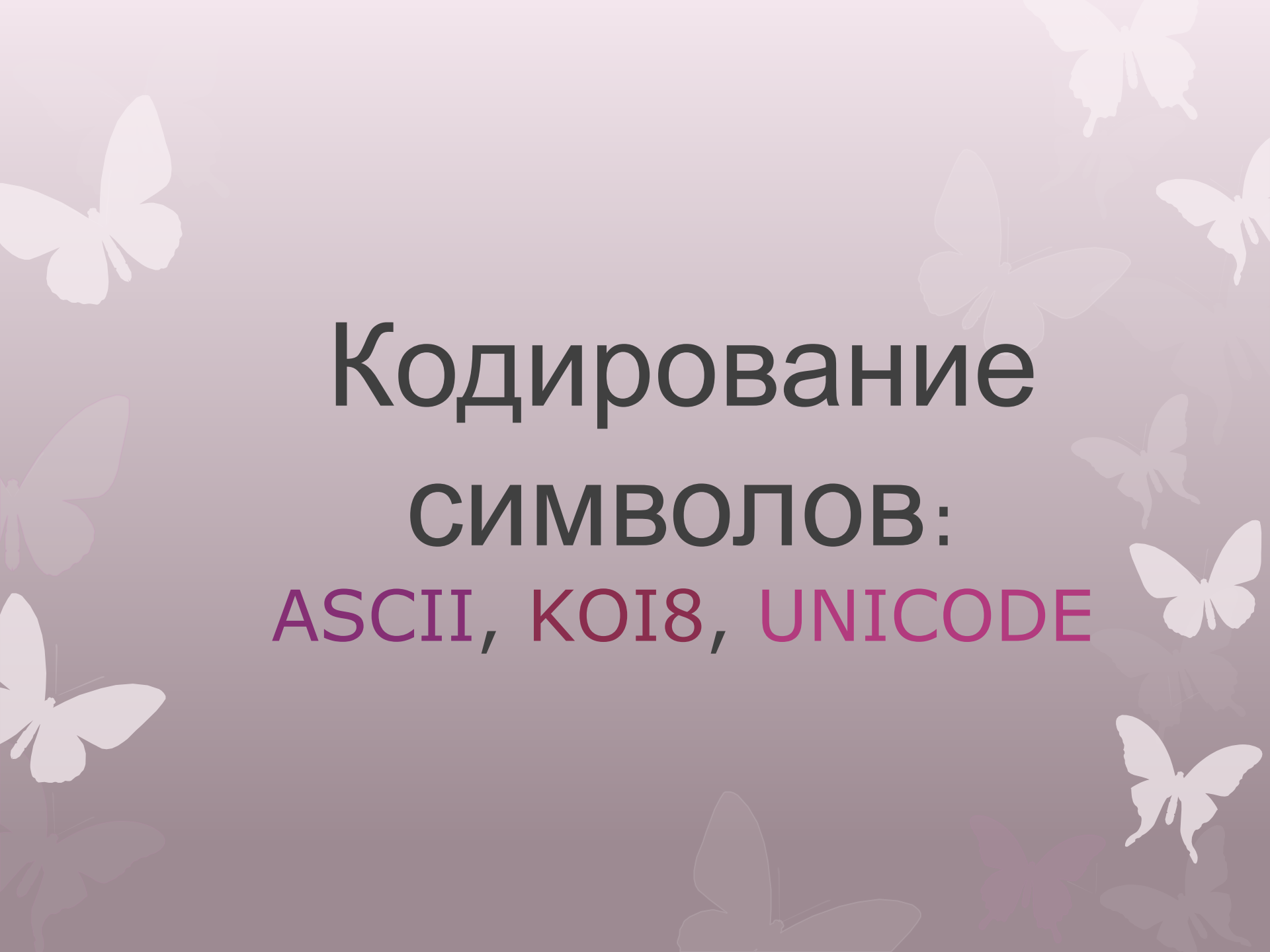




Кодирование СИМВОЛОВ: ASCII, KOI8, UNICODE

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 таблица **ASCII** (англ. *American Standard Code for Information Interchange*) — американский стандартный код для обмена информацией; произносится [эски].

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 «*коддовая 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	ENQuiry
6	ACK	ACKnoligement
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 ACKnoligement
16	SYN	SYNchronous idle
17	ETB	End of Transmission Block
18	CAN	CANcel
19	EM	End of Medium
1A	SUB	SUBstitute
1B	ESC	ESCAPE
1C	FS	File Separator
1D	GS	Groupe Separator
1E	RS	Record Separator
1F	DC1	Unit Separator

UNICODE

Юнико́д — 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

	040	041	042	043	044	045	046	047	048	049	04A	04B	04C	04D	04E	04F
0	È	А	Р	а	р	è	ᄆ	Ψ	Ɔ	Г	К	У	І	Ǻ	З	Û
	0400	0410	0420	0430	0440	0450	0460	0470	0480	0490	04A0	04B0	04C0	04D0	04E0	04F0
1	Ё	Б	С	б	с	ё	ᄇ	ψ	Ɔ	Г	к	у	Ж	ǻ	з	ü
	0401	0411	0421	0431	0441	0451	0461	0471	0481	0491	04A1	04B1	04C1	04D1	04E1	04F1
2	Ђ	В	Т	в	т	ђ	Ђ	Ө	Ɔ	Г	Х	ж	Ǻ	Й	Ў	
	0402	0412	0422	0432	0442	0452	0462	0472	0482	0492	04A2	04B2	04C2	04D2	04E2	04F2
3	Ѓ	Г	У	г	у	ѓ	ѣ	ө	ᄆ	Г	х	Ѓ	ǻ	й	ў	
	0403	0413	0423	0433	0443	0453	0463	0473	0483	0493	04A3	04B3	04C3	04D3	04E3	04F3
4	Є	Д	Ф	д	ф	є	Є	ᄇ	ᄆ	Б	Н	Ц	Ѓ	Æ	Й	Ї
	0404	0414	0424	0434	0444	0454	0464	0474	0484	0494	04A4	04B4	04C4	04D4	04E4	04F4
5	Ѕ	Е	Х	е	х	ѕ	Ѕ	ᄇ	ᄆ	Б	Н	Ц	Л	æ	й	č
	0405	0415	0425	0435	0445	0455	0465	0475	0485	0495	04A5	04B5	04C5	04D5	04E5	04F5
6	І	Ж	Ц	ж	ц	і	Ѓ	ᄇ	ᄆ	Ж	Ѓ	Ч	л	Ё	Ö	
	0406	0416	0426	0436	0446	0456	0466	0476	0486	0496	04A6	04B6	04C6	04D6	04E6	
7	Ї	З	Ч	з	ч	ї	Ѓ	ᄇ		ж	Ѓ	ч	Н	ё	ö	
	0407	0417	0427	0437	0447	0457	0467	0477			0497	04A7	04B7	04C7	04D7	04E7
8	Ј	И	Ш	и	ш	ј	Ѓ	Оу	ᄆ	З	ᄆ	Ч	н	Э	Ө	Ы
	0408	0418	0428	0438	0448	0458	0468	0478	0488	0498	04A8	04B8	04C8	04D8	04E8	04F8
9	Љ	Й	Щ	й	щ	љ	Ѓ	Оу	ᄆ	з	ᄆ	ч	Н	э	ө	Ы
	0409	0419	0429	0439	0449	0459	0469	0479	0489	0499	04A9	04B9	04C9	04D9	04E9	04F9
A	Њ	К	Ъ	к	ъ	њ	Ж	О	Й	К	С	Н	Н	Э	Ө	
	040A	041A	042A	043A	044A	045A	046A	047A	048A	049A	04AA	04BA	04CA	04DA	04EA	
B	Ђ	Л	Ы	л	ы	ђ	ж	О	й	к	с	н	Ч	э	ө	
	040B	041B	042B	043B	044B	045B	046B	047B	048B	049B	04AB	04BB	04CB	04DB	04EB	
C	Ќ	М	Ь	м	ь	ќ	Ж	О	ѣ	К	Т	е	ч	Ж	Э	
	040C	041C	042C	043C	044C	045C	046C	047C	048C	049C	04AC	04BC	04CC	04DC	04EC	
D	Й	Н	Э	н	э	й	Ж	О	ѣ	К	Т	е	М	ж	Э	
	040D	041D	042D	043D	044D	045D	046D	047D	048D	049D	04AD	04BD	04CD	04DD	04ED	
E	Ў	О	Ю	о	ю	ў	Ж	ᄆ	Р	К	У	е	М	Ж	У	
	040E	041E	042E	043E	044E	045E	046E	047E	048E	049E	04AE	04BE	04CE	04DE	04EE	
F	Ц	П	Я	п	я	ц	Ж	ᄇ	Р	к	у	е		э	у	
	040F	041F	042F	043F	044F	045F	046F	047F	048F	049F	04AF	04BF		04DF	04EF	

“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.