Глава 2

Кодирование и обработка текстовой информации

2.1. Кодирование текстовой информации

Двоичное кодирование текстовой информации в компьютере. Информация, выраженная с помощью естественных и формальных языков в письменной форме, обычно называется текстовой информацией.

Для представления текстовой информации (прописные и строчные буквы русского и латинского алфавитов, цифры, знаки и математические символы) достаточно 256 различных знаков. По формуле (1.1) можно вычислить, какое количество информации необходимо, чтобы закодировать каждый знак:

$$N = 2^I \Rightarrow 256 = 2^I \Rightarrow 2^8 = 2^I \Rightarrow I = 8 \text{ битов.}$$

1.3.2. Определение количества информации **Информатика и ИКТ-8**

Для обработки текстовой информации в компьютере необходимо представить ее в двоичной знаковой системе. Для кодирования каждого знака требуется количество информации, равное 8 битам, т. е. длина двоичного кода знака составляет восемь двоичных знаков. Каждому знаку необходимо поставить в соответствие уникальный двоичный код в интервале от 00000000 до 11111111 (в десятичном коде от 0 до 255) (табл. 2.1).

Человек различает знаки по их начертанию, а компьютер — по их двоичным кодам. При вводе в компьютер текстовой информации происходит ее двоичное кодирование, изображение знака преобразуется в его двоичный код. Пользователь нажимает на клавиатуре клавишу со знаком, и в компьютер поступает определенная последовательность из восьми электрических импульсов (двоичный код знака). Код знака хранится в оперативной памяти компьютера.

Таблица 2.1. **Кодировки знаков**

Двоичный код	Десятичный код	КОИ-8	Windows	MS-DOS	Mac	ISO
00000000	0					
...	...	...				
00001000	8	удаление последнего символа (клавиша {Backspace})				
...	...	...				
00001101	13	перевод строки (клавиша {Enter})				
...	...	...				
00100000	32	клавиша {Пробел}				
00100001	33	!				
...	...	...				
01011010	90	Z				
...	...	...				
10000000	128	—	Ъ	А	А	к
...	...	...	...	...	...	...
11000010	194	б	В	—	—	Т
...	...	...	...	...	...	...
11001100	204	л	М	|	|	Ь
...	...	...	...	...	...	
11011101	221	щ	Э	_	Ё	н
...	...	...	...	...	...	
11111111	255	ь	я	неразделяемый пробел	неразделяемый пробел	п

В процессе вывода знака на экран компьютера производится обратное кодирование, т. е. преобразование двоичного кода знака в его изображение.

Различные кодировки знаков. Присвоение знаку конкретного двоичного кода — это вопрос соглашения, которое фиксируется в кодовой таблице. Первые 33 кода в кодовой таблице (десятичные коды с 0 по 32) соответствуют не знакам, а операциям (перевод строки, ввод пробела и т. д.).

Десятичные коды с 33 по 127 являются интернациональными и соответствуют знакам латинского алфавита, цифрам, знакам арифметических операций и знакам препинания.

Десятичные коды с 128 по 255 являются национальными, т. е. в различных национальных кодировках одному и тому же коду соответствуют разные знаки. К сожалению, в настоящее время существуют пять различных кодовых таблиц для русских букв (*Windows*, *MS-DOS*, *КОИ-8*, *Mac*, *ISO*), поэтому тексты, созданные в одной кодировке, не будут правильно отображаться в другой.

Кодовые таблицы Windows-CD

В последние годы широкое распространение получил новый международный стандарт кодирования текстовых символов *Unicode*, который отводит на каждый символ 2 байта (16 битов). По формуле (1.1) определим количество символов, которые можно закодировать:

$$N = 2^I = 2^{16} = 65\,536.$$

Такого количества символов оказалось достаточно, чтобы закодировать не только русский и латинский алфавиты, цифры, знаки и математические символы, но и греческий, арабский, иврит и другие алфавиты.

Итак, в настоящее время имеется шесть различных кодировок для букв русского алфавита, в которых один и тот же знак имеет различные коды (табл. 2.2). К счастью, в большинстве случаев пользователь не должен заботиться о перекодировках текстовых документов, так как это делают специальные программы-конверторы, встроенные в операционную систему и приложения.

Таблица 2.2. **Десятичные коды некоторых знаков в различных кодировках**

Символ	Windows	MS-DOS	КОИ-8	Mac	ISO	Unicode
А	192	128	225	128	176	1040
В	194	130	247	130	178	1042
М	204	140	237	140	188	1052
Э	221	157	252	157	205	1069
я	255	239	241	223	239	1103

Например, в кодировке *Windows* последовательность числовых кодов 221 194 204 образует слово «ЭВМ» (см. табл. 2.2), тогда как в других кодировках это будет бессмысленный набор символов.