End of Transmission Block
18	CAN	CANcel
19	EM	End of Medium
1A	SUB	SUBstitude
1B	ESC	ESCape
1C	FS	File Separator
1D	GS	Groupe Separator
1E	RS	Record Separator
1F	DC1	Unit Separator

UNICODE

Юникóд — стандарт кодирования символов, позволяющий представить знаки практически всех письменных языков.

Это новая система кодирования символов, способная закодировать 1 114 112 символов (*code points*). Большинство символов, используемых в основных языках мира занимают 65 536 *code points*. Остальные (более миллиона) *code points* вполне достаточно для кодирования всех известных символов, включая даже исторические знаки и редкие языки. Стандарт **UNICODE** очень обширен, имеет три формы: 32-битную (**UTF-32**), 16-битную (**UTF-16**) и 8-битную (**UTF-8**). Весьма распространенная восьми битная форма **UTF-8** была создана для удобной совместимости с ASCII-ориентированными системами кодирования



“Автоматическое устройство осуществило перекодировку информационного сообщения на русском языке, первоначально записанного в 16-битном коде Unicode, в 8-битную кодировку КОИ-8. При этом информационное сообщение уменьшилось на 480 бит. Какова длина сообщения в символах?

Варианты:
1. 30
2. 60
3. 120
4. 480”

Решение примера.

При перекодировке в 8-битный код, каждый символ уменьшился в «объеме» в два раза (было 16 бит — стало 8). Следовательно, и все сообщение (сумма кодов символов) тоже уменьшилось в 2 раза. Т.к. полученное сообщение стало меньше на 480 бит, то умножив его на 2, мы получим длину исходного. Это 960 бит.
Изначально кодировка была 16-битная, значит разделив исходную длину 960 бит на 16 разрядов, получим кол-во символов. 960/16=60 символов (вариант 2).

Ответ:вариант 2 — 60 символов.