

Unicode 1.0.1

1. Introduction

As discussed in Volumes 1 and 2, small changes have been made to Unicode 1.0 in order to incorporate it into the international character encoding standard, ISO 10646, which was approved by ISO as an International Standard in June, 1992. The Unicode Consortium plans to issue Unicode 1.1 in early 1993. The character content and encoding will be identical to that of ISO 10646. To that end, Unicode 1.1 will include approximately 5,400 additional characters from ISO 10646 that are not already in Unicode 1.0.

In order to expedite use of Unicode in the interim, the Unicode Consortium is issuing an intermediate version, Unicode 1.0.1, which consists of Unicode 1.0 modified by the changes necessary to make the character codes a proper subset of ISO 10646.

This paper describes the differences between Unicode 1.0.1 and Unicode 1.0 (for more information, see Volume 1, pp. xix-xx and Volume 2, pp. 4-9 and 427-431). Implementations that use Unicode 1.0.1 as thus defined will be completely compatible with Unicode 1.1, and therefore fully compatible with ISO 10646.

Mapping of Unicode characters to the national and industry standards will be finalized in Unicode 1.1 to reflect comments from reviewers and alignment with ISO 10646. In early 1993 a technical report will be issued that defines the content of Unicode 1.1, including the complete revised mapping tables. The mapping tables will be available in soft form by anonymous FTP. The technical report will be sent to members of the Unicode Consortium (inc. associates & individuals); others may obtain copies or information about FTP by contacting:

The Unicode Consortium
1965 Charleston Road
Mountain View, California 94043 USA

Phone: (415) 961-4189
Fax: (415) 966-1637
E-mail: unicode-inc@hq.metaphor.com

2. Final Zone Allocations

The following zone reallocations do not affect any allocated Unicode 1.0 characters.

A. Unicode Allocation

<u>Range</u>	<u>Cells</u>	<u>Name/Contents</u>
U+0000 → U+4DFF	19,968	A-ZONE Alphabets, syllabaries, symbols (the 65 control codes are excluded)
U+4E00 → U+9FFF	20,992	I-ZONE Ideographs
U+A000 → U+DFFF	16,384	O-ZONE Reserved for future assignment
U+E000 → U+FFFF	8,192	R-ZONE Restricted use (FFFE & FFFF are excluded)

B. R-ZONE Allocation

<u>Range</u>	<u>Cells</u>	<u>Name/Contents</u>
U+E000 → U+F8FF	6,400	Private Use Area (Corporate Use starts at F8FF)
U+F900 → U+FFEF	1,776	Compatibility Zone (including presentation forms)
U+FFFO → U+FFFF	16	Specials (FFFE & FFFF are not character codes, and are excluded)

3. Characters deleted or withdrawn for further study:

A. Groups of characters deleted

<u>Range</u>	<u>Group Name</u>
U+0E70 → U+0E74	Thai Phonetic Order Vowel signs
U+0EF0 → U+0EF4	Lao Phonetic Order Vowel signs
U+1000 → U+104C	Tibetan script

B. Individual characters deleted

U+03DB ζ	GREEK SMALL LETTER STIGMA	U+03E1 ξ	GREEK SMALL LETTER SAMPI
U+03DD \digamma	GREEK SMALL LETTER DIGAMMA	U+2300 Ⓐ	APL COMPOSE
U+03DF Ϝ	GREEK SMALL LETTER KOPPA	U+2301 Ⓐ	APL OUT

4. Characters unified

<u>From</u>	<u>With</u>	<u>Image</u>	<u>Old Name</u>
U+0371	U+0314	◊	GREEK NON-SPACING DASIA PNEUMATA
U+0372	U+0313	◊	GREEK NON-SPACING PSILI PNEUMATA
U+0384	U+030D	◊	GREEK NON-SPACING TONOS
U+04C5	U+049A	К	CYRILLIC CAPITAL LETTER KA OGONEK
U+04C6	U+049B	к	CYRILLIC SMALL LETTER KA OGONEK
U+04C9	U+04B2	Х	CYRILLIC CAPITAL LETTER KHA OGONEK
U+04CA	U+04B3	х	CYRILLIC SMALL LETTER KHA OGONEK
U+3004	U+4EDD	全	IDEOGRAPHIC DITTO MARK

5. Characters moved

<u>From</u>	<u>To</u>	<u>Image</u>	<u>Old Name</u>
U+0370	U+0345	◊	GREEK NON-SPACING IOTA BELOW
U+0385	U+0344	◊	GREEK NON-SPACING DIAERESIS TONOS
U+03D7	U+037E	;	GREEK QUESTION MARK
U+03D8	U+0374	’	GREEK UPPER NUMERAL SIGN
U+03D9	U+0375	’	GREEK LOWER NUMERAL SIGN
U+03F3	U+0384	◊	GREEK SPACING TONOS
U+03F4	U+0385	◊	GREEK SPACING DIAERESIS TONOS
U+03F5	U+037A	◊	GREEK SPACING IOTA BELOW
U+05F5	U+FB1E	◊	HEBREW POINT VARIKA
U+32FF	U+3004	全	JAPANESE INDUSTRIAL STANDARD SYMBOL



6. Character blocks rearranged

The explicit list will be in Unicode 1.1.

<u>Range</u>	<u>Group Name</u>
U+32D0 → U+32FE	Circled Katakana: <i>The 1.1 characters will be arranged in modern order: e.g., A, I, U, E, O, KA, KI, ...</i>
U+FE80 → U+FEFC	Basic glyphs for Arabic language: <i>The 1.1 character shapes will be arranged in different order: Isolate, Final, Initial, Medial</i>

7. Character semantics changed

A. Zero Width Joining

U+200C		ZERO WIDTH NON-JOINER
U+200D		ZERO WIDTH JOINER

In the merger with ISO 10646, the semantics of these two characters have been given a narrow interpretation. This brings added precision to the explanation given in Volume 1, page 77.

The intent of these characters is to address cursive graphical connection between the glyphs of a script, e.g. in scripts like Arabic whose printed form emulates handwriting. NON-JOINER and JOINER are best thought of as behaving like tiny letters that neighboring glyphs may connect to (JOINER) or avoid connecting to (NON-JOINER). They are thus processed as ordinary cursive letters rather than as control characters.

NON-JOINER and JOINER affect how the two neighboring glyphs connect to *them*, not to *each other*. As such, they have no direct relationship with ligature formation; in particular, JOINER does not in any way request that its two neighbors be ligatured to each other. Indeed, both NON-JOINER and JOINER may break up ligatures by interrupting the character sequence required to form the ligature.







The precise relationship between cursive appearance and ligatured appearance may differ from script to script, and therefore the precise usage of these characters is script-dependent. In the case of Latin typography, cur-

siveness (handwriting, emulation) and ligaturing are independent. Thus the text on Volume 1, page 77, may be clarified as follows:

f + JOINER + i will not form the ligature fi. Instead, if cursive versions of the f and i are available in the font, each will independently connect to the JOINER on the appropriate side (having the same appearance as f + i).

Usage of optional ligatures such as fi is not controlled by any codes within the Unicode standard, but is determined by protocols or resources external to the text sequence.


As further illustration, let a hyphen stand for a cursive connection to a preceding or following letter. Then in a cursive Latin font we would get the following results:

<u>Unicode</u>	<u>Rendering</u>
f i s h	f- -i- -s- -h (optionally using a ligature: fi- -s- -h)
f  i s h	f- -i- -s- -h
f  i s h	f i- -s- -h
f   i s h	f- i- -s- -h
f   i s h	f -i- -s- -h

With regard to the Arabic script, the statements in Volume 1, page 77, remain correct. In Volume 2, page 390, Arabic rules L2 and L3, the JOINER can be used to get the appearance in parentheses.

With regard to conjuncts in Indic scripts, the statements in Volume 1, pp. 53-56, and Volume 2, pp. 399-414, remain correct. However for clarity, in pp. 399-414 the term *ligature* should be replaced by the term *conjunct*.

B. Byte Order Mark














U+FEFF  ZERO WIDTH NO-BREAK SPACE

In addition to the meaning of BYTE ORDER MARK, as defined in Volume 1 of the Unicode standard, the code value U+FEFF may now also be used as ZERO WIDTH NO-BREAK SPACE (ZWNBS). For convenience in discussion, it can also be referred to by this name (which is the ISO 10646/Unicode 1.1 name for U+FEFF).


ZWNBS behaves like a U+00A0 NO-BREAK SPACE in that it indicates the absence of word boundaries; however, ZWNBS has no width. For example, this character can be inserted after the fourth character in the text "base+delta" to indicate that there should be no line break between the "e" and the "+" (for more information, see Volume 2, pp. 6-7).

8. Characters added

There are a large number of characters that will be added to Unicode 1.1 that will be included in the technical report, as explained above. These will include the following characters, which were omitted from Unicode 1.0.

U+0A4D 	GURMUKHI SIGN VIRAMA	U+FFE9 	HALFWIDTH LEFTWARDS ARROW
U+0A8D 	GUJARATI VOWEL CANDRA E	U+FFEA 	HALFWIDTH UPWARDS ARROW
U+0A91 	GUJARATI VOWEL CANDRA O	U+FFEB 	HALFWIDTH RIGHTWARDS ARROW
U+0AC9 	GUJARATI VOWEL SIGN CANDRA O	U+FFEC 	HALFWIDTH DOWNWARDS ARROW
U+0B56 	ORIYA AI LENGTH MARK	U+FFED 	HALFWIDTH BLACK SQUARE
U+25EF 	LARGE CIRCLE	U+FFEE 	HALFWIDTH WHITE CIRCLE
U+FFE8 	HALFWIDTH FORMS LIGHT VERTICAL		

9. Character mapping changed

<u>From</u>	<u>To Image</u>	<u>XJIS Name</u>
U+00AD	U+2010 -	815D JIS HYPHEN
U+20DD	U+25EF 	81FC JIS COMPOSITION CIRCLE