4.0 Character Properties

The chapter on Character Properties discusses in detail the attributes of several character types and how they are dealt with in the Unicode encoding scheme.

Disclaimer

The content of all character property tables has been verified as far as possible by the Unicode Consortium. However, the Unicode Consortium does not guarantee that the tables are correct in every detail. The character property tables are provided for informational purposes only. The Unicode Consortium is not responsible for errors that may occur either in the character property tables or in software which implements those tables as they are printed in this volume.

4.1 Numeric

Numeric is a general classification of characters that represent numbers. This includes characters such as fractions, subscripts, superscripts, Roman numerals, currency numerators, encircled numbers, and script-specific digits. In many traditional numbering systems, letters are used with a numeric value. Examples include Greek and Hebrew letters, and Latin letters used in outlines (II.A.1.b). These are special cases, and are not included here as numeric.

Digits form a large subcategory of numerics consisting of those numerics which can combine in sequence to form numbers. This includes characters such as subscripts, superscripts, Roman numerals, and script-specific digits. Digits do not include characters such as encircled numbers or fractions.

Decimal digits form a large subcategory of digits consisting of those digits which can be used to form decimal-radix numbers. This includes characters such as subscripts, superscripts, and script-specific digits. Decimal digits do not include characters such as Roman numerals (1 + 5 = 15 = fifteen, but I + V = IV = four).

The Unicode standard assigns distinct codes to native forms of digits, specific to a given script or language. Examples are the digits used with the Arabic script, Chinese numbers or those of the Indic languages. (For naming conventions, see the introduction to the Arabic block.) An alternative would have been to provide codes only for the European digits (ASCII 0–9) and make the the other forms presentation variants only. The latter choice favors applications which focus on number parsing over word processing or display oriented applications. To help programs reduce the complexity of their parsers, the Unicode standard provides mapping tables which can be used to fold any of its digits into the ASCII range.

Decimal Digits

Decimal digits are digits which can be concatenated to build decimal numbers. This list is a proper subset of the *Digits* list.

UNIC	*Unicode character name*	*Value*
0030	DIGIT ZERO	0
0031	DIGIT ONE	1
0032	DIGIT TWO	2
0033	DIGIT THREE	3
0034	DIGIT FOUR	4
0035	DIGIT FIVE	5
0036	DIGIT SIX	6
0037	DIGIT SEVEN	7
0038	DIGIT EIGHT	8
0039	DIGIT NINE	9
00B2	SUPERSCRIPT DIGIT TWO	2
00B3	SUPERSCRIPT DIGIT THREE	3
00B9	SUPERSCRIPT DIGIT ONE	1
0660	ARABIC-INDIC DIGIT ZERO	0
0661	ARABIC-INDIC DIGIT ONE	1
0662	ARABIC-INDIC DIGIT TWO	2
0663	ARABIC-INDIC DIGIT THREE	3
0664	ARABIC-INDIC DIGIT FOUR	4
0665	ARABIC-INDIC DIGIT FIVE	5
0666	ARABIC-INDIC DIGIT SIX	6
0667	ARABIC-INDIC DIGIT SEVEN	7
0668	ARABIC-INDIC DIGIT EIGHT	8
0669	ARABIC-INDIC DIGIT NINE	9
06F0	EASTERN ARABIC-INDIC DIGIT ZERO	0
06F1	EASTERN ARABIC-INDIC DIGIT ONE	1
06F2	EASTERN ARABIC-INDIC DIGIT TWO	2
06F3	EASTERN ARABIC-INDIC DIGIT THREE	3
06F4	EASTERN ARABIC-INDIC DIGIT FOUR	4
06F5	EASTERN ARABIC-INDIC DIGIT FIVE	5
06F6	EASTERN ARABIC-INDIC DIGIT SIX	6
06F7	EASTERN ARABIC-INDIC DIGIT SEVEN	7
06F8	EASTERN ARABIC-INDIC DIGIT EIGHT	8
06F9	EASTERN ARABIC-INDIC DIGIT NINE	9
0966	DEVANAGARI DIGIT ZERO	0
0967	DEVANAGARI DIGIT ONE	1
0968	DEVANAGARI DIGIT TWO	2
0969	DEVANAGARI DIGIT THREE	3
096A	DEVANAGARI DIGIT FOUR	4
096B	DEVANAGARI DIGIT FIVE	5
096C	DEVANAGARI DIGIT SIX	6
096D	DEVANAGARI DIGIT SEVEN	7
096E	DEVANAGARI DIGIT EIGHT	8
096F	DEVANAGARI DIGIT NINE	9
09E6	BENGALI DIGIT ZERO	0
09E7	BENGALI DIGIT ONE	1
09E8	BENGALI DIGIT TWO	2
09E9	BENGALI DIGIT THREE	3
09EA	BENGALI DIGIT FOUR	4
09EB	BENGALI DIGIT FIVE	5
09EC	BENGALI DIGIT SIX	6

UNIC	Unicode character name	Value
09ED	BENGALI DIGIT SEVEN	7
09EE	BENGALI DIGIT EIGHT	8
09EF	BENGALI DIGIT NINE	9
0A66	GURMUKHI DIGIT ZERO	0
0A67	GURMUKHI DIGIT ONE	1
0A68	GURMUKHI DIGIT TWO	2
0A69	GURMUKHI DIGIT THREE	3
0A6A	GURMUKHI DIGIT FOUR	4
0A6B	GURMUKHI DIGIT FIVE	5
0A6C	GURMUKHI DIGIT SIX	6
0A6D	GURMUKHI DIGIT SEVEN	7
0A6E	GURMUKHI DIGIT EIGHT	8
0A6F	GURMUKHI DIGIT NINE	9
0AE6	GUJARATI DIGIT ZERO	0
0AE7	GUJARATI DIGIT ONE	1
0AE8	GUJARATI DIGIT TWO	2
0AE9	GUJARATI DIGIT THREE	3
0AEA	GUJARATI DIGIT FOUR	4
0AEB	GUJARATI DIGIT FIVE	5
0AEC	GUJARATI DIGIT SIX	6
0AED	GUJARATI DIGIT SEVEN	7
0AEE	GUJARATI DIGIT EIGHT	8
0AEF	GUJARATI DIGIT NINE	9
0B66	ORIYA DIGIT ZERO	0
0B67	ORIYA DIGIT ONE	1
0B68	ORIYA DIGIT TWO	2
0B69	ORIYA DIGIT THREE	3
0B6A	ORIYA DIGIT FOUR	4
0B6B	ORIYA DIGIT FIVE	5
0B6C	ORIYA DIGIT SIX	6
0B6D	ORIYA DIGIT SEVEN	7
0B6E	ORIYA DIGIT EIGHT	8
0B6F	ORIYA DIGIT NINE	9
0BE7	TAMIL DIGIT ONE	1
0BE8	TAMIL DIGIT TWO	2
0BE9	TAMIL DIGIT THREE	3
0BEA	TAMIL DIGIT FOUR	4
0BEB	TAMIL DIGIT FIVE	5
0BEC	TAMIL DIGIT SIX	6
0BED	TAMIL DIGIT SEVEN	7
0BEE	TAMIL DIGIT EIGHT	8
0BEF	TAMIL DIGIT NINE	9
0C66	TELUGU DIGIT ZERO	0
0C67	TELUGU DIGIT ONE	1
0C68	TELUGU DIGIT TWO	2
0C69	TELUGU DIGIT THREE	3
0C6A	TELUGU DIGIT FOUR	4
0C6B	TELUGU DIGIT FIVE	5
0C6C	TELUGU DIGIT SIX	6
0C6D	TELUGU DIGIT SEVEN	7
0C6E	TELUGU DIGIT EIGHT	8
0C6F	TELUGU DIGIT NINE	9
0CE6	KANNADA DIGIT ZERO	0
0CE7	KANNADA DIGIT ONE	1
0CE8	KANNADA DIGIT TWO	2
0CE9	KANNADA DIGIT THREE	3

UNIC	Unicode character name	Value
0CEA	KANNADA DIGIT FOUR	4
0CEB	KANNADA DIGIT FIVE	5
0CEC	KANNADA DIGIT SIX	6
0CED	KANNADA DIGIT SEVEN	7
0CEE	KANNADA DIGIT EIGHT	8
0CEF	KANNADA DIGIT NINE	9
0D66	MALAYALAM DIGIT ZERO	0
0D67	MALAYALAM DIGIT ONE	1
0D68	MALAYALAM DIGIT TWO	2
0D69	MALAYALAM DIGIT THREE	3
0D6A	MALAYALAM DIGIT FOUR	4
0D6B	MALAYALAM DIGIT FIVE	5
0D6C	MALAYALAM DIGIT SIX	6
0D6D	MALAYALAM DIGIT SEVEN	7
0D6E	MALAYALAM DIGIT EIGHT	8
0D6F	MALAYALAM DIGIT NINE	9
0E50	THAI DIGIT ZERO	0
0E51	THAI DIGIT ONE	1
0E52	THAI DIGIT TWO	2
0E53	THAI DIGIT THREE	3
0E54	THAI DIGIT FOUR	4
0E55	THAI DIGIT FIVE	5
0E56	THAI DIGIT SIX	6
0E57	THAI DIGIT SEVEN	7
0E58	THAI DIGIT EIGHT	8
0E59	THAI DIGIT NINE	9
0ED0	LAO DIGIT ZERO	0
0ED1	LAO DIGIT ONE	1
0ED2	LAO DIGIT TWO	2
0ED3	LAO DIGIT THREE	3
0ED4	LAO DIGIT FOUR	4
0ED5	LAO DIGIT FIVE	5
0ED6	LAO DIGIT SIX	6
0ED7	LAO DIGIT SEVEN	7
0ED8	LAO DIGIT EIGHT	8
0ED9	LAO DIGIT NINE	9
1040	TIBETAN DIGIT ZERO	0
1041	TIBETAN DIGIT ONE	1
1042	TIBETAN DIGIT TWO	2
1043	TIBETAN DIGIT THREE	3
1044	TIBETAN DIGIT FOUR	4
1045	TIBETAN DIGIT FIVE	5
1046	TIBETAN DIGIT SIX	6
1047	TIBETAN DIGIT SEVEN	7
1048	TIBETAN DIGIT EIGHT	8
1049	TIBETAN DIGIT NINE	9
2070	SUPERSCRIPT DIGIT ZERO	0
2074	SUPERSCRIPT DIGIT FOUR	4
2075	SUPERSCRIPT DIGIT FIVE	5
2076	SUPERSCRIPT DIGIT SIX	6
2077	SUPERSCRIPT DIGIT SEVEN	7
2078	SUPERSCRIPT DIGIT EIGHT	8
2079	SUPERSCRIPT DIGIT NINE	9
2080	SUBSCRIPT DIGIT ZERO	0
2081	SUBSCRIPT DIGIT ONE	1
2082	SUBSCRIPT DIGIT TWO	2

UNIC	Unicode character name	Value
2083	SUBSCRIPT DIGIT THREE	3
2084	SUBSCRIPT DIGIT FOUR	4
2085	SUBSCRIPT DIGIT FIVE	5
2086	SUBSCRIPT DIGIT SIX	6
2087	SUBSCRIPT DIGIT SEVEN	7
2088	SUBSCRIPT DIGIT EIGHT	8
2089	SUBSCRIPT DIGIT NINE	9

Digits

Digits include all of the characters labeled in the *Decimal Digits* list, plus the following characters, which cannot be concatenated to form decimal numbers. This list is a proper subset of the *Numbers* list.

UNIC	*Unicode character name*	*Value*
2460	CIRCLED DIGIT ONE	1
2461	CIRCLED DIGIT TWO	2
2462	CIRCLED DIGIT THREE	3
2463	CIRCLED DIGIT FOUR	4
2464	CIRCLED DIGIT FIVE	5
2465	CIRCLED DIGIT SIX	6
2466	CIRCLED DIGIT SEVEN	7
2467	CIRCLED DIGIT EIGHT	8
2468	CIRCLED DIGIT NINE	9
2474	PARENTHESIZED DIGIT ONE	1
2475	PARENTHESIZED DIGIT TWO	2
2476	PARENTHESIZED DIGIT THREE	3
2477	PARENTHESIZED DIGIT FOUR	4
2478	PARENTHESIZED DIGIT FIVE	5
2479	PARENTHESIZED DIGIT SIX	6
247A	PARENTHESIZED DIGIT SEVEN	7
247B	PARENTHESIZED DIGIT EIGHT	8
247C	PARENTHESIZED DIGIT NINE	9
2488	DIGIT ONE PERIOD	1
2489	DIGIT TWO PERIOD	2
248A	DIGIT THREE PERIOD	3
248B	DIGIT FOUR PERIOD	4
248C	DIGIT FIVE PERIOD	5
248D	DIGIT SIX PERIOD	6
248E	DIGIT SEVEN PERIOD	7
248F	DIGIT EIGHT PERIOD	8
2490	DIGIT NINE PERIOD	9
24EA	CIRCLED DIGIT ZERO	0
2776	INVERSE CIRCLED DIGIT ONE	1
2777	INVERSE CIRCLED DIGIT TWO	2
2778	INVERSE CIRCLED DIGIT THREE	3
2779	INVERSE CIRCLED DIGIT FOUR	4
277A	INVERSE CIRCLED DIGIT FIVE	5
277B	INVERSE CIRCLED DIGIT SIX	6
277C	INVERSE CIRCLED DIGIT SEVEN	7
277D	INVERSE CIRCLED DIGIT EIGHT	8
277E	INVERSE CIRCLED DIGIT NINE	9
2780	CIRCLED SANS-SERIF DIGIT ONE	1
2781	CIRCLED SANS-SERIF DIGIT TWO	2
2782	CIRCLED SANS-SERIF DIGIT THREE	3
2783	CIRCLED SANS-SERIF DIGIT FOUR	4
2784	CIRCLED SANS-SERIF DIGIT FIVE	5
2785	CIRCLED SANS-SERIF DIGIT SIX	6
2786	CIRCLED SANS-SERIF DIGIT SEVEN	7
2787	CIRCLED SANS-SERIF DIGIT EIGHT	8
2788	CIRCLED SANS-SERIF DIGIT NINE	9
278A	INVERSE CIRCLED SANS-SERIF DIGIT ONE	1
278B	INVERSE CIRCLED SANS-SERIF DIGIT TWO	2

UNIC	*Unicode character name*	*Value*
278C	INVERSE CIRCLED SANS-SERIF DIGIT THREE	3
278D	INVERSE CIRCLED SANS-SERIF DIGIT FOUR	4
278E	INVERSE CIRCLED SANS-SERIF DIGIT FIVE	5
278F	INVERSE CIRCLED SANS-SERIF DIGIT SIX	6
2790	INVERSE CIRCLED SANS-SERIF DIGIT SEVEN	7
2791	INVERSE CIRCLED SANS-SERIF DIGIT EIGHT	8
2792	INVERSE CIRCLED SANS-SERIF DIGIT NINE	9

Numbers

This list includes all of the *Digits* list, plus other characters which can be interpreted as having a numerical value associated with them. All of the Roman numerals are listed here, although some of them can be considered *Digits* within the Roman numeration scheme.

UNIC	*Unicode character name*	*Value*
00BC	FRACTION ONE QUARTER	1/4
00BD	FRACTION ONE HALF	1/2
00BE	FRACTION THREE QUARTERS	3/4
09F4	BENGALI CURRENCY NUMERATOR ONE	1
09F5	BENGALI CURRENCY NUMERATOR TWO	2
09F6	BENGALI CURRENCY NUMERATOR THREE	3
09F7	BENGALI CURRENCY NUMERATOR FOUR	4
09F8	BENGALI CURRENCY NUMERATOR ONE LESS THAN THE DENOMINATOR	—
09F9	BENGALI CURRENCY DENOMINATOR SIXTEEN	16
0BF0	TAMIL NUMBER TEN	10
0BF1	TAMIL NUMBER ONE HUNDRED	100
0BF2	TAMIL NUMBER ONE THOUSAND	1000
2153	FRACTION ONE THIRD	1/3
2154	FRACTION TWO THIRDS	2/3
2155	FRACTION ONE FIFTH	1/5
2156	FRACTION TWO FIFTHS	2/5
2157	FRACTION THREE FIFTHS	3/5
2158	FRACTION FOUR FIFTHS	4/5
2159	FRACTION ONE SIXTH	1/6
215A	FRACTION FIVE SIXTHS	5/6
215B	FRACTION ONE EIGHTH	1/8
215C	FRACTION THREE EIGHTHS	3/8
215D	FRACTION FIVE EIGHTHS	5/8
215E	FRACTION SEVEN EIGHTHS	7/8
215F	FRACTION NUMERATOR ONE	1
2160	ROMAN NUMERAL ONE	1
2161	ROMAN NUMERAL TWO	2
2162	ROMAN NUMERAL THREE	3
2163	ROMAN NUMERAL FOUR	4
2164	ROMAN NUMERAL FIVE	5
2165	ROMAN NUMERAL SIX	6
2166	ROMAN NUMERAL SEVEN	7
2167	ROMAN NUMERAL EIGHT	8
2168	ROMAN NUMERAL NINE	9
2169	ROMAN NUMERAL TEN	10
216A	ROMAN NUMERAL ELEVEN	11
216B	ROMAN NUMERAL TWELVE	12
216C	ROMAN NUMERAL FIFTY	50
216D	ROMAN NUMERAL ONE HUNDRED	100
216E	ROMAN NUMERAL FIVE HUNDRED	500
216F	ROMAN NUMERAL ONE THOUSAND	1000
2170	SMALL ROMAN NUMERAL ONE	1
2171	SMALL ROMAN NUMERAL TWO	2
2172	SMALL ROMAN NUMERAL THREE	3
2173	SMALL ROMAN NUMERAL FOUR	4
2174	SMALL ROMAN NUMERAL FIVE	5
2175	SMALL ROMAN NUMERAL SIX	6

UNIC	Unicode character name	Value
2176	SMALL ROMAN NUMERAL SEVEN	7
2177	SMALL ROMAN NUMERAL EIGHT	8
2178	SMALL ROMAN NUMERAL NINE	9
2179	SMALL ROMAN NUMERAL TEN	10
217A	SMALL ROMAN NUMERAL ELEVEN	11
217B	SMALL ROMAN NUMERAL TWELVE	12
217C	SMALL ROMAN NUMERAL FIFTY	50
217D	SMALL ROMAN NUMERAL ONE HUNDRED	100
217E	SMALL ROMAN NUMERAL FIVE HUNDRED	500
217F	SMALL ROMAN NUMERAL ONE THOUSAND	1000
2180	ROMAN NUMERAL ONE THOUSAND C D	1000
2181	ROMAN NUMERAL FIVE THOUSAND	5000
2182	ROMAN NUMERAL TEN THOUSAND	10000
2469	CIRCLED NUMBER TEN	10
246A	CIRCLED NUMBER ELEVEN	11
246B	CIRCLED NUMBER TWELVE	12
246C	CIRCLED NUMBER THIRTEEN	13
246D	CIRCLED NUMBER FOURTEEN	14
246E	CIRCLED NUMBER FIFTEEN	15
246F	CIRCLED NUMBER SIXTEEN	16
2470	CIRCLED NUMBER SEVENTEEN	17
2471	CIRCLED NUMBER EIGHTEEN	18
2472	CIRCLED NUMBER NINETEEN	19
2473	CIRCLED NUMBER TWENTY	20
247D	PARENTHESIZED NUMBER TEN	10
247E	PARENTHESIZED NUMBER ELEVEN	11
247F	PARENTHESIZED NUMBER TWELVE	12
2480	PARENTHESIZED NUMBER THIRTEEN	13
2481	PARENTHESIZED NUMBER FOURTEEN	14
2482	PARENTHESIZED NUMBER FIFTEEN	15
2483	PARENTHESIZED NUMBER SIXTEEN	16
2484	PARENTHESIZED NUMBER SEVENTEEN	17
2485	PARENTHESIZED NUMBER EIGHTEEN	18
2486	PARENTHESIZED NUMBER NINETEEN	19
2487	PARENTHESIZED NUMBER TWENTY	20
2491	NUMBER TEN PERIOD	10
2492	NUMBER ELEVEN PERIOD	11
2493	NUMBER TWELVE PERIOD	12
2494	NUMBER THIRTEEN PERIOD	13
2495	NUMBER FOURTEEN PERIOD	14
2496	NUMBER FIFTEEN PERIOD	15
2497	NUMBER SIXTEEN PERIOD	16
2498	NUMBER SEVENTEEN PERIOD	17
2499	NUMBER EIGHTEEN PERIOD	18
249A	NUMBER NINETEEN PERIOD	19
249B	NUMBER TWENTY PERIOD	20
277F	INVERSE CIRCLED NUMBER 10	10
2789	CIRCLED SANS-SERIF NUMBER 10	10
2793	INVERSE CIRCLED SANS-SERIF NUMBER 10	10
3007	IDEOGRAPHIC NUMBER ZERO	0
3021	HANGZHOU NUMERAL ONE	1
3022	HANGZHOU NUMERAL TWO	2
3023	HANGZHOU NUMERAL THREE	3
3024	HANGZHOU NUMERAL FOUR	4
3025	HANGZHOU NUMERAL FIVE	5
3026	HANGZHOU NUMERAL SIX	6

UNIC	Unicode character name	Value
3027	HANGZHOU NUMERAL SEVEN	7
3028	HANGZHOU NUMERAL EIGHT	8
3029	HANGZHOU NUMERAL NINE	9
3280	CIRCLED IDEOGRAPH ONE	1
3281	CIRCLED IDEOGRAPH TWO	2
3282	CIRCLED IDEOGRAPH THREE	3
3283	CIRCLED IDEOGRAPH FOUR	4
3284	CIRCLED IDEOGRAPH FIVE	5
3285	CIRCLED IDEOGRAPH SIX	6
3286	CIRCLED IDEOGRAPH SEVEN	7
3287	CIRCLED IDEOGRAPH EIGHT	8
3288	CIRCLED IDEOGRAPH NINE	9
3289	CIRCLED IDEOGRAPH TEN	10

4.2 Space Characters

Eight-bit character sets contain two space characters U+0020 SPACE and U+00A0 NON-BREAKING SPACE. The Unicode standard has several additional space characters which provide explicit control over their width (from zero-width or non-printing, on upward). U+2007 FIGURE SPACE is intended to be used as a thousands separator in those countries that use a space to separate groups of digits. It behaves like a numeric separator for the purposes of bidirectional layout (See Appendix A for a detailed discussion of bidirectional coding.)

Note that not all space characters have word- or line-breaking properties.

Space characters include:

U+0020	SPACE
U+00A0	NON-BREAKING SPACE
U+2000	EN QUAD
U+2001	EM QUAD
U+2002	EN SPACE
U+2003	EM SPACE
U+2004	THREE-PER-EM SPACE
U+2005	FOUR-PER-EM SPACE
U+2006	SIX-PER-EM SPACE
U+2007	FIGURE SPACE
U+2008	PUNCTUATION SPACE
U+2009	THIN SPACE
U+200A	HAIR SPACE
U+200B	ZERO WIDTH SPACE
U+3000	IDEOGRAPHIC SPACE

4.3 Dashes

In addition to spaces, the Unicode standard encodes several dashes. Here the semantics of the ASCII *hyphen-minus* (U+002D) is ambiguous. The Unicode standard provides two explicit codes, *hyphen* and *minus* for those applications that need to distinguish these two. Dashes of various length are provided as well. In a few cases, the Unicode standard makes a distinction purely on the basis of the intended semantics without a corresponding visual difference. For example, typographers typically use the *en-dash* to typeset the *minus*, but the Unicode character encoding has two different codes, so that it is possible to distinguish which one has the numeric quality.

Dash characters include

U+002D	HYPHEN-MINUS
U+2010	HYPHEN
U+2011	NON-BREAKING HYPHEN
U+2012	FIGURE DASH
U+2013	EN DASH
U+2014	EM DASH
U+2015	QUOTATION DASH
U+207B	SUPERSCRIPT HYPHEN-MINUS
U+208B	SUBSCRIPT HYPHEN-MINUS
U+2212	MINUS
U+301C	WAVE DASH
U+3030	WAVY DASH

4.4 Line Breaking

Rules of line breaking differ substantially from script to script and language to language. The rules for determining correct line break can be quite complex (especially when hyphenation is included) and are beyond the scope of the Unicode standard.

However, there are certain characters with distinguished semantics vis-a-vis line break. Certain characters are word delimiters, and always allow line break. These include all spaces except U+00A0 NON-BREAKING SPACE and U+2007 FIGURE SPACE.

Certain other characters generally disallow word-breaking on either side, including U+00A0 NON-BREAKING SPACE and U+2011 NON-BREAKING HYPHEN. These characters are included for compatibility (proper control of line break cannot be accomplished by simply cloning a small number of characters).

4.5 Non-spacing Marks

When rendered, the non-spacing marks are attached to the preceding base character in some manner, and do not to occupy a spacing position by themselves.

All of the characters in the range U+0300 → U+0348, U+20D0 → U+20E1, and U+302A → U+302F are non-spacing marks. In addition, the following characters are also non-spacing marks:

UNIC	*Unicode character name*
0370	GREEK NON-SPACING IOTA BELOW
0371	GREEK NON-SPACING DASIA PNEUMATA
0372	GREEK NON-SPACING PSILI PNEUMATA
0384	GREEK NON-SPACING TONOS
0385	GREEK NON-SPACING DIAERESIS TONOS
0483	CYRILLIC NON-SPACING TITLO
0484	CYRILLIC NON-SPACING PALATALIZATION
0485	CYRILLIC NON-SPACING DASIA PNEUMATA
0486	CYRILLIC NON-SPACING PSILI PNEUMATA
05B0	HEBREW POINT SHEVA
05B1	HEBREW POINT HATAF SEGOL
05B2	HEBREW POINT HATAF PATAH
05B3	HEBREW POINT HATAF QAMATS
05B4	HEBREW POINT HIRIQ
05B5	HEBREW POINT TSERE
05B6	HEBREW POINT SEGOL
05B7	HEBREW POINT PATAH
05B8	HEBREW POINT QAMATS
05B9	HEBREW POINT HOLAM
05BB	HEBREW POINT QUBUTS
05BC	HEBREW POINT DAGESH
05BD	HEBREW POINT METEG
05BF	HEBREW POINT RAFE
05C1	HEBREW POINT SHIN DOT
05C2	HEBREW POINT SIN DOT
05F5	HEBREW POINT VARIKA
064B	ARABIC FATHATAN
064C	ARABIC DAMMATAN
064D	ARABIC KASRATAN
064E	ARABIC FATHAH
064F	ARABIC DAMMAH
0650	ARABIC KASRAH
0651	ARABIC SHADDAH
0652	ARABIC SUKUN
0670	ARABIC ALEF ABOVE
0901	DEVANAGARI SIGN CANDRABINDU
0902	DEVANAGARI SIGN ANUSVARA
093C	DEVANAGARI SIGN NUKTA
0941	DEVANAGARI VOWEL SIGN U
0942	DEVANAGARI VOWEL SIGN UU
0943	DEVANAGARI VOWEL SIGN VOCALIC R

UNIC	Unicode character name
0944	DEVANAGARI VOWEL SIGN VOCALIC RR
0945	DEVANAGARI VOWEL SIGN CANDRA E
0946	DEVANAGARI VOWEL SIGN SHORT E
0947	DEVANAGARI VOWEL SIGN E
0948	DEVANAGARI VOWEL SIGN AI
094D	DEVANAGARI SIGN VIRAMA
0951	DEVANAGARI STRESS SIGN UDATTA
0952	DEVANAGARI STRESS SIGN ANUDATTA
0953	DEVANAGARI GRAVE ACCENT
0954	DEVANAGARI ACUTE ACCENT
0962	DEVANAGARI VOWEL SIGN VOCALIC L
0963	DEVANAGARI VOWEL SIGN VOCALIC LL
0981	BENGALI SIGN CANDRABINDU
09BC	BENGALI SIGN NUKTA
09C1	BENGALI VOWEL SIGN U
09C2	BENGALI VOWEL SIGN UU
09C3	BENGALI VOWEL SIGN VOCALIC R
09C4	BENGALI VOWEL SIGN VOCALIC RR
09CD	BENGALI SIGN VIRAMA
09E2	BENGALI VOWEL SIGN VOCALIC L
09E3	BENGALI VOWEL SIGN VOCALIC LL
0A02	GURMUKHI SIGN BINDI
0A3C	GURMUKHI SIGN NUKTA
0A41	GURMUKHI VOWEL SIGN U
0A42	GURMUKHI VOWEL SIGN UU
0A47	GURMUKHI VOWEL SIGN EE
0A48	GURMUKHI VOWEL SIGN AI
0A4B	GURMUKHI VOWEL SIGN OO
0A4C	GURMUKHI VOWEL SIGN AU
0A70	GURMUKHI TIPPI
0A71	GURMUKHI ADDAK
0A81	GUJARATI SIGN CANDRABINDU
0A82	GUJARATI SIGN ANUSVARA
0ABC	GUJARATI SIGN NUKTA
0AC1	GUJARATI VOWEL SIGN U
0AC2	GUJARATI VOWEL SIGN UU
0AC3	GUJARATI VOWEL SIGN VOCALIC R
0AC4	GUJARATI VOWEL SIGN VOCALIC RR
0AC5	GUJARATI VOWEL SIGN CANDRA E
0AC7	GUJARATI VOWEL SIGN E
0AC8	GUJARATI VOWEL SIGN AI
0ACD	GUJARATI SIGN VIRAMA
0B01	ORIYA SIGN CANDRABINDU
0B3C	ORIYA SIGN NUKTA
0B3F	ORIYA VOWEL SIGN I
0B41	ORIYA VOWEL SIGN U
0B42	ORIYA VOWEL SIGN UU
0B43	ORIYA VOWEL SIGN VOCALIC R
0B4D	ORIYA SIGN VIRAMA
0BC0	TAMIL VOWEL SIGN II
0BCD	TAMIL SIGN VIRAMA
0C3E	TELUGU VOWEL SIGN AA
0C3F	TELUGU VOWEL SIGN I
0C40	TELUGU VOWEL SIGN II
0C46	TELUGU VOWEL SIGN E
0C47	TELUGU VOWEL SIGN EE

UNIC	Unicode character name
0C48	TELUGU VOWEL SIGN AI
0C4A	TELUGU VOWEL SIGN O
0C4B	TELUGU VOWEL SIGN OO
0C4C	TELUGU VOWEL SIGN AU
0C4D	TELUGU SIGN VIRAMA
0C55	TELUGU LENGTH MARK
0C56	TELUGU AI LENGTH MARK
0CBF	KANNADA VOWEL SIGN I
0CC6	KANNADA VOWEL SIGN E
0CCC	KANNADA VOWEL SIGN AU
0CCD	KANNADA SIGN VIRAMA
0D41	MALAYALAM VOWEL SIGN U
0D42	MALAYALAM VOWEL SIGN UU
0D43	MALAYALAM VOWEL SIGN VOCALIC R
0D4D	MALAYALAM SIGN VIRAMA
0E31	THAI VOWEL SIGN MAI HAN-AKAT
0E34	THAI VOWEL SIGN SARA I
0E35	THAI VOWEL SIGN SARA II
0E36	THAI VOWEL SIGN SARA UE
0E37	THAI VOWEL SIGN SARA UEE
0E38	THAI VOWEL SIGN SARA U
0E39	THAI VOWEL SIGN SARA UU
0E3A	THAI VOWEL SIGN PHINTHU
0E47	THAI VOWEL SIGN MAI TAI KHU
0E48	THAI TONE MAI EK
0E49	THAI TONE MAI THO
0E4A	THAI TONE MAI TRI
0E4B	THAI TONE MAI CHATTAWA
0E4C	THAI THANTHAKHAT
0E4D	THAI NIKKHAHIT
0EB1	LAO VOWEL SIGN MAI KAN
0EB4	LAO VOWEL SIGN I
0EB5	LAO VOWEL SIGN II
0EB6	LAO VOWEL SIGN Y
0EB7	LAO VOWEL SIGN YY
0EB8	LAO VOWEL SIGN U
0EB9	LAO VOWEL SIGN UU
0EBB	LAO VOWEL SIGN MAI KON
0EBC	LAO SEMIVOWEL SIGN LO
0EC8	LAO TONE MAI EK
0EC9	LAO TONE MAI THO
0ECA	LAO TONE MAI TI
0ECB	LAO TONE MAI CATAWA
0ECC	LAO CANCELLATION MARK
0ECD	LAO NIGGAHITA
1026	TIBETAN VOWEL SIGN I
1027	TIBETAN VOWEL SIGN SHORT I
1028	TIBETAN VOWEL SIGN U
1029	TIBETAN VOWEL SIGN E
102A	TIBETAN VOWEL SIGN O
102E	TIBETAN ANUSVARA
1030	TIBETAN UNDER RING
1036	TIBETAN CANDRABINDU
1037	TIBETAN CANDRABINDU WITH ORNAMENT
103B	TIBETAN HONORIFIC UNDER RING
103D	TIBETAN VOWEL SIGN AI

UNIC	*Unicode character name*
103E	TIBETAN VOWEL SIGN AU
104B	TIBETAN VIRAMA
104C	TIBETAN LENITION MARK
3099	NON-SPACING KATAKANA-HIRAGANA VOICED SOUND MARK
309A	NON-SPACING KATAKANA-HIRAGANA SEMI-VOICED SOUND MARK

4.6 Directional Character Types

All Unicode characters without exception are directional. The directional types left-to-right and right-to-left are called *strong types*, and characters of those types are called strong directional characters. In addition, the Bidirecional Algorithm uses *weak types* and *neutrals*. The table below shows these types.

Strong Types

L *Left-Right*	Left-to-right types include most alphabetic, syllabic, and Han ideographic characters.
Latin letters	U+0041 → U+005A, U+0061 → U+007A, U+00C0 → U+00D6, U+00D8 → U+00F6, U+00F8 → U+00FF
European Latin → Modifier Letters	U+0100 → U+02FF
General Diacriticals	U+0300 → U+036F
Greek → Armenian	U+0370 → U+058F
Devanagari → Georgian	U+0900 → U+10FF
Hiragana → Han	U+3040 → U+8BFF
Roman Numerals	U+2160 → U+2182
Left-Right Mark	U+200E
Symbol Diacritics	U+20D0 → U+20FF
Miscellaneous	U+0026, U+0040
R *Right-Left*	Right-to-left types include Arabic, Hebrew, and punctuation specific to those scripts.
Arabic and Hebrew	U+0590 → U+065F, U+066D → U+06EF
Right-Left Mark	U+200F

Weak Types

EN *European Number*	
European digits	U+0030 → U+0039
Eastern Arabic digits	U+06F0 → U+06F9
Super/Sub digits	U+2070, U+00B9, U+00B2 → U+00B3, U+2074 → U+2079, U+2080 → U+2089

ES European Number Separator

Figure Space	U+2007
Period	U+002E
Slash	U+002F

ET European Number Terminator

Plus sign	U+002B
Minus Sign	U+2212
Superscript plus and minus	U+207A, U+207B
Subscript plus and minus	U+208A, U+208B
Hyphen-Minus	U+002D
Plus-Minus	U+00B1
Minus-Plus	U+2213
Percents	U+0025, U+066A, U+2030, U+2031
Degree	U+00B0
Minute (Prime)	U+2032
Second (Double Prime)	U+2033
Currency symbols	U+00A2 → U+00A5, U+20A0 → U+20CF, U+0024
Number sign	U+0023

AN Arabic Number

Arabic-based digits	U+0660 → U+0669
Arabic decimal & thousands separators	U+066B, U+066C

CS Common Number Separator

Colon	U+003A
Comma	U+002C

Neutrals

B Block Separator

Paragraph separator (PS)	U+2029
Line separator (LS)	U+2028

S Segment Separator

(see below)

WS	*Whitespace*
Space	U+0020
NBSP	U+00A0
General Punctuation Spaces	U+2000 → U+2006, U+2008 → U+200B, U+3000

ON	*Other Neutrals*
All other characters	punctuation, symbols

As with other character properties, the Compatibility Zone characters have the same directional properties as their corresponding canonical characters. The directional type of all unassigned characters is not defined. This is also true of unassigned characters falling within the ranges used in the above tables; for brevity, not all unassigned characters in the ranges are called out separately where there are gaps. As unassigned characters are assigned in future versions of the Unicode standard, the new character properties will be documented.

Where unassigned characters are bidirectionally ordered for display (for example, as replacement glyphs), conformant processes are free to choose different directional properties. However, for best compatibility with future versions of the Unicode standard, it is recommended that unassigned characters be generally given the directional property *neutral* (N). The unassigned range U+0700 → U+08FF is reserved for use by future right-to-left scripts, however, so that a reasonable default in that case is the directional property *right-left* (R).

The definition of control code semantics is outside of the scope of the Unicode standard. Implementers should interpret the type of characters such as CR, LF, GS, and so on according the closest semantics to the types given here, such as interpreting CR (when used as paragraph separator) as being a block separator. *Horizontal tab* would generally be interpreted as a segment separator, which indicates that in a line containing tabs, the tab-delimited segments go in the base level direction (see the Bidirectional Algorithm, Appendix A).

Since horizontal Han ideographic characters are generally left-to-right, they have the *left-right* (L) directional character type. When they are written from right-to-left, their direction can overridden, as discussed in Bidirectional Algorithm, Appendix A.