7. Alphabets and character sets

7.1. Transliteration/transcription of languages with Non-Latin scripts

7.1.1. Transliteration/transcription

All material from languages that are written in a writing system other than the Latin alphabet must be transliterated/transcribed.

As a rule, one transliterates when the original spelling reflects the pronunciation fairly closely, and one transcribes when the orignal spelling does not represent the sounds well. In Europe, only Modern Greek is generally transcribed. All other languages that use non-Latin writing systems are transliterated.

7.1.2. The Cyrillic alphabet

7.1.2.1. The Slavic languages

Six modern Slavic languages are written in a version of the Cyrillic alphabet: Russian, Ukrainian, Belorussian, Bulgarian, Macedonian, and eastern Serbo-Croatian (as well as Old Church Slavonic, which is written in Cyrillic in most textbooks and editions).

The following letters are common to all Cyrillic alphabets and are uniformly transcribed: note

а	a	о	o
б	b	п	p
в	v	р	r
д	d	с	s
е	e	т	t
ж	z	у	u
з	z	ф	f
к	k	ц	c
л	l	ч	č
м	m	ш	š
н	n		

Each language has some special letters that do not exist in all languages or are transliterated differently in different languages.

Russian

г	g	ъ	"
ё	ë	ы	y
и	i	ь	′
й	j	э	è

х	x	ю	ju
щ	šč	я	ja

Ukrainian

		є	ji
г	h	й	j
ґ	g	х	x
ѓ	je	щ	šč
и	y	ю	ju
і	i	я	ja

Belorussian

		ы	y
г	h	ь	´
і	i	э	è
й	i	ю	ju
ў	w	я	ja
х	x		

Bulgarian

		щ	št
г	g	ъ	â, ă
и	i	ю	ju
й	j	я	ja
х	x		

Serbo-Croatian (eastern)

		љ	lj
г	g	њ	nj
ћ	ć	х	h
и	i	џ	dž
ј	j	ђ	đ

Macedonian

		ѕ	dz
г	g	и	i
ѓ	ǵ	ј	j
ќ	ć	х	h
љ	lj	џ	dž
њ	nj		

Old Church Slavonic

		лъ	ĭ
г	g	ь	ĕ
s	dz	ю	ju
и	i	ѣ	ja
і	i	є	je
ђ	ǵ	ѧ	ę
оу	u	ѫ	ǫ
х	x	ѩ	ję
щ	št	ѭ	jǫ
ъ	ŭ		
ы	y		

7.1.2.2. Non-Slavic languages

Many non-Slavic languages spoken in Russia and other parts of the former Soviet Union use the Cyrillic alphabet. No transliteration conventions for these languages are given here because

(a) most of these languages are not widely studied outside of Russia and the former Soviet Union

(b) some of them (especially the languages of Central Asia) are in the process of shifting to other alphabets, especially the Latin and the Arabic alphabets

7.1.3. The Greek alphabet

The Greek alphabet is used for Ancient Greek and for Modern Greek. For Ancient Greek, a transliteration is used because the spelling reflected the pronunciation fairly closely. For Modern Greek, linguists usually use a transcription because many spelling conventions from Ancient Greek are still used in Modern Greek spelling.

7.1.3.1. Ancient Greek

Cf. Martinet, André. 1953. "A project of transliteration of Classical Greek." *Word* 9.2: 152-161.

α	a	ν	n
β	b	ξ	x
γ	g	ο	o
δ	d	π	p
ε	e	ρ	r
ζ	z	σ	s
η	ē	τ	t

θ	th	υ	u
ι	i	φ	ph
κ	k	χ	kh
λ	l	ψ	ps
μ	m	ω	ō

Special conventions:

(a) accent marks and trema are written as in Greek,

e.g. ά = á, έ = è, ΐ = ĩ, αϋ = aü
αί = aí, εῦ = eũ

(b) spiritus asper is transliterated as *h*, e.g. ἁ = ha

(c) iota subscriptum is not subscript, e.g. ῃ = ēi, ῳ = ōi; ᾳ is āi

(d) γ before nasal stop can be transliterated as *n*, e.g. ἄγγελος= ángelos

7.1.3.2. Modern Greek

Cf. Joseph, Brian & Philippaki-Warburton, Irene. 1987. *Modern Greek*. London: Routledge.

(a) Vowel inventory: /i e a o u/, transcribed as in IPA.

(b) Consonant inventory:

/p	t		c	k
b	d			g
f	θ	s	ç	x
v	ð	z	j	γ
m		n		
		l	λ	
		r/		

(c) The sixteen consonants /p, t, k, b, d, g, f, s, x, v, z, j, m, n, l, r/ are transcribed as in IPA.

(d) The fricatives /θ, ð, γ/ may be transcribed as in IPA; or alternatively they may be transcribed by the digraphs <th>, <dh>, <gh> for typographic convenience.

(e) The palatal sonorants /ɲ, ʎ/ are always transcribed by the digraphs <nj>, <lj>.

(f) The palatal obstruents (c, ɟ, ç/ contrast with /k, g, x/ only before back vowels, so they need to be distinguished from these only in this environment; before front vowels, dorsal obstruents are always palatal. Thus,

και	/ce/	<ke>

κιόλας /colas/ <kjólas>

before back vowels:

/c/ = <kj>

/ɟ/ = <gj>

/ç/ = <xj>

before front vowels:

/c/ = <k>

/ɟ/ = <g>

/ç/ = <x>

Stress should also be indicated in Modern Greek transcriptions.

7.1.4. The Hebrew alphabet

Among European languages that are widely studied, only Yiddish is written in the Hebrew alphabet. The transliteration of YIVO (New York center for Yiddish studies) should be used.

א	a	ײ	ey
ב	b	ײ	ay
ג	g	כ,ך	kh
ד	d	ל	l
ה	h	מ,ם	m
ו	u	נ,ן	n
וי	oy	ס	s
וו	v	ע	e
ז	z	פ	p
זש	zh	פ,ף	f
ח	(Hebrew)	צ,ץ	ts
ט	t	ק	k
טש	tsh	ר	r
י	i	ש	sh
יי	y	ת	(Hebrew)

Two letters are only used in Hebrew loanwords. Note that Hebrew loanwords are not transliterated, but transcribed.

7.1.5. Others

7.1.5.1.

For the Armenian alphabet, see

> Minassian, Martiros. 1980. *Grammaire d'Arménien oriental*. Delmar, N.Y.: Caravan Books.

7.1.5.2.

For the Georgian alphabet, see

Vogt, Hans. 1971. *Grammaire de la language géorgienne*. Oslo.

For both Armenian and Georgian, see also

> Comrie, Bernard. 1981. *The languages of the Soviet Union*. Cambridge: Cambridge University Press, p. 288-289.

7.2. Survey of systems in use

Rendering other languages than English (and possibly Latin and Dutch) has been a problem as long as computers have existed, and it has been particularly acute for linguists, who more often than not need to give examples from several languages within one text. Gradually, computer hardware and software has become more suited to fulfil these needs, but at present, we are still far from a general and well-working solution. At least two proposals exist for a 'universal character set' which would comprise virtually all characters needed to render the languages of the world and in addition, the special characters used e.g. in phonetic transcriptions and mathematical and logical formulae. The most promising one at present seems to be Unicode, a two-byte encoding system for characters developed by an international consortium (Unicode Inc., 1965 Charleston Rd, Mountain View, CA 94043, USA.) Representing each character as two bytes means theoretically that there is room for 65,536 characters, which makes it possible to include not only the Roman alphabet and its extensions but also e.g. the essential parts of the Han characters that form the basis for the Chinese, Japanese and Korean writing systems. The fact that major software companies are members of the Unicode consortium gives good hope that it will be adopted in the future. However, this is of little help in the present situation: virtually all existing software packages build on one-byte representations of characters, which makes it impossible to exploit Unicode's principles. (For details on Unicode, see Sheldon 1991).

The development of word-processing systems has now got so far that representing West European languages is usually no problem, and files can also relatively easily be transferred between the major word-processors and between the IBM and Macintosh worlds, provided that you follow the instructions in the manuals. Also, it is usually possible, with varying degrees of difficulty, to render at least the more common diacritics and special characters in the major word-processors. Within Eurotyp, problems arise above all when files containing data in 'non-EC languages' are exchanged between different participants, or when other software (such as data-base programs) are used which does not provide for special characters. The following solutions are recommended:

1. If possible, the persons involved should agree on one word-processing system.

2. As a second alternative, there should be an agreement on a simple way of representing characters that are not in the 7-bit Ascii character set. This ensures that data can be transferred without loss of information between virtually any two computers and by virtually any channel of transmission, and that it can be entered using any existing text editor. Below, an example of such a standard is given; it has been used in Eurotyp Theme Group 6 and seems to work fairly well for most purposes.

A general principle is also to choose representations which do not contain more 'fancy' features than necessary. Standard orthography should be used when possible; languages which use non-Roman writing systems should be transliterated rather than transcribed phonetically.

7.3. Coding diacritics and special characters

Diacritics and special characters not found in the standard or extended ascii character set should be rendered as in Table 1. *x* stands for any character. Example: *šāh* should be entered as s\a5h. Other special characters are given in Table 2.

Notice the following: The Ascii code of # is 35, that of $ is 36. If you are using a 7-byte system, please check that you are using the correct characters, since some national standards have other characters in those places. Avoid the national accented characters if you are not certain that the end-user can identify the standard you have been using.

Table 1.Codes for diacritics

ẋ	x1
́x	x2
̂x	x3
̃x	x4
̄x	x5
̌x	x6
̇x	x7
̈x	x8
̲x	x9
̊x	x0

Table 2.Codes for other special characters

ð	d#
þ	t#
ʔ	?#
ɔ	c#
ə	e#
ɬ	l#
ŋ	n#

x̰ x$

x̥ x%

x̋ x&

x̯ x*

x̆ x\

Reference

Sheldon, Kenneth M. (1991). "Ascii goes global." *Byte*, July 1991, 108-116.

Back to index

Note: if you cannot see the symbols download Arial Unicode MS place it in your font folder (please note that the file is BIG > 15 MB)