# Кодирование текстовой информации

Компьютеры третьего поколения «научились» работать с текстовой информацией. Текстовая информация по своей природе дискретна, т. к. представляется последовательностью отдельных символов.

Для компьютерного представления текстовой информации достаточно:

1) определить множество всех символов (алфавит), требуемых для представления текстовой информации;

2) выстроить все символы используемого алфавита в некоторой последовательности (присвоить каждому символу алфавита свой номер);

3) получить для каждого символа n-разрядный двоичный код ($n \leq 2^n$), переведя номер этого символа в двоичную систему счисления.

В памяти компьютера хранятся специальные кодовые таблицы, в которых для каждого символа указан его двоичный код. Все кодовые таблицы, используемые в любых компьютерах и любых операционных системах, подчиняются международным стандартам кодирования символов.

# Кодировка ASCII и её расширения

Основой для компьютерных стандартов кодирования символов послужил код ASCII (American Standard Code for Information Interchange) — американский стандартный код для обмена информацией, разработанный в 1960-х годах в США и применявшийся для любых, в том числе и некомпьютерных, способов передачи информации (телеграф, факсимильная связь и т. д.). Этот код 7-битовый: общее количество символов составляет $2^7 = 128$, из них первые 32 символа — управляющие, а остальные — изображаемые, т. е. имеющие графическое изображение. К изображаемым символам в ASCII относятся буквы латинского алфавита (прописные и строчные), цифры, знаки препинания и арифметических операций, скобки и некоторые специальные символы. Кодировка ASCII приведена в табл. 3.8.

Таблица 3.8

**Кодировка ASCII**

| | 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 |

Хотя для кодирования символов в ASCII достаточно 7 битов, в памяти компьютера под каждый символ отводится ровно 1 байт (8 битов), при этом код символа помещается в младшие биты, а в старший бит заносится 0.

Например, 01000001 — код прописной латинской буквы «A»; с помощью шестнадцатеричных цифр его можно записать как 41.

Стандарт ASCII рассчитан на передачу только английского текста. Со временем возникла необходимость кодирования и неанглийских букв. Во многих странах для этого стали разрабатывать расширения ASCII -кодировки, в которых применялись однобайтовые коды символов. При этом первые 128 символов кодовой таблицы совпадали с кодировкой ASCII, а остальные (со 128-го по 255-й) использовались для кодирования букв национального алфавита, символов национальной валюты и т. п. Из-за несогласованности этих разработок для многих языков было создано несколько вариантов кодовых таблиц (например, для русского языка их было создано около десятка!).

Впоследствии использование кодовых таблиц было несколько упорядочено: каждой кодовой таблице было присвоено особое название и номер. Для русского языка наиболее распространёнными стали однобайтовые кодовые таблицы CP-866, Windows-1251 (табл. 3.9) и КОИ-8 (табл. 3.10). В них первые 128 символов совпадают с ASCII-кодировкой, а русские буквы размещены во второй части таблицы. Обратите внимание на то, что коды русских букв в этих кодировках различны.

Таблица 3.9

**Кодировка Windows-1251**

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | Ђ | Ѓ | ‚ | ѓ | „ | … | † | ‡ | € | ‰ | Љ | ‹ | Њ | Ќ | Ћ | Џ |
| 9 | ђ | ‘ | ’ | “ | ” | • | – | — | | ™ | љ | › | њ | ќ | ћ | џ |
| A | | Ў | ў | Ј | ¤ | Ґ | ¦ | § | Ё | © | Є | « | ¬ | | ® | Ї |
| B | ° | ± | І | і | ґ | µ | ¶ | · | ё | № | є | » | ј | Ѕ | ѕ | ї |
| C | А | Б | В | Г | Д | Е | Ж | З | И | Й | К | Л | М | Н | О | П |
| D | Р | С | Т | У | Ф | Х | Ц | Ч | Ш | Щ | Ъ | Ы | Ь | Э | Ю | Я |
| E | а | б | в | г | д | е | ж | з | и | й | к | л | м | н | о | п |
| F | р | с | т | у | ф | х | ц | ч | ш | щ | ъ | ы | ь | э | ю | я |

Таблица 3.10

**Кодировка КОИ-8**

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

Мы выяснили, что при нажатии на алфавитно-цифровую клавишу в компьютер посылается некоторая цепочка нулей и единиц. В текстовых файлах хранятся не изображения символов, а их коды.

При выводе текста на экран монитора или принтера необходимо восстановить изображения всех символов, составляющих данный текст, причём изображения эти могут быть разнообразны и достаточно причудливы. Внешний вид выводимых на экран символов кодируется и хранится в специальных шрифтовых файлах. Современные текстовые процессоры умеют внедрять шрифты в файл. В этом случае файл содержит не только коды символов, но и описание используемых в этом документе шрифтов. Кроме того, файлы, создаваемые с помощью текстовых процессоров, включают в себя и такие данные о форматировании текста, как его размер, начертание, размеры полей, отступов, межстрочных интервалов и другую дополнительную информацию.

## Стандарт Unicode

Ограниченность 8-битной кодировки, не позволяющей одновременно пользоваться несколькими языками, а также трудности, связанные с необходимостью преобразования одной кодировки в другую, привели к разработке нового кода. В 1991 году был разработан новый стандарт кодирования символов, получивший название Unicode (Юникод), позволяющий использовать в текстах любые символы любых языков мира.

Unicode — это «уникальный код для любого символа, независимо от платформы, независимо от программы, независимо от языка» (www.unicode.org).

В Unicode на кодирование символов отводится 31 бит. Первые 128 символов (коды 0-127) совпадают с таблицей ASCII. Далее размещены основные алфавиты современных языков: они полностью умещаются в первой части таблицы, их коды не превосходят 65 536 = $2^{16}$.

Стандарт Unicode описывает алфавиты всех известных, в том числе и «мёртвых», языков. Для языков, имеющих несколько алфавитов или вариантов написания

(например, японского и индийского), закодированы все варианты. В кодировку Unicode внесены все математические и иные научные символьные обозначения и даже некоторые придуманные языки (например, язык эльфов из трилогии Дж. Р. Р. Толкина «Властелин колец»).

Всего современная версия Unicode позволяет закодировать более миллиона различных знаков, но реально используется чуть менее 110 000 кодовых позиций.

Для представления символов в памяти компьютера в стандарте Unicode имеется несколько кодировок.

В операционных системах семейства Windows используется кодировка UTF-16. В ней все наиболее важные символы кодируются с помощью 2 байт (16 бит), а редко используемые — с помощью 4 байт.

В операционной системе Linux применяется кодировка UTF-8, в которой символы могут занимать от 1 (символы, входящие в таблицу ASCII) до 4 байт. Если значительную часть текста составляют цифры и латинские буквы, то это позволяет в несколько раз уменьшить размер файла по сравнению с кодировкой UTF-16.

 Кодировки Unicode позволяют включать в один документ символы самых разных языков, но их использование ведёт к увеличению размеров текстовых файлов.

## Информационный объём текстового сообщения

Мы уже касались этого вопроса, рассматривая алфавитный подход к измерению информации.

 **Информационным объёмом** текстового сообщения называется количество бит (байт, килобайт, мегабайт и т. д.), необходимых для записи этого сообщения путём заранее оговоренного способа двоичного кодирования.

Оценим в байтах объём текстовой информации в современном словаре иностранных слов из 740 страниц, если на одной странице размещается в среднем 60 строк по 80 символов (включая пробелы).

Будем считать, что при записи используется кодировка «один символ — один байт». Количество символов во всем словаре равно:

$$80 \cdot 60 \cdot 740 = 3\,552\,000.$$

Следовательно, объём равен

$$3\,552\,000 \text{ байт} = 3\,468{,}75 \text{ Кбайт} \approx 3{,}39 \text{ Мбайт}.$$

Если же использовать кодировку UTF-16, то объём этой же текстовой информации в байтах возрастёт в 2 раза и составит 6,78 Мбайт.

## Самое главное

Текстовая информация по своей природе дискретна, т. к. представляется последовательностью отдельных символов.

В памяти компьютера хранятся специальные кодовые таблицы, в которых для каждого символа указан его двоичный код. Все кодовые таблицы, используемые в любых компьютерах и любых операционных системах, подчиняются международным стандартам кодирования символов.

Основой для компьютерных стандартов кодирования символов послужил код ASCII, рассчитанный на передачу только английского текста. Расширения ASCII — кодировки, в которых первые 128 символов кодовой таблицы совпадают с кодировкой ASCII, а остальные (со 128-го по 255-й) используются для кодирования букв национального алфавита, символов национальной валюты и т. п.

В 1991 году был разработан новый стандарт кодирования символов, получивший название Unicode (Юникод), позволяющий использовать в текстах любые символы любых языков мира. Кодировки Unicode позволяют включать в один документ символы самых разных языков, но их использование ведёт к увеличению размеров текстовых файлов.

### Вопросы и задания

1. Какова основная идея представления текстовой информации в компьютере?

2. Что представляет собой кодировка ASCII? Сколько символов она включает? Какие это символы?

3. Как известно, кодовые таблицы каждому символу алфавита ставят в соответствие его двоичный код. Как, в таком случае, вы можете объяснить вид таблицы 3.8 «Кодировка ASCII»?

4. С помощью таблицы 3.8:

1) декодируйте сообщение 64 65 73 6B 74 6F 70;
2) запишите в двоичном коде сообщение TOWER;
3) декодируйте сообщение

01101100 01100001 01110000 01110100 01101111 01110000

5. Что представляют собой расширения ASCII-кодировки? Назовите основные расширения ASCII-кодировки, содержащие русские буквы.

6. Сравните подходы к расположению русских букв в кодировках Windows-1251 и КОИ-8.

7. Представьте в кодировке Windows-1251 текст «Знание — сила!»:

1) шестнадцатеричным кодом;
2) двоичным кодом;
3) десятичным кодом.

8. Представьте в кодировке КОИ-8 текст «Дело в шляпе!»:

1) шестнадцатеричным кодом;
2) двоичным кодом;
3) десятичным кодом.

9. Что является содержимым файла, созданного в современном текстовом процессоре?

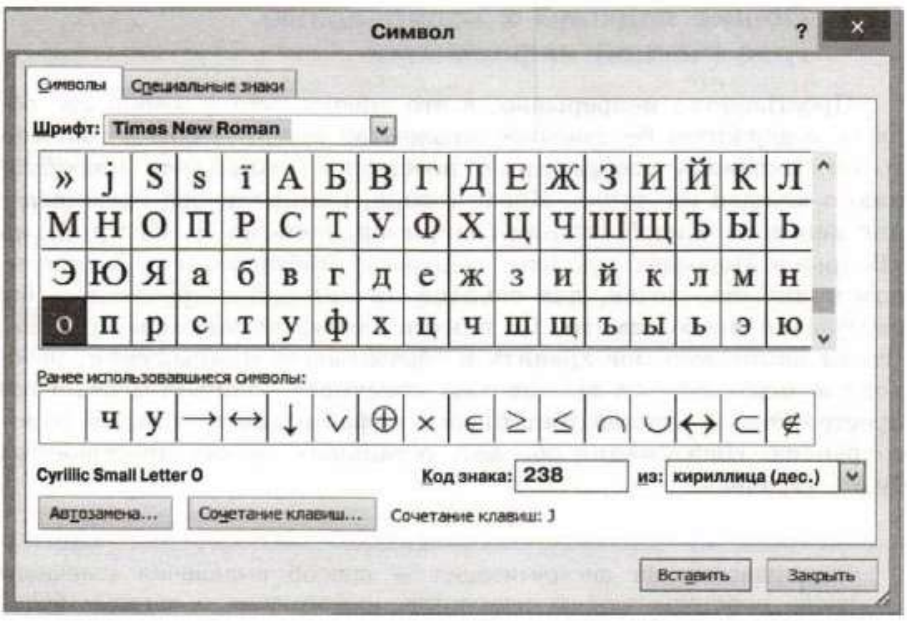
10. В кодировке Unicode на каждый символ отводится 2 байта. Определите в этой кодировке информационный объём следующей строки:

Где родился, там и сгодился.

11. Набранный на компьютере текст содержит 2 страницы. На каждой странице 32 строки, в каждой строке 64 символа. Определите информационный объём текста в кодировке Unicode, в которой каждый символ кодируется 16 битами.

12. Текст на русском языке, первоначально записанный в 8-битовом коде Windows, был перекодирован в 16-битную кодировку Unicode. Известно, что этот текст был распечатан на 128 страницах, каждая из которых содержала 32 строки по 64 символа в каждой строке. Каков информационный объём этого текста?

13. В текстовом процессоре MS Word откройте таблицу символов (вкладка **Вставка → Символ → Другие символы**):



В поле **Шрифт** установите **Times New Roman**, в поле **из — кириллица** (дес.).

Вводя в поле **Код знака** десятичные коды символов, декодируйте сообщение:

| | | | | | | |
|---|---|---|---|---|---|---|
| 196 | 238 | 240 | 238 | 227 | 243 | 32 |
| 238 | 241 | 232 | 235 | 232 | 242 | 32 |
| 232 | 228 | 243 | 249 | 232 | 233 | 46 |