КОДУВАННЯ ТЕКСТОВИХ ДАНИХ

Руденко Віктор Дмитрович,
провідний науковий співробітник Інституту педагогіки НАПН України,
кандидат педагогічних наук, доцент.

Анотація. Наведені основні поняття і терміни кодування даних, розглянута еволюція систем кодування даних, описані принципи кодування текстових даних.

Ключові слова: кодування даних, текстові дані, еволюція систем кодування даних, скан-код, ASCII.

Сутність кодування й основні поняття

Інформація у суспільстві, науці, техніці, мистецтві тощо представляється за допомогою відповідної мови. Існують дві основні групи мов: природні і штучні.

Природні — це мови, за допомогою яких спілкуються між собою люди. Таких мов налічується у світі близько 2000. Вони функціонують в основному в межах однієї нації. Шість із них (англійська, російська, французька, німецька, іспанська і китайська) визнані офіційними мовами Організації Об'єднаних Націй (ООН). Будь-яка природня мова містить символи алфавіту, з яких за допомогою граматики конструюються слова і речення.

Алфавіт — це набір певної кількості знаків (символів), за допомогою яких створюються повідомлення. У природних мовах такими знаками є букви і синтаксичні знаки. На папері вони мають певне накреслення (наприклад, а, б, ?, f), а під час розмови озвучуються.

Граматика — сукупність правил будови слів, словосполучень та речень. Граматика складається з морфології й синтаксису.

Морфологія — частина граматики, що вивчає природу й будову слів, а також правила зміни слів.

Синтаксис — сукупність правил будови словосполучень і речень.

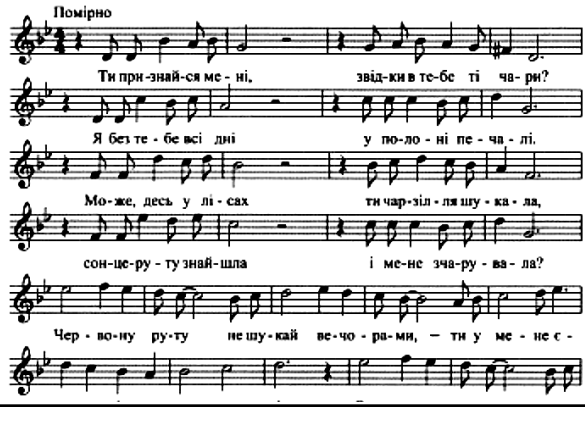
Штучні (або формальні) мови створені для представлення деякого виду інформації у певній сфері людської діяльності. У них максимально враховані правила й традиції природних мов. Штучні мови мають міжнаціональний характер. До таких мов можна віднести:

- мову науки (математики, хімії, біології тощо)

$$(a+b)^2=a^2+2ab+b^2,$$

$$S(t)=a_0+\sum_{n=1}^{+\infty}a_n\cos(n\omega t)+b_n;$$

- мову мистецтва (музики, скульптури, живопису);



- мову міміки й жестів;

- мову дорожнього руху;

- спеціальні мови (азбука Брайля для незрячих, азбука Морзе та інші). До спеціальних мов належать і мови програмування.

Одну й ту саму інформацію можна подати різними способами. Це залежить від багатьох факторів, на-

самперед, від обраної мови представлення інформації, а також типів джерела й приймача інформації. Наприклад, інформація про місце, дату й час проведення футбольного матчу Динамо-Шахтар може розповсюджуватися за допомогою повідомлень по радіо чи телебаченню, плакатів тощо.

Надалі будемо розглядати способи й методи представлення інформації лише в пристроях обчислювальної техніки. У сучасних комп'ютерних системах для подання інформації застосовуються такі типи даних: числові, текстові, графічні, звукові, відео.

Усі наведені типи даних у комп'ютерних системах кодуються за відповідними правилами. Кодування — це процес представлення даних у такій формі, яку сприймає й опрацьовує комп'ютер. Кодування даних у комп'ютерних системах має низку принципових особливостей, до яких можна віднести такі.

Перша. Кожний тип даних має власні правила і системи кодування. Але будь-який тип даних зберігається й опрацьовується лише двійковими символами 0 і 1. Таке кодування називають двійковим. Наприклад, літера R може мати такий двійковий код: 1010010. Зворотний процес, тобто процес перетворення даних із двійкового коду у форму, яку сприймає людина, називають декодуванням.

Друга. Різні програмні засоби комп'ютера використовують різні системи кодування. Наприклад, системи кодування текстових редакторів, графічних систем, баз даних та інших суттєво відрізняються одна від одної. У той же час незалежно від типу програмного засобу кожна клавіша клавіатури має постійний код, який називається скан-кодом. Під час уведення даних драйвер клавіатури перетворює скан-коди натиснутих клавіш у внутрішні коди комп'ютера, найчастіше в код ASCII. Потім код ASCII перетворюється в інші коди.

Третя. У комп'ютері зберігаються й опрацьовуються не абстрактні двійкові символи 0 і 1, за допомогою яких закодовано певний тип даних, а реальні фізичні сигнали, що відповідають цим двійковим символам. Оскільки комп'ютер складається з різних фізичних пристроїв, то застосовуються й різні фізичні сигнали. Наприклад, в основній пам'яті символу 1 відповідає напруга 2,4–2,7В, а символу 0 — 0,4–0,6В. На оптичних носіях цим символам відповідають темні й світлі крапки, а на магнітних пристроях — позитивна і негативна остаточна магнітна індукція. Під час передавання даних з одного пристрою в інший виникає проблема узгодження різних типів фізичних сигналів.

Нижче наведені основні терміни, що найчастіше застосовуються у процесі кодування даних у комп'ютерних системах.

Код — це завідомо узгоджене співвідношення між символами різних типів даних і фізичними сигналами, наприклад між символами кирилиці і двійковими символами. У процесі кодування даних часто використовуються таблиці, за допомогою яких встановлюється відповідність між різними знаковими системами. Узгодження може бути на рівні декількох корпорацій, на державному й міжнародному рівнях. Коли певний код набуває масового застосування, він вважається стандартом.

Символи — це букви, цифри, знаки пунктуації та інші знаки. У комп'ютері символи подаються різними типами фізичних сигналів.

Сигнал — фізична величина, яка використовується для подання символу. Сигнали ототожнюються з двійковими величинами 0 і 1, якими кодуються символи.

Еволюція кодування в системах передавання даних

Способи й методи кодування даних у сучасних комп'ютерах не виникли самі собою. Вони засновані на ідеях кодування символів у системах передавання даних, які існують вже декілька століть. У цих системах способи і методи кодування неодноразово змінювалися в міру того, як розвивалися апаратні засоби передавання даних. Нижче стисло описана еволюція методів кодування в системах передавання даних.

Розвиток систем кодування символів фактично розпочався з появою електрики. У 1753 році в одному з шотландських журналів був запропонований метод кодування за принципом — кожній букві окремий провід. Пропонувалося між містами прокласти 26 ліній зв'язку (для кожної букви окремий провід), якими будуть передаватися електричні сигнали. Однак, складність прокладання ліній зв'язку і значні економічні витрати не дозволили втілити його в життя.

У 1833 році Фредерік Гаус запропонував метод кодування 25 символів (букви I і J були об'єднані) за допомогою матриці 5×5. Ідея полягала в такому: по одному проводу передавався електричний сигнал 5-ти градацій, від якого стрілка відхилялася на певне значення вправо, а потім — електричний сигнал також 5-ти градацій, від якого стрілка відхилялася вліво. Перше значення визначало номер рядка в матриці, а друге — номер стовпця. Наприклад, якщо перше відхилення стрілки було зафіксовано на значенні 3, а друге — на значенні 2, то це визначало, що була передана літера M (рис. 1).

У XIX столітті американський винахідник Семюель Морзе запропонував систему кодування символів за допомогою коротких і довгих сигналів (крапка, тире). Наприклад, буква A кодується коротким і довгим сигналом (• – –), буква S трьома короткими сигналами (• • •), цифра 1 коротким і трьома довгими сигналами (• – – –). Максимальна кількість сигналів для одного символа дорівнює 5. Короткі кодові комбінації використані для символів, що часто зустрічаються, а довгі — для тих, що зустрічаються рідко.

Метод кодування Морзе заснований на знаннях людини самої системи кодування і його здібностях розпізнавати (декодувати) сигнали, які він чує. У 1844 році за допомогою апарату Морзе було передане перше телеграфне повідомлення з Вашингтона до

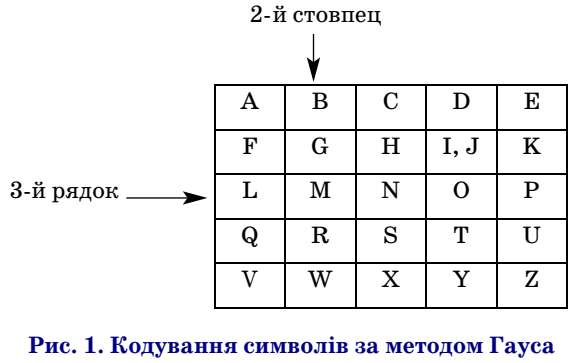


Рис. 1. Кодування символів за методом Гауса

Балтимору, а в 1866 році розпочалася ера міжнародного телеграфу, коли між США і Францією був прокладений трансатлантичний кабель.

Зазначимо, що код Морзе має суттєвий недолік. У ньому використовується різна довжина кодових комбінацій. Кодувати символи під час передавання й декодування їх під час приймання могла тільки спеціально підготовлена людина. Для автоматичного кодування й декодування він непридатний. Тому в другій половині дев'ятнадцятого століття було розроблено декілька кодів, у яких кожний символ кодувався однаковим числом сигналів. Сам сигнал мав два стани: увімкнено-вимикнемо, тобто кодування здійснювалося на основі двійкових сигналів.

Найбільш відомий код запропонував у 1870 році французький інженер Еміль Бодо. Оскільки на той час для передавання електричних сигналів використовувалися повільні електромеханічні пристрої, то Е. Бодо обмежив кількість сигналів у кодовій комбінації п'ятьма. Це давало можливість мати 32 кодові комбінації (2^5=32), що недостатньо для 26 літер, 10 десяткових цифр і синтаксичних знаків. Тому Е. Бодо використав 26 комбінацій для літер і 6 комбінацій для керуючих символів. Зокрема, кодова комбінація 11111 використовувалася для перемикання на регістр літер (LTRS), а кодова комбінація 11011 — для перемикання на регістр цифр (FIGS). Зазначимо, що коди керуючих символів LTRS і FIGS, як і інші керуючі символи завжди інтерпретуються однаково, незалежно від того, у якому регістрі знаходиться приймальний пристрій. Фрагмент таблиці Бодо наведено у табл. 1.

Код Бодо став основою для розробки стандартного телеграфного апарату, у якому кожна кодова комбінація починалася сигналом СТАРТ і яка закінчувалася сигналом СТОП. Телеграфні апарати на основі коду Бодо розвивалися більше 50 років. Для того щоб розрізняти великі й малі букви, пізніше став використовуватися 6-розрядний код.

Нині в епоху розвитку комп'ютерної техніки існують різні коди для передавання символів. Найрозповсюдженішими з них є такі:

- міжнародний алфавіт №2 CCIEE — простий 5-розрядний код, який використовується для передавання телексних повідомлень;
- код EBCDIC — використовується, в основному у синхронних системах зв'язку з мейнфреймами;
- код ASCII (American Standard Code for Information Interchange). Це 7-розрядний код, який затверджений стандартом ANSI X3.4-1977. За його допомогою кодується 128 символів: усі великі й малі букви англійського алфавіту, цифри від 0 до 9, синтаксичні й інші символи, а також 32 керуючих символів, які не виводяться на екран і друк.

Таблиця 1

Сигнали коду					У регістрі LTRS	У регістрі FIRS
1	2	3	4	5		
•	•				A	C
•			•	•	B	?
	•	•	•		C	:
•	•	•	•	•	LTRS	
•	•		•	•	FIGS	
		•			Проміжок	
		•		•	Повернення каретки	
	•	•		•	Перехід на новий рядок	

Крім перерахованих кодів у системах передавання інформації використовуються й інші системи кодування символів. Наприклад, популярними в країнах колишнього Радянського Союзу були семирозрядний код обміну інформацією (КОИ-7), восьмирозрядний код обміну інформацією (КОИ-8), двійковий код для обробки інформації (ДКОИ). Вони використовуються й нині, зокрема, в Інтернеті.

Такою є стисла історія розвитку методів кодування символьної інформації в системах передавання даних.

Кодування текстових даних

Текстові дані уводяться в комп'ютер, зазвичай, за допомогою клавіатури. На будь-якій сучасній клавіатурі комп'ютера розміщені символи англійського і національного алфавітів, синтаксичні і деякі спеціальні знаки, десяткові цифри, символи функціональних і спеціалізованих клавіш (**Ctrl**, **Enter** та інші). Кожній клавіші присвоєний так званий скан-код, що реєструється контролером клавіатури під час натиснення і відпускання клавіші. Скан-код — це фактично порядковий номер клавіші, він генерується на апаратному рівні і не залежить ні від програми, що виконується у даний момент, ні від символу, що нанесений на клавішу. У клавіатурах старого зразка (IBM PC/XT) скан-код — це двійковий код довжиною в один байт, сім молодших розрядів якого відведені під власне номер клавіші, а восьмий розряд приймає значення 0 при натисненні клавіші й 1 при її відпусканні. Отже, при відпусканні клавіші її скан-код збільшується на 80 (тут і надалі числа наведені у шістнадцятковій системі числення). Наприклад, під час натиснення клавіші **A** на клавіатурі типу XT генерується код **1E**, а після відпускання — код **9E**.

У клавіатурах типу IBM PC/AT для скан-коду найчастіше відводяться два байти. У разі натиснення й відпускання клавіші її скан-код не змінюється, але після відпускання старший байт набуває значення **F0**. Наприклад, на клавіатурі AT скан-код натиснутої клавіші A має значення **1C**, а скан-код відпускання — **F01C**. Для деяких службових клавіш скан-код займає більше двох байтів.

Якщо натиснутою була клавіша керування або комбінація цих клавіш, то відповідний сигнал керування передається в ОС. Якщо натиснута буквено-цифрова клавіша або клавіша синтаксичного чи спеціального знаку, то з пам'яті відеоконтролера вибирається відповідний код знакогенерації для відображення знаку на екрані монітора.

Отже, під час натиснення і відпускання клавіші контролер клавіатури реєструє її скан-код. Але для обробки даних у комп'ютері використовуються не скан-коди, а внутрішні коди символів. Перетворення скан-коду клавіші з урахуванням стану клавіш **Shift**, **Ctrl**, **Alt**, **CapsLook** та інших у внутрішній код здійснюється за допомогою драйвера клавіатури. Для цього використовуються спеціальні таблиці кодування, які встановлюють

однозначну відповідність між скан-кодами клавіш і символами клавіатури. Нині для кодування символів клавіатури застосовуються різні методи, тобто різні кодові таблиці. Наприклад, для кирилиці використовуються такі кодові таблиці: КОІ-8, CP1251, CP866, Mac, ISO. Відзначимо, що символи, закодовані за допомогою однієї таблиці, не будуть однаково відображатися на екрані у разі використання іншої таблиці. У табл. 2 наведено приклад значення двійкового коду 11000010 і відповідні йому символи для різних кодових таблиць.

У сучасних персональних комп'ютерах найрозповсюдженішею є таблиця кодування ASCII. У базовому варіанті коду ASCII для будь-якого символу клавіатури використовується один байт. Нижня половина кодової таблиці, тобто коди молодших 7-ми біт (коди 0–127) у всьому світі застосовуються для кодування стандартного набору символів (символів латинського алфавіту, цифр, знаків арифметичних і логічних операцій та деяких інших), наразі перші 32 комбінацій (код 0–31) відведені для символів керування комп'ютером. Друга половина таблиці (коди 128–255) використовуються для кодування символів національного алфавіту, а також символів псевдографіки. На рис. 2 наведена нижня половина кодової таблиці ASCII за стандартом ANSI X3.4-1977, яку змінювати не можна.

Цю таблицю часто називають основною або базовою таблицею ASCII. З рис. 2 видно, що, наприклад, символ F має код 1000110 (у шіснадцятковій системі 46), символ 5 — код 0110101 (35), символ h — 1101000 (68). Великі й малі літери відрізняються тільки значенням 6-го розряду. Наприклад, символ A має код 1000001 (41), а символ a — 1100001 (61). Перші 32 кодові комбінації таблиці, тобто комбінації від 0000000 (00) до 0011111 (1F), відведені під символи керування. Вони не виводяться на екран, а використовуються для спеціальних цілей, зокрема для передавання команд периферійним пристроям, наприклад, для управління друкарськими пристроями. Наведемо призначення деяких керуючих кодових комбінацій: CR — повертання каретки, LF — переведення рядка, FF — перехід на початок наступної сторінки, HT — горизонтальна табуляція, SI, SO — перемикання між англійськими і національними символами. Для посилання на символи керування часто використовуються позначення CHR(N) або CHR$(N). Таке позначення розуміється як символ, код якого дорівнює N.

Загальна схема формування зображена на рис. 3.

Базова таблиця ASCII (див. рис. 2) була застосована в комп'ютерах IBM PC для внутрішнього подання символів. Однак, оскільки комп'ютери PC заполонили фактично весь світ, то система кодування ANSI фактично стала загальновизнаною в усьому світі. У той же час 7-розрядний код ASCII не забезпечував кодування національних алфавітів. Тому стандартом ISO 646 була введена нова 8-розрядна версія коду ASCII. Восьмий розряд надав ще 128 кодових комбінацій, які могли використовуватися в різних цілях, у тому числі для подання національного алфавіту.

Отже, можна вважати, що кодова таблиця ASCII складається з двох половин: перша половина містить кодові комбінації від 00 до 7F і є базовою, обов'язковою для всіх розробників персональних комп'ютерів і користувачів. Друга половина містить кодові комбінації від 80 до FF. Її можуть використовувати розробники і користувачі у своїх цілях, у тому числі для подання національних алфавітів.

Таблиця 2

Двійковий код	КОІ8	CP1251	CP866	Mac	ISO
11000010	Б	В	-	-	Т

Номери розрядів 7 6 5 4 3 2 1				0	0	0	0	1	1	1	1
				0	0	1	1	0	0	1	1
				0	1	0	1	0	1	0	1
0	0	0	0	NUT	DLE	Проміжок	0	@	P	.	P
0	0	0	1	SOH	DC1	!	1	A	Q	A	Q
0	0	1	0	STX	DC2	“	2	B	R	B	R
0	0	1	1	ETX	DC3	#	3	C	S	C	S
0	1	0	0	EOT	DC4	$	4	D	T	D	T
0	1	0	1	ENQ	NAK	%	5	E	U	E	U
0	1	1	0	ACK	SYN	&	6	F	V	F	V
0	1	1	1	BEL	ETB	’	7	G	W	G	W
1	0	0	0	BS	CAN	(	8	H	X	H	X
1	0	0	1	HT	EM	)	9	I	Y	I	Y
1	0	1	0	LF	SUB	*	:	J	Z	J	Z
1	0	1	1	VT	ESC	+	;	K	[	K	{
1	1	0	0	FF	FS	,	<	L		L	|
1	1	0	1	CR	GS	-	=	M	]	M	}
1	1	1	0	SO	RS	.	>	N	^	N	-
1	1	1	1	SI	US	/	?	O	C	O	DEL

Рис. 2. Система кодування ASCII символів клавіатури

Однак, щоб не було повного хаосу з використанням другої половини кодової таблиці ASCII (її часто називали розширеним кодом ASCII), стандартом ISO 8859 було введено поняття кодової сторінки, кожна з яких містила 16 кодових комбінацій. Були також розроблені рекомендації щодо кодування певної мови або групи мов, у тому числі для кириличного алфавіту. Однак, фірма Microsoft запропонувала свій варіант кодування цього алфавіту, так звану кодову таблицю Windows CP1251. Ураховуючи широке розповсюдження операційної системи Windows на ринку країн СНД та інших програмних засобів компанії Microsoft, ця кодова таблиця є більш популярною, ніж кодування, яке рекомендовано ISO.

Кодування кириличних символів у системі Windows-1251 наведено в табл. 3. Перша половина таблиці використовується для кодування деяких синтаксичних і спеціальних символів.

Таблиця 3

	8	9	A	B	C	D	E	F
0					А	Р	а	р
1					Б	С	б	с
2					В	Т	в	т
3					Г	У	г	у
4					Д	Ф	д	ф
5					Е	Х	е	х
6					Ж	Ц	ж	ц
7					З	Ч	з	ч
8					И	Ш	и	ш
9					Й	Щ	й	щ
A					К	Ъ	к	ъ
B					Л	Ы	л	ы
C					М	Ь	м	ь
D					Н	Э	н	э
E					О	Ю	о	ю
F					П	Я	п	я

Другий варіант коду ASCII має довжину 2 байти. Його називають розширеним кодом. Він застосовується для функціональних клавіш, а також комбінацій клавіш з клавішею **Alt**.

Оскільки нині персональні комп'ютери розповсюджені по всьому світу, то, безумовно, бажано було мати єдину систему кодування символів мов народів усього світу. Але система кодування ASCII цього не забезпечує, тому що має тільки 256 кодових комбінацій, а мови народів світу мають близько 200 000 символів. Як крок для розв'язання цієї проблеми є запроваджена групою комп'ютерних компаній нового 16-розрядного міжнародного стандарту ISO 10646 під назвою Unicode (Юнікод), який має 65536 кодових комбінацій. Програми MS Windows Office підтримують це кодування починаючи з 1997 року. Основна ідея Unicode — поставити кожному символу єдине 16-розрядне значення, яке називається вказівником коду. Юнікод має декілька версій, але найпоширенішими є: (UTF (UnicodeTransformationFormat — формат перетворення Юнікоду) і (UCS (UniversalCharacterSet — універсальна таблиця символів). Число після UTF вказує на кількість бітів, виділених на один символ, а число після UCS — кількість байтів. UTF широко застосовується для передавання символів через Інтернет (чат, електронна пошта тощо).

Коди в Юнікоді поділені на області (рис. 3). Коди від 0000 до 007F — це символи набору ASCII, а коди від 0400 до 052F відведені для символів кирилиці.

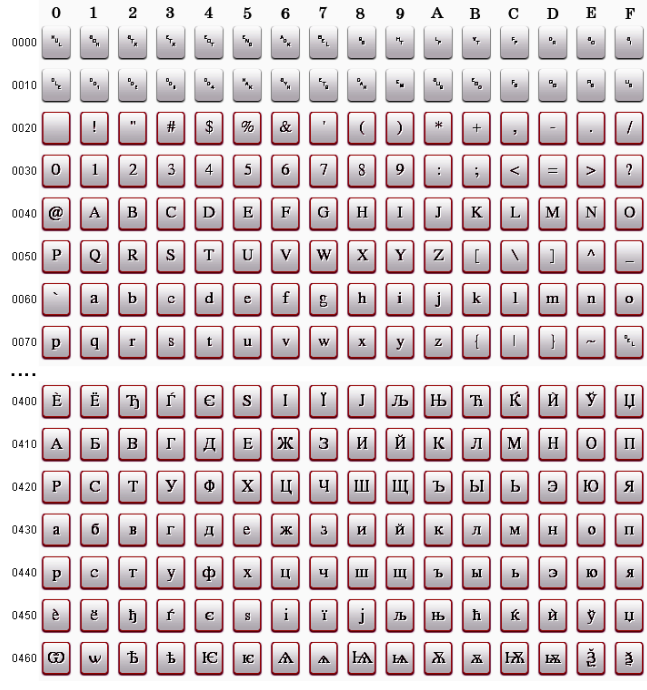


Рис. 3. Коди символів в Юнікоді

Для визначення шістнадцяткового коду символу в системі Unicode необхідно у системі MS Word на вкладці Вставлення натиснути кнопку Символ, потім — Інші символи. У результаті відріється вікно Символ (рис. 4). У нижній частині цього вікна у шістнадцатковій системі числення висвітлюється код вибраного символу.

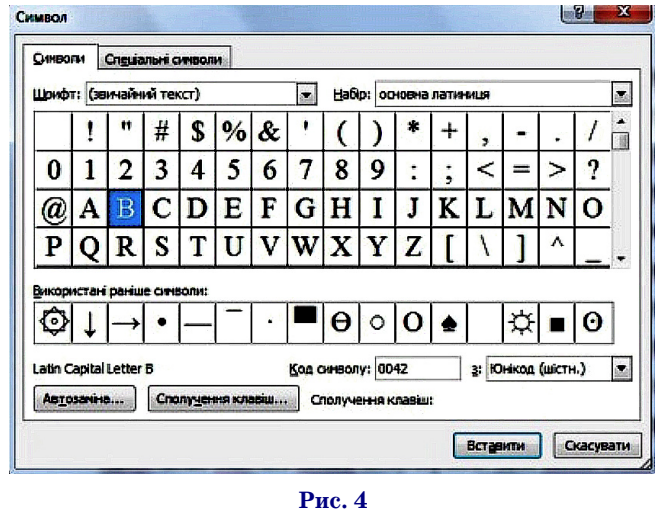


Рис. 4

Нині більша частина кодової таблиці UNICODE вже розподілена. Причому перші кодові комбінації від 0000h до 00FF відведені під код ASCII (базова таблиця і таблиця з англійськими літерами зі штрихами, так званий набір Latin-1). Ця таблиця поділена на блоки, кожний по 16 кодів. Декілька блоків створюють зону. Зона може мати різну кількість блоків. Зона закріплюється за певною мовою. Російська зона має 256 кодів.

Література

1. Технологии передачи данных. 7-е изд. /Г. Хелд. — СПб.: Питер, К.: Издательская група BHV, 2003. — 720 с.