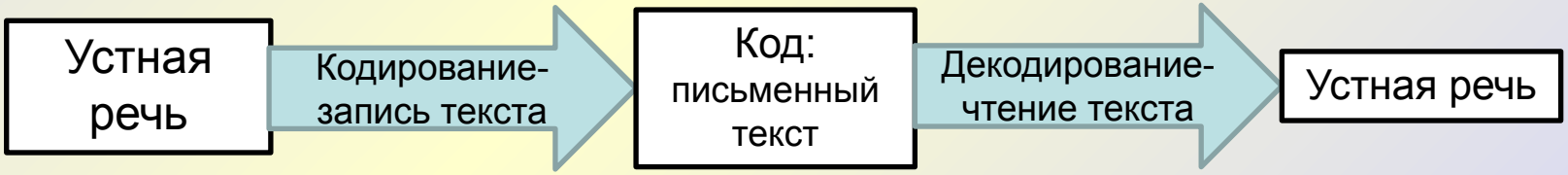
Кодирование текстовой информации

Кодирование - процесс представления информации в виде последовательности условных обозначений

Код – множество слов –последовательность символов из некоторого алфавита, используемых при кодировании информации

Письменность – это способ кодирования устной речи на естественном языке.

Процесс письменного обмена между людьми

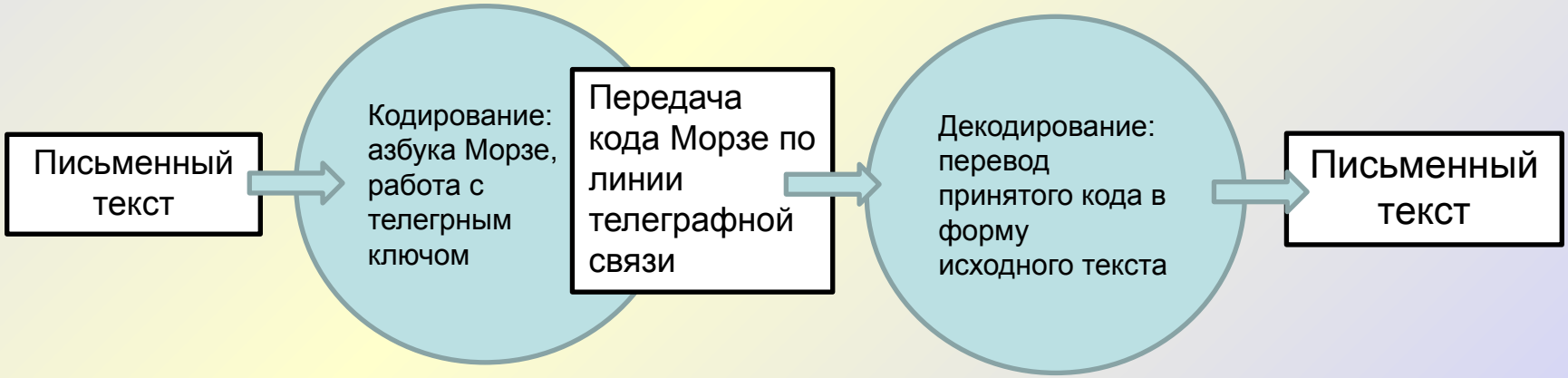


Способ кодирования зависит от назначения кода

Если код предназначен для передачи текста по технической системе связи, то он должен быть приспособлен к возможностям этой системы

Примером технического кода является азбука Морзе

Процесс передачи телеграфного сообщения с использованием азбуки Морзе:



Алфавит телеграфного кода Морзе состоит из три символов: точка, тире, пропуск. Это троичный код.

Во второй половине XX века появляются компьютеры. Для компьютерной обработки текстов потребовалось создать стандарт кодирования.

Первый разработчик **ANSII** Американский национальный институт стандартизации. Впоследствии была создана Международная организация стандартизации **ISO**

В 1963был принят стандарт ASCII Американский стандартный код информационного обмена (American Standard Code for Information Interchang).

sp 32	! 33	" 34	# 35	$ 36	% 37	& 38	' 39	(40	) 41	* 42	+ 43	, 44	- 45	. 46	/ 47
0 48	1 49	2 50	3 51	4 52	5 53	6 54	7 55	8 56	9 57	: 58	; 59	< 60	= 61	> 62	? 63
@ 64	A 65	B 66	C 67	D 68	E 69	F 70	G 71	H 72	I 73	J 74	K 75	L 76	M 77	N 78	O 79
P 80	Q 81	R 82	S 83	T 84	U 85	V 86	W 87	X 88	Y 89	Z 90	[91	\ 92	] 93	^ 94	_ 95
` 96	a 97	b 98	c 99	d 100	e 101	f 102	g 103	h 104	i 105	j 106	k 107	l 108	m 109	n 110	o 111
p 112	q 113	r 114	s 115	t 116	u 117	v 118	w 119	x 120	y 121	z 122	{ 123	| 124	} 125	~ 126	

ASCII - это семиразрядный двоичный код.
Общее количество символов 128, из них 32 символа – управляющие, а остальные «изображаемые», т.е. имеющие графическое изображение

sp 32	! 33	" 34	# 35	$ 36	% 37	& 38	' 39	(40	) 41	* 42	+ 43	, 44	- 45	. 46	/ 47
0 48	1 49	2 50	3 51	4 52	5 53	6 54	7 55	8 56	9 57	: 58	; 59	< 60	= 61	> 62	? 63
@ 64	A 65	B 66	C 67	D 68	E 69	F 70	G 71	H 72	I 73	J 74	K 75	L 76	M 77	N 78	O 79
P 80	Q 81	R 82	S 83	T 84	U 85	V 86	W 87	X 88	Y 89	Z 90	[91	\ 92	] 93	^ 94	_ 95
` 96	a 97	b 98	c 99	d 100	e 101	f 102	g 103	h 104	i 105	j 106	k 107	l 108	m 109	n 110	o 111
p 112	q 113	r 114	s 115	t 116	u 117	v 118	w 119	x 120	y 121	z 122	{ 123	| 124	} 125	~ 126	

Код	Действие	Английское название
7	Подача стандартного звукового сигнала	Beep
8	Удаление предыдущего символа	Back Space (BS)
13	Перевод строки	Line Feed (LF)
26	Признак «Конец текстового файла»	End Of File (EOF)
27	Отмена предыдущего ввода	Escape (Esc)

sp 32	! 33	" 34	# 35	$ 36	% 37	& 38	' 39	(40	) 41	* 42	+ 43	, 44	- 45	. 46	/ 47
0 48	1 49	2 50	3 51	4 52	5 53	6 54	7 55	8 56	9 57	: 58	; 59	< 60	= 61	> 62	? 63
@ 64	A 65	B 66	C 67	D 68	E 69	F 70	G 71	H 72	I 73	J 74	K 75	L 76	M 77	N 78	O 79
P 80	Q 81	R 82	S 83	T 84	U 85	V 86	W 87	X 88	Y 89	Z 90	[91	\ 92	] 93	^ 94	_ 95
` 96	a 97	b 98	c 99	d 100	e 101	f 102	g 103	h 104	i 105	j 106	k 107	l 108	m 109	n 110	o 111
p 112	q 113	r 114	s 115	t 116	u 117	v 118	w 119	x 120	y 121	z 122	{ 123	| 124	} 125	~ 126	

ASCII Code Chart

	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	HT	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

Коды символов могут быть двоичными, десятичными и шестнадцатеричными

Символ	Десятичный код	Двоичный код	Символ	Десятичный код	Двоичный код
Пробел	32	00100000	0	48	00110000
!	33	00100001	1	49	00110001
#	35	00100011	2	50	00110010
$	36	00100100	3	51	00110011
*	42	00101010	4	52	00110100
+	43	00101011	5	53	00110101
,	44	00101100	6	54	00110110
–	45	00101101	7	55	00110111
.	46	00101110	8	56	00111000
/	47	00101111	9	57	00111001
A	65	01000001	N	78	01001110
B	66	01000010	O	79	01001111
C	67	01000011	P	80	01010000
D	68	01000100	Q	81	01010001
E	69	01000101	R	82	01010010
F	70	01000110	S	83	01010011
G	71	01000111	T	84	01010100
H	72	01001000	U	85	01010101
I	73	01001001	V	86	01010110
J	74	01001010	W	87	01010111
K	75	01001011	X	88	01011000
L	76	01001100	Y	89	01011001
M	77	01001101	Z	90	01011010

Символы в ASCII кодируются семью битами, но в памяти компьютера под каждый символ отводится ровно 1 байт (старший бит не используется).
Важным свойством **ASCII** является соблюдение алфавитной последовательности кодировки строчных и прописных букв и десятичных цифр.

ASCII Code Chart

	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	HT	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

Символ	Десятичный код	Двоичный код	Символ	Десятичный код	Двоичный код
Пробел	32	00100000	0	48	00110000
!	33	00100001	1	49	00110001
#	35	00100011	2	50	00110010
$	36	00100100	3	51	00110011
*	42	00101010	4	52	00110100
+	43	00101011	5	53	00110101
,	44	00101100	6	54	00110110
–	45	00101101	7	55	00110111
.	46	00101110	8	56	00111000
/	47	00101111	9	57	00111001
A	65	01000001	N	78	01001110
B	66	01000010	O	79	01001111
C	67	01000011	P	80	01010000
D	68	01000100	Q	81	01010001
E	69	01000101	R	82	01010010
F	70	01000110	S	83	01010011
G	71	01000111	T	84	01010100
H	72	01001000	U	85	01010101
I	73	01001001	V	86	01010110
J	74	01001010	W	87	01010111
K	75	01001011	X	88	01011000
L	76	01001100	Y	89	01011001
M	77	01001101	Z	90	01011010

Вопрос. Почему сдвиг, с помощью которого по коду прописной английской буквы можно получить код соответствующей строчной, равен 32, а не 26?

В чем главный недостаток **ASCII?**

sp 32	! 33	" 34	# 35	$ 36	% 37	& 38	' 39	(40	) 41	* 42	+ 43	, 44	– 45	. 46	/ 47
0 48	1 49	2 50	3 51	4 52	5 53	6 54	7 55	8 56	9 57	: 58	; 59	< 60	= 61	> 62	? 63
@ 64	A 65	B 66	C 67	D 68	E 69	F 70	G 71	H 72	I 73	J 74	K 75	L 76	M 77	N 78	O 79
P 80	Q 81	R 82	S 83	T 84	U 85	V 86	W 87	X 88	Y 89	Z 90	[91	\ 92	] 93	^ 94	_ 95
` 96	a 97	b 98	c 99	d 100	e 101	f 102	g 103	h 104	i 105	j 106	k 107	l 108	m 109	n 110	o 111
p 112	q 113	r 114	s 115	t 116	u 117	v 118	w 119	x 120	y 121	z 122	{ 123	| 124	} 125	~ 126	

Впоследствии стали разрабатывать расширения ASCII, в которых применялись однобайтовые коды символов, первые 128 совпадали с кодировкой **ASCII, остальные для кодирования букв национального алфавита. Из-за несогласованности этих разработок было создано по несколько вариантов таких таблиц.**

Для русского языка наиболее распространенные однобайтовые кодовые таблицы CP-866, Windows-1251(CP-1251), КОИ-8 . Первая часть 0-127 совпадает с ASCII, во второй половине коды русских букв, но они не совпадают в этих таблицах.

К чему приводит несовпадение кодовых таблиц?

КОИ8 (koi-8r)

(Код обмена информацией, 8-битный, ОС Unix).

Для представления букв других языков СССР использовался блок псевдографических символов. Эту кодировку легко отличить от других по необычному порядку русских букв. Этот порядок приближен к порядку букв в латинском алфавите.

	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	² 157	· 158	÷ 159
A	═ 160	║ 161	╒ 162	ё 163	╓ 164	╔ 165	╕ 166	╖ 167	╗ 168	╘ 169	╙ 170	╚ 171	╛ 172	╜ 173	╝ 174	╞ 175
B	╟ 176	╠ 177	╡ 178	Ё 179	╢ 180	╣ 181	╤ 182	╥ 183	╦ 184	╧ 185	╨ 186	╩ 187	╪ 188	╫ 189	╬ 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

CP1251 ("Code Page", «кодовая страница» или Windows-1251)

Эта кодировка использовалась и используется до сих пор в операционных системах семейства Windows. Основные особенности заключаются в порядке русских букв удобном для сортировки и отсутствии псевдографики.

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

Кодировка CP866 используется в ОС MS DOS).

К особенностям данной кодировки можно отнести наличие псевдографики, которая врезается в середину списка строчных букв русского языка. Использовалась в так же в советских клонах IBM PC.

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

- **Задание.** Какое слово отобразится в кодировках СР-866, Windows-1251, если в кодировке КОИ-8 набрано слово «ДИСК»

В начале 90-ых годов появился новый международный стандарт **Unicode,** в котором на кодирование символов отводится 31 бит.

- **0-127** коды полностью совпадают с **ASCII,**
- **128-65 536** основные алфавиты современных языков
- **> 65536** все существующие, вымершие и искусственно созданные алфавиты мира, а также множество математических, музыкальных, химических и прочих символов

В современных компьютерах и операционных системах используется укороченная 16-битовая версия Unicode, в которую входят все современные алфавиты. Эта часть Unicode называется **базовой многоязыковой страницей BMP –Base Multilingual Plane**

Ответьте на вопрос:
Что такое стандарт ASCII. Принцип кодирования.

Задание.
1.Представьте в форме шестнадцатеричного кода слово «ЭВМ» во трех кодировках.

2. С помощью кодировочной таблицы **ASCII** декодируйте шестнадцатеричную запись:

494E464F524D4154494F4E20544543484E4F4C4F4759

Символ	Десятичный код	Двоичный код	Символ	Десятичный код	Двоичный код
Пробел	32	00100000	0	48	00110000
!	33	00100001	1	49	00110001
#	35	00100011	2	50	00110010
$	36	00100100	3	51	00110011
*	42	00101010	4	52	00110100
+	43	00101011	5	53	00110101
,	44	00101100	6	54	00110110
–	45	00101101	7	55	00110111
.	46	00101110	8	56	00111000
/	47	00101111	9	57	00111001
A	65	01000001	N	78	01001110
B	66	01000010	O	79	01001111
C	67	01000011	P	80	01010000
D	68	01000100	Q	81	01010001
E	69	01000101	R	82	01010010
F	70	01000110	S	83	01010011
G	71	01000111	T	84	01010100
H	72	01001000	U	85	01010101
I	73	01001001	V	86	01010110
J	74	01001010	W	87	01010111
K	75	01001011	X	88	01011000
L	76	01001100	Y	89	01011001
M	77	01001101	Z	90	01011010

Вопрос. Почему сдвиг, с помощью которого по коду прописной английской буквы можно получить код соответствующей строчной, равен 32, а не 26?

Ответ1

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

КОИ8-Р: 252 247 237
СР1251: 221 194 204
СР866: 157 130 140

- Переводим последовательности кодов из десятичной системы в шестнадцатеричную:

КОИ8-Р: FC F7 ED
СР1251: DD C2 CC
СР866: 9D 82 8C

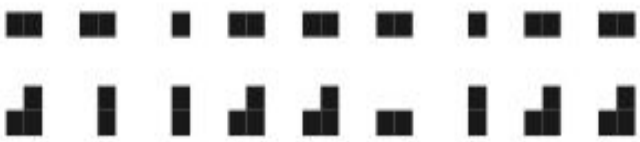
Ответ2

- **INFORMATION TECHNOLOGY**

Задача 1

Цепочка ПТИУААМДЛ получена перестановкой букв в некотором слове. Имеется последовательность цифр, задающая порядок, в котором надо выписать буквы цепочки для получения исходного слова. Каждая цифра записывалась в прямоугольный шаблон размера 5 на 3 пикселей по образцу

При передаче часть пикселей на местах, одинаковых для каждой цифры, стерлись. Получилось вот что:



Восстановите исходное слово и перехваченную перестановку.

Кодирование

равномерное

при равномерном кодировании все символы кодируются **кодами равной длины**;

неравномерное

при неравномерном кодировании разные символы могут кодироваться **кодами разной длины,** это затрудняет декодирование

Однозначное декодирование

- закодированное сообщение можно однозначно декодировать с начала, если выполняется *условие Фано*: **никакое кодовое слово не является началом другого кодового слова;**

- закодированное сообщение можно однозначно декодировать с конца, если выполняется *обратное условие Фано*: **никакое кодовое слово не является окончанием другого кодового слова;**

- условие Фано – это достаточное, но не необходимое условие однозначного декодирования.

Однозначное декодирование

- Для кодирования некоторой последовательности, состоящей из букв А, Б, В, Г и Д, решили использовать неравномерный двоичный код, позволяющий однозначно декодировать двоичную последовательность, появляющуюся на приёмной стороне канала связи. Использовали код:
  А–1, Б–000, В–001, Г–011. Укажите, каким кодовым словом должна быть закодирована буква Д. Длина этого кодового слова должна быть наименьшей из всех возможных. Код должен удовлетворять свойству однозначного декодирования.
  1) 00 ***2)*** 01 ***3)***11 ***4)*** 010

Пример 2.47. Представить в пяти различных кодировках слово «Кодировка». Выполним это задание с спользованием 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.