Unicode Technical Report #14

Line Breaking Properties

Revision	5
Authors	Asmus Freytag
Date	August 20, 1999
This Version	http://www.unicode.org/unicode/reports/tr14-5
Previous Version	http://www.unicode.org/unicode/reports/tr14-4.html
Latest Version	http://www.unicode.org/unicode/reports/tr14

Summary

This report presents the specification of line breaking properties for Unicode characters.

Status of this document

This document contains informative material and normative specifications which have been considered and approved by the Unicode Technical Committee for publication as a Technical Report and as part of the Unicode Standard, Version 3.0 (forthcoming). Any reference to version 3.0 of the Unicode Standard automatically includes this technical report.

The content of all technical reports must be understood in the context of the appropriate version of the Unicode Standard. References in this technical report to sections of the Unicode Standard refer to the Unicode Standard, Version 3.0. See http://www.unicode.org/unicode/standard/versions/ for more information.

This technical report may undergo further editorial work before the release of the Unicode Standard, Version 3.0. The contents of section 7.2 are preliminary. Please mail corrigenda and other comments to the authors.

1.0 Overview and Scope

The Unicode Standard presents a summary of basic line-breaking behavior, but does not give a complete specification. This technical report provides the needed information in a way that reflects best practices. Normative line-breaking properties are assigned to those characters whose line breaking behavior must be identical across all implementations. Default, informative line-breaking properties for all other classes of characters are supplied as well.

2.0 Definitions

All terms not defined here shall be as defined in the Unicode Standard.

Line fitting - the process of determining the how much text will fit on a line of text, given the available space between the margins and the actual display width of the text.

Overfull - a line that contains so much text that it does not fit in the space allotted, or only after unacceptable compression of the text.

Underfull - a line that contains so little text that it ends too far from the margin, or one that would require unacceptable expansion when lines are justified.

Line Break - the position in the text where one line ends and the next one starts.

Line Break Opportunity - a place where a line is allowed to end. Whether a given position in the text is a valid line break opportunity depends on the line breaking rules in force, as well as on context.

Line Breaking Property - A character property with mutually exclusive values, as set out in Table 1. Line breaking properties are used to determine the type of break.

Mandatory Break - a line must break following a character that has the mandatory break property. Also known as a *forced* break.

Direct Break - a line break opportunity exists between two characters of given line breaking properties. In addition, if they are separated by one or more spaces, a break opportunity exists after the last space. This is indicated in the text as **B** ‡ **A**, where **B** is the character class of the character *before* and **A** is the character class of the character *after* the break.

Indirect Break - no line breaking opportunity exists between two adjacent characters of given line breaking properties. However, if they are separated by one or more space characters, a break opportunity exists after the last space. This indicated in the text as **B** % **A**.

Prohibited Break - no line breaking opportunity exists between two characters of given line breaking properties. This is indicated in the text as **B** ^ **A**.

Line Breaking - the process of selecting that part of a text that can be displayed on a line. In other words, selecting one among several line breaking opportunities such that the resulting line is optimal (unless the user requested an explicit line break).

Hyphenation — Hyphenation uses language specific rules to provide additional line breaking opportunities within a word. Hyphenation improves the layout of narrow columns, especially for languages with many longer words, such as German or Finnish. For the purpose of this document, it is assumed that hyphenation is equivalent to insertion of *soft hyphen* characters. All other aspects of hyphenation are outside the scope of this document.

Value	Line Breaking Property	Examples	Characters with this property
BK *	*Mandatory Break*	NL, PS	cause a line break.
CR *	*Carriage Return*	CR	cause a line break, except between CR and LF
LF *	*Line Feed*	LF	cause a line break, except between CR and LF
CM *	*Combining Marks*	Combining Marks, Conjoining Jamo	prohibit a line break between the character and the preceding character
SG *	*Surrogates*	High Surrogates	prohibit a break from a following low surrogate
ZW *	*Zero Width Space*	ZWSP	optional break
IN	*Inseparable*	Leaders	allow only indirect line breaks between pairs.
GL *	*Non-breaking ("Glue")*	NBSP, ZWNSP	prohibit line breaks before or after.
CB *	*Contingent Break Opportunity*	Inline Objects	provide a line break opportunity contingent on additional information.
SP *	*Space*	Space	generally provide a line break opportunity after the character, enables indirect breaks
BA	*Break Opportunity After*	Spaces, Hyphens	generally provide a line break opportunity after the character
BB	*Break Opportunity Before*	Punctuation used in dictionaries	generally provide a line break opportunity before the character.
B2	*Break Opportunity Before and After*	EM Dash	provide a line break opportunity before and after the character
HY	*Hyphen*	Hyphen-Minus	provide a line break opportunity after the character, except in numeric context
NS	*Non Starter*	small kana	allow only indirect line break before
OP	*Opening*	"(", "[", "{", etc.	prohibit a line break after
CL	*Closing*	")", "]", "}", etc.	prohibit a line break before
QU	*Ambiguous Quotation*	Quotation marks	act like they are both opening and closing
EX	*Exclamation*	"!", "?" etc.	prohibit line break before
ID	*Ideographic*	Ideographs	break before or after
NU	*Numeric*	Digits	form numeric expressions for line breaking purposes
IS	*Infix Separator (Numeric)*	. ,	prevent breaks after any and before numeric
SL	*Symbols allowing breaks*	/	prevent a break before, and allow a break after
AL	*Ordinary Alphabetic*	Alphabets and regular symbols	are alphabetic characters
PR	*Prefix (Numeric)*	$, £, ¥, etc.	don't break in front of a numeric expression
PO	*Postfix (Numeric)*	%, ¢, ‰, º	don't break following a numeric expression

SA	*Complex Context*	South East Asian: Thai, Lao, Khmer	Provide a line break opportunity contingent on additional, language specific context analysis
XX	*Unknown*	Unassigned	are all characters with (as yet) unknown line breaking behavior or unassigned code positions

Table 1 Line Breaking Properties (* = normative)

3.0 Description

Lines are broken as result of either of two conditions. The first condition is the presence of an explicit line breaking character. The second condition results from a formatting algorithm having selected among available line breaking opportunities the particular one that results in the optimal layout of the text.

The definition of optimal line break is outside the scope of this document. Different formatting algorithms may use different methods of determining an optimal break. For example, simple implementations just consider a line at a time, trying find a *locally optimal* line break. A common approach is to allow no compression and consider the longest line that fits. When compression is allowed, a locally optimal line break seeks to balance the relative merits of the resulting amounts compression and expansion for different line break candidates.

More complex algorithms may take into account the interaction of line breaking decisions for the whole paragraph. TEX is a well known example of such a *globally optimal* strategy that may make complex tradeoffs to avoid unnecessary hyphenation and other legal, but inferior breaks. For the purpose of this document, what is important is not so much what defines the optimal amount of text on the line, but how line breaking *opportunities* are defined.

Three principal styles of context analysis determine line-breaking opportunities.

1. Western (spaces and hyphens are used to determine breaks)
2. East Asian (lines can break anywhere, unless prohibited)
3. South East Asian (require morphological analysis)

The first is commonly used for scripts employing the space character. The second is used with East Asian ideographic scripts. The third is used for scripts such as Thai, which do not use spaces, but which restrict word-breaks to syllable boundaries, the determination of which requires knowledge of the language comparable to that required by a hyphenation algorithm.

NOTE: Korean may alternately use a space-based (style 1) instead of the style 2 context analysis.

Space-based line breaking is often augmented by hyphenation. Some Unicode characters have explicit line breaking properties assigned to them. These can be used for the first and second type context analysis for line break opportunities. For multilingual text, styles one and two can be unified into a single set of specifications.

NOTE: Interpretation of line breaking properties in bi-directional text takes place before applying rule L1. of the Unicode Bidirectional Algorithm. However, it is strictly independent of directional properties of the characters or of any auxiliary information determined by the application of rules of that algorithm.

4.0 Conformance

- The line breaking behavior of characters with normative line breaking properties are described in The Unicode Standard. (See The Unicode Standard, Version 3.0, Chapters 6 and 13). Unless otherwise stated, the information in this technical report is not intended to supersede the normative specifications found in The Unicode Standard, but to organize the description in a different context and provide additional informative detail.
- All line breaking properties are informative, except for the line breaking properties marked with a * in Table 1 *Line Breaking Properties*. The behavior for characters with normative line breaking properties must be the same for all conformant implementations. For the purpose of determining conformance, all informative line breaking properties are equivalent, as if they all had been merged into a single category ‘other’. Conformance can then be determined by comparing the resulting breakpoints for string of these normative character classes.
- Conformant implementations must not tailor characters with normative line breaking properties to any of the informative properties, but may tailor characters with informative line breaking properties to one of the normative line breaking properties.

- Higher level protocols may further restrict, override, or extend the line breaking properties of certain characters in some contexts.

5.0 Specification

The following sections list all Unicode characters grouped by their line breaking property and provide additional description of their line breaking behavior.

The main emphasis in this section is to fix the membership of character classes for each line breaking property. The classification by properties defined here is used as input into two algorithms defined below that implement workable default line breaking methods.

Precedence

Little attempt has been made to make the narrative descriptions of the property self-consistent with all the other descriptions. However a rough precedence level is provided by the order of their appearance below.

Where line breaking properties are mutually exclusive of each other, the earlier one in the list applies. For example an explicitly breaking character provides an unconditional line break even when following a 'no-break' character, because explicitly breaking characters appear earlier in the list.

Properties

Each property is marked with an annotation for easy reference showing that

A - the property allows a break opportunity *after* in specified contexts

XA - the property prevents a break opportunity *after* in specified contexts

B - the property allows a break opportunity *before* in specified contexts

XB - the property prevents a break opportunity *before* in specified contexts

P - the property allows a break opportunity for a p*a*ir of same characters

XP - the property prevents a break opportunity for a *pair* of same characters

BK - Explicitly breaking characters (A) – (normative)

Explicit breaks act independently of the surrounding characters.

000C PAGE SEPARATOR (FF)

Form Feed separates a page. The text on the new page starts at the beginning of the line. No paragraph formatting is applied.

2028 LINE SEPARATOR (LS)

The text after the Line Separator starts at the beginning of the line. No paragraph formatting is applied.

This is similar to HTML

2029 PARAGRAPH SEPARATOR (PS)

The text of the new paragraph starts at the beginning of the line. Paragraph formatting is applied. This is similar to HTML <P>

"NEW LINE FUNCTION (NLF)"

New line functions provide additional explicit breaks. They are not individual characters, but are expressed as sequences of control characters NL, LF, and CR. What particular sequence(s) form a NLF depends on the implementation and other circumstances as described in Unicode Technical Report 13, *Unicode Newline Policy.*

If a the character sequence for a *new line function* contains more than one character, it is kept together. The default behavior is to break after LF or CR, but not between CR and CR, or CR and LF. Two additional line break classes have been added for convenience in this operation.

CR – Carriage Return (A) – (normative)

000D CARRIAGE RETURN (CR)

Do not break if followed by a LF, mandatory break after otherwise

LF – Line Feed (A) – (normative)

000A LINE FEED (LF)

There is a mandatory break after any LF character.

CM - Attached characters (XB) – (normative)

Combining characters

Combining character sequences are treated as units for the purposes of line breaking. The line-breaking behavior of the sequence is that of the base character. If U+0020 SPACE is used as a base character, it is treated as AL instead of SP.

All characters with general category Mn, Mc, and Me.

Conjoining Jamos

1160..11F9 Conjoining Jamos

A sequence of conjoining Jamos is used to make up a Hangul syllable. Breaks are only allowed around the entire Hangul syllable, and then the line break properties are the same for precomposed Hangul syllables as for conjoined sequence of Jamos.

> **NOTE:** non-initial conjoining Jamos thus behave like combining marks, while the initial combining Jamos have the same property as Hangul Syllables.

Control and formatting characters

Most controls and formatting characters are ignored in line breaking. All characters of General Category Cc and Cf, unless listed explicitly elsewhere.

SG – Surrogates (XP) – (normative)

All characters with General Category Cs. There is no break between a High surrogate and a low surrogate.

ZW – Zero Widths Space (A) – (normative)

200B ZERO WIDTH SPACE (ZWSP)

This character does not have width. It is used to enable additional (invisible) break opportunities wherever SPACE cannot be used.

GL - Non-breaking or "glue" characters (XB/XA) – (normative)

The action of these characters is to glue together both left and right neighbor character such that they are kept on the same line. If they follow a space character, they still allow a break.

FEFF ZERO WIDTH NO-BREAK SPACE (ZWNBSP)

Since this character is not visible, it is the preferred choice for keeping characters together that would otherwise be split across the line at a direct break.

00A0 NOBREAK SPACE (NBSP)

202F NARROW NO-BREAK SPACE (NBSP)

NO-BREAK SPACE is the preferred character to use where two words should be visually separated but kept on the same line, as in the case of a title and a name "Dr.<NBSP>Joseph Becker". NARROW NO-BREAK SPACE is used in Mongolian.

2007 FIGURE SPACE

This is the preferred space to use in numbers. It has the same width as a digit and keeps the number together for the purpose of line breaking.

2011 NON-BREAKING HYPHEN (NBHY)

This is the preferred character to use where words must be hyphenated but may not be broken at the hyphen.

0F0C TIBETAN MARK DELIMITER TSHEG BSTAR

This looks exactly like a Tibetan *tsheg*, but can be used to prevent a break. It inhibits breaking on either side, like *no-break space*.

CB - Contingent break opportunity characters (B/A)

Contingent Break Opportunity Before and After

FFFC OBJECT REPLACEMENT CHARACTER

By default there is a break opportunity both *before* and *after* the object. Object-specific line break behavior is implemented in the associated object itself, and where available can override the default to prevent either or both of the break opportunities.

IN - Inseparable characters (XP)

Leaders

These characters are intended to be used in consecutive sequence. They therefore prevent line breaks absolutely in a series of two character of this class.

2024 ONE DOT LEADER

2025 TWO DOT LEADER

2026 HORIZONTAL ELLIPSIS

Horizontal ellipsis can be used as a three dot leader.

SP - Break opportunity after characters (A)

Breaking Spaces

0020 SPACE (SP)

The space characters are explicit break opportunities, but spaces at the end of a line are not measured for fit. If there is a sequence of space characters, and breaking after any of the space characters would result in the same visible line, the line breaking position after the last space character in the sequence is the locally most optimal one. In other words, since the last character measured for fit is BEFORE the space character, any number of space characters are kept together invisibly on the previous line and the first non-space character starts the next line.

NOTE: SPACE, but none of the other breaking spaces, is used in determining an indirect break.

BA - Break opportunity after characters (A)

Like SP, but are not part in determining indirect breaks.

These characters with General category Zs

2000	EN QUAD
2001	EM QUAD
2002	EN QUAD
2003	EM QUAD
2004	THREE-PER-EM SPACE
2005	FOUR-PER-EM SPACE
2006	SIX-PER-EM SPACE
2008	PUNCTUATION SPACE
2009	THIN SPACE
200A	HAIR SPACE

The preceding list of space characters all have a specific width, but behave otherwise as breaking spaces.

Tabs

Except for the effect of the location of the tabstops, the tab character acts similarly to a space for the purpose of line breaking.

0009	TAB

Breaking Hyphens

Breaking hyphens establish explicit break opportunities immediately after each occurrence.

There are three types of hyphens: Explicit hyphens, conditional hyphens, and dictionary-inserted hyphens (as a result of a hyphenation process). There is no character code for the third kind of hyphen; therefore if it is desired to make the distinction, the dictionary-inserted hyphens must be represented out of band, or with a privately assigned control code.

2010	HYPHEN
058A	ARMENIAN HYPHEN

Hyphens are graphic characters with width. Since, unlike spaces, they print, they are included in the measured part of the preceding line

00AD	SOFT HYPHEN (SHY)

SHY is rendered invisibly and has no width, *except* at a line break. The rendering of the soft hyphen depends on the script. For the Latin script it is rendered as a hyphen, however, some languages require a change in spelling surrounding an optional hyphen, if it occurs at a line break. For example in German “Becker” changes to “Bek-ker” when hyphenated.

The action of a hyphenation algorithm is equivalent to the insertion of a SHY. However, when a word contains an explicit SHY it is customarily treated as overriding the action of the hyphenator for that word.

0F0B	TIBETAN MARK INTERSYLLABIC TSHEG
1361	ETHIOPIC WORDSPACE
1680	OGHAM SPACE MARK
17D5	KHMER SIGN BARIYOOSAN

The Tibetan *thseg* is a visible mark, but it functions effectively like a space to separate words (or other units) in Tibetan. It provides a break opportunity after itself, like space.

Ethiopian word space is visible word delimiter and is kept on the line before.

The Ogham space mark is rendered visibly between words but should be elided at the end of a line.

BB – Break opportunity before characters (BB)

1806 MONGOLIAN TODO SOFT HYPHEN

The Mongolian Todo soft hyphen provides a line break opportunity, but it stays with the following line

B2 – Break opportunity before and after characters (BA/XP)

2014 EM DASH

The em dash character is used to set off parenthetical text, normally without spaces. Line breaks can occur before and after an em dash, but not between two em dashes. Pairs of em dashes are often used instead of quotation dash.

HY - Hyphen (XA)

002D HYPHEN-MINUS

Some additional context analysis is required to distinguish usage of this character as a hyphen from the use as minus sign (or indicator of numerical range). If used as hyphen, it acts like HYPHEN.

> **NOTE**: In some practice runs of HYPHEN-MINUS are used to stand in for longer dashes or horizontal rules. If it is desired to treat them like the characters or layout elements they stand for, and actual character code conversion is not performed, line breaking will need to support these special cases explicitly.

OP - Opening characters (XA)

The opening character of any set of paired punctuation must be kept with the following character

Characters of general category Ps in the Unicode Character Database.

CL - Closing characters (XB)

The closing character of any set of paired punctuation must be kept with the preceding character

3001..3002 IDEOGRAPHIC COMMA..IDEOGRAPHIC FULL STOP

FF0C FULLWIDTH COMMA

FF0E FULLWIDTH FULL STOP

FE50 SMALL COMMA

FE52 SMALL FULL STOP

FF61 HALFWIDTH IDEOGRAPHIC FULL STOP

FF64 HALFWIDTH IDEOGRAPHIC COMMA

plus any characters of general category Pe in the Unicode Character Database.

SL - Solidus with break after (A)

URLs are common enough now in regular plain text, that they must be taken into account when assigning general purpose line breaking properties. The SL line break property is intended to provide a break after, but not in front of digits so as to not break “1/2” or “06/07/99”.

002F SOLIDUS

Slash (SOLIDUS) is allowed as an additional, limited break opportunity to improve layout of web addresses

> **NOTE:** Normally, symbols are treated as **AL**. If it is desired to allow other breaks, more symbols can be added to this category, or category BA, BB, B2 by tailoring, for example “=”. Mathematics requires additional specifications for line breaking, which are outside the scope of this document.

QU - Ambiguous Quotation mark Characters (XB/XA)

Some paired characters can be either opening or closing depending on usage. The default is to treat them as both opening and closing.

Note: If language information is available, it can be used to determine which character is used as opening and which as closing quote. (See the information in the Unicode Standard, Version 3.0, Chapter 6.)

Characters of general category Pf or Pi in the Unicode Character Database as well as,

0022	QUOTATION MARK
0027	APOSTROPHE

NS - Non-starters (XB)

Some characters cannot start a line, but unlike CL they may allow a break in some context when they are following one or more space characters.

All characters with the following combination of General Category and East Asian Width

Sk(w) + Lm(w) + Lm(h)

plus the following characters

0E5A..0E5B	THAI CHARACTER ANGKHANKHU..THAI CHARACTER KHOMUT
17D4	KHMER SIGN KHAN
17D6..17DA	KHMER SIGN CAMNUC PII KUUH..KHMER SIGN KOOMUUT
203C	DOUBLE EXCLAMATION MARK
2044	FRACTION SLASH
301C	WAVE DASH
30FB	KATAKANA MIDDLE DOT
FE54..FE55	SMALL SEMICOLON..SMALL COLON
FF1A	FULLWIDTH COLON.. FULLWIDTH SEMICOLON
FF65	HALFWIDTH KATAKANA MIDDLE DOT
FF70	HALFWIDTH KATAKANA-HIRAGANA PROLONGED SOUND MARK

Plus all Hiragana, Katakana, and Halfwidth Katakana “small” characters

Note: Optionally, the NS restriction may be relaxed and characters treated like ID, to achieve a more permissive style of line breaking.

EX - Exclamation / Interrogation (XB)

These behave like closing characters, except in relation to postfix and ‘non-starter’ characters

0021	EXCLAMATION MARK
003F	QUESTION MARK
FE56..FE57	SMALL QUESTION MARK..SMALL EXCLAMATION MARK
FF01	FULLWIDTH EXCLAMATION MARK
FF1F	FULLWIDTH QUESTION MARK

ID - Ideographic characters (B/A)

Do not require other characters to provide break opportunities, can ordinarily break before and after and between pairs.

4E00..9FAF	CJK UNIFIED IDEOGRAPHS
3400..4DBF	CJK UNIFIED IDEOGRAPHS EXTENSION A
F900..FAFF	CJK COMPATIBILITY IDEOGRAPHS
3000	IDEOGRAPHIC SPACE

AC00..D7AF	HANGUL SYALLABLES
3130..318F	HANGUL COMPATIBILITY JAMO
1100..115F	HANGUL JAM0 (ONLY THE INITIALS)
	HIRAGANA (except small characters)
	KATAKANA (except small characters)
A000..A4C8	YI SYLLABLES
A490-ACFF	YI RADICALS
2E80.. 2FFF	CJK, KANGXI RADICALS, DESCRIPTION SYMBOLS
FF10-FF19	WIDE DIGITS

plus all of the 3000-33FF blocks not covered elsewhere

NOTE: use ZWNBSP as a manual override to prevent break opportunities around ideographs.

IS - Infix Numeric Separator characters (XB)

Characters that usually occur inside a numerical expression, may not be separated from following numeric characters, unless space character intervenes. Since they are otherwise sentence ending punctuation, they prevent breaks before.

There is no break in “100.00” or “10,000”, nor in “12:59”

002C	COMMA
002E	FULL STOP
003A	COLON
003B	SEMICOLON
0589	ARMENIAN FULL STOP

PR - Prefix characters (Numeric) (XA)

Characters that usually precede a numerical expression, may not be separated from following numeric characters or following opening characters, EVEN if space character intervenes.

There is no break in “$ (100.00)”

All currency symbols (General Category Sc) except as listed explicitly in PO and the following:

002B	PLUS
005C	REVERSE SOLIDUS
00B1	PLUS-MINUS
2116	NUMERO SIGN
2213	MINUS-PLUS

PO - Postfix characters (Numeric) (XB)

Characters that usually follow a numerical expression, may not be separated from preceding numeric characters or preceding closing characters, EVEN if space character intervenes.

There is no break in “(12.00) %”

0025	PERCENT SIGN
00A2	CENT SIGN
00B0	DEGREE SIGN
2030	PER MILLE SIGN
2031	PER TEN THOUSAND SIGN
2032..2035	PRIME..REVERSED TRIPLE PRIME

20A7	PESETA SIGN
2103	DEGREE CELSIUS
2109	DEGREE FAHRENHEIT
2126	OHM SIGN
FE6A	SMALL PERCENT SIGN
FF05	FULLWIDTH PERCENT SIGN
FFE0	FULLWIDTH CENT SIGN

NU - Numeric characters (XP)

Behave like ordinary characters in the context of ordinary characters, activate the prefix and postfix behavior of prefix and postfix characters

DECIMAL DIGITS (All characters of General Category Nd. except FULL WIDTH)

SA - Complex-context dependent characters (P)

Runs of these characters require morphological analysis to determine break opportunities. This is similar to e.g. a hyphenation algorithm. For the characters that have this property, **no** line breaks will be found otherwise, therefore complex context analysis is mandatory.

> **Note:** These characters can be mapped into their equivalent line break classes as result of dictionary lookup, thus permitting a logical separation of this algorithm from the morphological analysis.

If dictionary lookup is not available they should be treated as XX.

All characters of General Category Lo or Lm in these ranges:

0E00..0EFF	THAI / LAO
1780..17FF	KHMER

AL - Ordinary alphabetic and symbol characters (XP)

Require other characters to provide break opportunities, otherwise no breaking between pairs of ordinary characters. However, this is tailorable. In some Far Eastern documents it may be desirable to allow breaking between pairs of ordinary characters.

> **NOTE:** use ZWSP as a manual override to provide break opportunities around alphabetic or symbol characters.

ALPHABETIC — all characters of General Category Lx, except as they appear above
SYMBOLS — all characters of General Category Sx, except as they appear above

AI – Ambiguous (Alphabetic or Ideograph)

Characters with East Asian Width property A (ambiguous width), and which would otherwise be AL in this classification, take on the AL line break class only when their *resolved* width is N (narrow) and take the ID line break class, when their resolved width is W (wide).

XX - Unknown (XP)

Unassigned code positions and characters for which reliable line breaking information is not available (e.g. Private use characters) are assigned this default line breaking property. The behavior is otherwise identical to class AL. Implementations can override or tailor this default behavior, e.g. by assigning private use characters the property ID if that is more likely to give the correct default behavior for their users. Users can manually insert ZWSP or ZWNBSP around characters of class XX to force or prevent breaks as needed.

6.0 Additional information

Dictionary usage

Dictionaries follow strict standards that guide their use of characters to indicate features of the terms listed. Some of these conventions mark places that can also serve as line breaking opportunities and therefore interact with line breaking and are described here. If implemented, these characters would be inserted in the corresponding property above.

GL - Non-breaking or "glue" characters (XA/XB)

Some dictionaries use character that looks like a vertical series of four dots to indicate places where there is a syllable, but no break. This character has not been encoded in Unicode.

BA - Break opportunities after characters (A)

2027 HYPHENATION POINT

Hyphenation point is primarily used to visibly indicate syllabification of words. Syllable breaks are potential line breaking opportunities in the middle of words. The hyphenation point It is mainly used in dictionaries and similar works. When an actual line breaking opportunity falls inside a word containing hyphenation point characters, the hyphenation point is rendered as a regular hyphen at the end of the line.

00B4 ACCUTE ACCENT

In dictionaries, stressed syllables are indicated with a spacing acute accent instead of the hyphenation point. In this case the accent would move to the next line, and the preceding line ended with a hyphen.

007C VERTICAL LINE

In some dictionaries, a vertical bar is used instead of a hyphenation point. In this usage, U+0323 COMBINING DOT BELOW is used to mark stressed syllables, so all breaks are marked by the vertical bar. For an actual break opportunity, the vertical bar is rendered as a hyphen.

BB - Break opportunities before characters (B)

02C8 MODIFIER LETTER VERTICAL LINE

02CC MODIFIER LETTER LOW VERTICAL LINE

These characters are used in dictionaries to indicate stress and secondary stress when IPA is used. Both are prefixes to the stressed syllable in IPA. Therefore, the only sensible way to break them is to keep them with the syllable. The line breaker should break *before* them.

> **NOTE**: It is hard to find actual examples in most dictionaries, since the pronunciation fields usually occur right after the head word, and the columns are wide enough to prevent line breaks in the pronunciations.

6.1 Additional Details on Dictionary usage

Chambers Twentieth Century, new edition 1972. Puts stress accent after syllable. Uses hyphens between syllables. Accent stays with syllable, followed by hyphen when splitting line. E.g. apocalypse. No special convention if splitting at hyphen. I have also encountered a dictionary where a natural hyphen in a word becomes a tilde dash if the word is split. Looking up the noun “syllable” in eight dictionaries yields eight different conventions!

Dictionary of the English Language, Samuel Johnson 1843 **SY´LLABLE** where ´ is a U+02B9 (and a large one at that) and follows the vowel of the main syllable (not the syllable itself).

Oxford English Dictionary 1st Edition **si·lâ'bl** where · is a slightly above middle dot indicating the vowel of the stressed syllable (similar to Johnson's acute). â is really U+0103. The ' is an apostrophe.

Oxford English Dictionary 2nd Edition Has gone to IPA **’slleb(e)l** where **’** is U+02C8, l is U+026A, e is U+0259 (both times). The **’** comes before the stressed syllable. The () indicates the schwa may be omitted.

Chambers English Dictionary 7th Edition **sil´e-bl** where the stressed syllable is followed by ´ U+02B9, **e** is U+0259, **-** is a hyphen when splitting a word like **abate´- ment** the stress mark ´ goes after stressed syllable followed by the hyphen.

*BBC English Dictionar*y **sIlebl** where I is U+026A U+0332, e is U+0259. The vowel of the stressed syllable is underlined.

Collins Cobuild English Language Dictionary **sIlebe°l** where I is U+026A U+0332, and means the same as the BBC. The e is U+0259 (both times). The **°** is a U+2070 and indicates the schwa may be omitted.

Readers Digest Great Illustrated Dictionary. **syl·la·ble** (**sílleb'l**) The spelling of the word has hyphenation points (· is a U+2027) followed by phonetic spelling. The vowel of the stressed syllable is given an accent (rather than being followed by an accent). e is schwa and ' is apostrophe.

Webster's 3rd New International Dictionary. **syl·la·ble /'silebel/** The spelling of the word has hyphenation points (· is a U+2027) and is followed by phonetic spelling. The stressed syllable is preceded by **'** U+02C8. The e's are schwas as usual. Webster splits words at the end of a line with a normal hyphen. When a hyphenated word is split at the hyphen this is indicated by a double hyphen which looks like a light version of the German Fraktur hyphen (short equals sign with a slight slope up to the right).

7. 0 Implementation notes

The Unicode Standard, Version 2.0, describes a particular method for boundary detection in Chapter 5. It is based on a set of hierarchical rules and character classifications. That algorithm would be well suited for implementation of some of the advanced heuristics.

A simpler algorithm can be devised that uses a two dimensional table to resolve break opportunities between pairs or characters.

7.1 Rule based Algorithm

The linebreaking algorithm can be expressed in a series of rules which take line breaking classes as input.

Linebreaking rules

The rules are applied in order. That is, there is an implicit "otherwise" at the front of each rule.

> ! Mandatory break
>
> ^ No break allowed at the indicated position
>
> ‡ Break allowed at the indicated position

They are stated in terms of LB classes. The examples use representative characters for clarity instead of the acronyms. 'H' stands for an ideographs, 'h' for small kana, '9' for digits.

Resolve line break classes:

LB1. Assign a line break category to each character of the input. Resolve CB, SA, XX, SG into other line break classes depending on criteria outside this algorithm.

Start and end of text:

LB2a. Never break at the start of text

> ^ sot

LB2b. Always break at the end of text

> ! eot

These two rules are designed to deal with degenerate cases. Their effect is to have at least one character on each line, and at least one linebreak for the whole text. "Emergency line breaking behavior usually also allows line breaks anywhere on the line if a legal line break cannot be found. This has the effect of preventing text to run over the margins.

Mandatory breaks:

LB 3a. Always break after hard line breaks (but never between CR and LF). There is a break opportunity after every ZWSP, but not a hard break.

> CR ^ LF

LF !

CR !

BK !

LB 3b. Don't break before hard line breaks.

^ (BK | CR | LF)

Explicit breaks and non-breaks:

LB4. Don't break before spaces.

^ SP

^ ZW

LB 5.Don't break before or after ZWNBSP

^ GL

GL ^

Combining Marks:

At any possible break opportunity between CM and a following character, CM behaves as if it had the type of its base character. If there is no base, the CM behaves like AL. Virama and non-initial Jamo are treated as CM and initial Jamo are merged with class ID so they work correctly.

LB 6. Don't break graphemes (before combining marks, around virama or on sequences of conjoining Jamos.

^ CM

Treat X { CM * } as if it was X

LB 7. In all of the following rules, if a space is the base character for a combining mark, the space is changed to type AL.

Treat SP CM * as if it was AL

Opening and closing:

these have special behavior with respect to spaces.

LB 8.Don't break before '!' or '/' or ',' or ']', even after spaces.

{ SP * } ^ CL

LB 9.Don't break after '[', even after spaces.

OP { SP * } ^

LB 10. Don't break within '"[', , even with intervening spaces.

QU { SP * } ^ OP

LB 11. Don't break within ']h', even with intervening spaces.

CL { SP* } ^ NS

Spaces:

LB 12. Break after spaces

SP ‡

LB 13. Break after hyphens and before BB

HY ‡

‡ BB

Special case rules:

LB 14.Don't break before or after ""

^ QU

QU ^

LB 15. Don't break before small kana and other non starters, or before Hyphen-Minus, or before other

hyphens or after BB

^ NS

^ HY

^ BA

BB

LB 16. Don't break between two ellipses, or between letters or numbers and ellipsis:

IN ^ IN

NU ^ IN

AL ^ IN

Examples: '9...', 'a...', 'H...'

LB 18. Don't break within 'a9', '3a', or 'H%'

AL ^ NU

NU ^ AL

ID ^ PO

Numbers:

Numbers are of the form PO SP * (CL | HY) ? NU+) ? PO

i.e.'$' SP* ('(' | '-')? '9'+ ')'? '%',

Examples:

$(12.35)

2,1234

(12)¢

12.54¢

This is approximated with the following rule. (Some cases already handled above,like '9,', '[9'.)

LB 18. Don't break between the follwowing pairs of classes

PR ^ NU

PR ^ OP

PR ^ HY

HY ^ NU

SL ^ NU

NU ^ NU

IS ^ NU

NU ^ PO

CL ^ PO

Example pairs:

'$9', '$[', '$-', '-9', '/9', '99', ',9', '9%' ']%'

Finally, join alphabetic letters and break everything else.

LB 20. Don't break between alphabetics ("at")

AL ^ AL

LB 22.- Break everywhere else.

B ‡ A

where B is any class of character before, and A is any class of character after the break position.

7.2 Pair table based algorithm

A two dimensional table can be used to resolve break opportunities between pairs of characters. The rows of the table are labeled by the possible values of the line breaking property of the leading character in the pair, the columns are labeled by the line breaking property for the following character of the pair. Each intersection is labeled with the resulting line breaking opportunity.

The Japanese standard JIS X 4051-1995 provides an example of such a table-based definition. However, it uses line breaking classes whose membership is not solely determined by line breaking property (as in this report), but in some cases by heuristic analysis or markup of the text.

Minimal table

If two rows of the table have identical values and the corresponding columns also have identical values, the two line breaking classes can be coalesced. The JIS standard uses 20 classes of which only 14 appear to be unique.

Extended context

By broadening the definition of pair from BA to B {SP*} A where A and B are characters and SP* is an optional run of space characters, the same table can be used to distinguish between cases where SPACE can or cannot provide a line breaking opportunity (i.e. direct and indirect breaks). (Equivalent rules to the ones above can be formulated to the ones above, not using SP, but using % to express indirect breaks. These rules can then be simplified to involve only pairs of classes, e.g. only constructions of the form

B‡A

B%A

B ^ A

where either A or B may be empty. These simplified rules can then be automatically translated into a pair table, as in the example below. Line break analysis then proceeds by pair table lookup.

Example table

The following example table implements the line breaking behavior described in this Technical Report, within the limitation that only context of the form B {SP * } A is considered. BK and SP classes are handled explicitly in the outer loop as given in the code sample below. B {CM*} can be handled in the table, or explicitly in the outer loop. Using the table entries is equivalent to making the simplifying assumption that combining marks are always behave as if applied to **AL.**

	‘After’ class																
	OP	**CL**	**QU**	**GL**	**NS**	**EX**	**SL**	**IS**	**PR**	**PO**	**NU**	**AL**	**ID**	**IN**	**HY**	**CM**	‘Before’ class
OP	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	S	OP — open
CL		X	S	S	X	X	X	X		S					s	S	CL — close
QU	X	X	S	S	S	X	X	X	S	S	S	S	S	S	S	S	QU — quotation
GL	S	X	S	S	S	X	X	X	S	S	S	S	S	S	S	S	GL — glue
NS		X	S	S	S	X	X	X							s	S	NS — no-start
EX		X	S	S	S	X	X	X							s	S	EX — exclamation/interrogation
SL		X	S	S	S	X	X	X			S				s	S	SL — Syntax (slash)
IS		X	S	S	S	X	X	X			S				s	S	IS — infix (numeric) separator
PR	S	X	S	S	S	X	X	X			S	S	S		s	S	PR — prefix
PO		X	S	S	S	X	X	X							s	S	PO — postfix
NU		X	S	S	S	X	X	X		S	S	S		S	s	S	NU — numeric
AL		X	S	S	S	X	X	X			S	S		S	s	S	AL — alphabetic
ID		X	S	S	S	X	X	X		S				S	s	S	ID — ideograph (atomic)
IN		X	S	S	S	X	X	X						S	s	S	IN — inseparable
HY		X	S	S	S	X	X	X							S	S	HY — hyphens and spaces
CM		X	S	S	S	X	X	X			S	S		S	s	S	CM — Combining Marks

X denotes a prohibited break: Never break here, even if one or more spaces intervene (**^** avove)
S denotes an indirect break opportunity: Don’t break here, unless one of more spaces intervene (**%** in rule notation)
_ an empty cell denotes a direct break opportunity (‡ above)

{ED note: update pair table and consolidate sample code below.}

The following two functions demonstrate how the pair table is used.

```
// placeholder function for complex break analysis
int findComplexBreak(int *pcls, int *pbrk, int cch)
{
	if (!cch)  return 0;
	int cls = pcls[0];
	for(int ich = 0; ich < cch; ich++) {

		// .. do complex break analysis here

		if (pcls[ich] != SA)
			break;
	}
	return ich;
}

// handle spaces separately, all others by table
int findLineBrk1(int *pcls, int *pbrk, int cch)
{
	if (!cch) return;

	int  cls = pcls[0];
	for (int ich = 1; ich < cch && cls != BK; ich++) {
		if (pcls[ich] == SP) {
			pbrk[ich-1] = XX;
			continue;
		}

		if (pcls[ich] == SA) {
			ich += findComplexBreak(&pcls[ich-1], &pbrk[ich-1], cch - (ich-1));
			if (ich < cch)
				cls = pcls[ich];
			continue;
		}

		// lookup pair table information
		int brk = brkPairs[cls][pcls[ich]];

		if (brk == SS)	{
			pbrk[ich-1] = ((pcls[ich - 1] == SP) ? SS : XX);
		} else {
			pbrk[ich-1] = brk;
		}
		cls = pcls[ich];
	}
	pbrk[ich-1] = 0;

	return ich;
}
```

If one makes the simplifying assumption that combining marks are only applied to AL, or that applying a combining mark turns the combination into AL, then CM can be handled in the table as shown. Otherwise a simple statement in the outer loop

```
if (pcls[i] == CM) {
	pbrk[ich-1] = 0;
	continue;
}
```

would have the effect of letting the CM take on the class of the preceding non-CM characters. This also requires a special rule to cover the case of a missing base character in the setup part before the loop:

```
if (pcls[i] == CM)
	cls = SP;
```

7.3 Customization

A real world line breaking algorithm must be tailorable to some degree. There are three principle ways of

tailoring a pair-table based algorithm:

1. Change the line breaking class assignment for some characters
2. Change the table value assigned to a pair of character classes
3. Change the interpretation of the line breaking actions
4. Augment the algorithm.

- The first is useful for cases where the line breaking properties of one class of characters are occasionally lumped together with the properties of another class to achieve a less restrictive line breaking behavior.
- The second method is particularly useful if the behavior can be expressed by a change at a limited number of pair intersections. These intersections can be labeled with special values that cause different actions for different customizations.
- The third method is equivalent to the second, but instead of changing table values, an additional indirection is performed. This is most suitable when customizations need to be done at run time.
- The fourth method is the most open ended...

7.4 Examples of customization:

1. Korean uses either implicit breaking around Hangul and ideographs or uses spaces. Reference [1] shows how this can be elegantly handled by the second or third method. Only the intersection of **ID**/**ID**, **AL**/**ID** and **ID**/**AL** are affected.For alphabetic style line breaking, breaks for these four cases require space, for ideographic style line breaking, these four cases don't require spaces.
2. Sometimes allowing alphabetic characters and digit strings to break anywhere is required in Far Eastern context. According to reference [1] this can be done by the second or third method, affecting the intersections of **NU**/**NU**, **NU**/**AL**, **AL**/**AL**, and **AL**/**NU**.
3. Force a keep on Kana syllables, i.e. kyu, spelled KI yu would be kept together even though KI and yu are normally atomic. This can be handled via the first method, by changing the classification of the Kana small characters between **ID** and **NS** as needed.

8.0 Further Information

[1] Michel Suignard, *Worldwide Typography and How to Apply JIS X 4051-1995 to Unicode,* Proceedings of the Twelfth International Unicode/ISO 10646 Conference, Tokyo, Japan, 1998

[2] Cy Cedar, David Veintimilla, Michel Suignard and Asmus Freytag, *Report from the Trenches: Microsoft Publisher goes Unicode,* Proceedings of the Eleventh International Unicode Conference, San Jose, CA 1997

[3] *The Unicode Standard, Version 2.0,* (Reading, Massachusetts: Addison-Wesley 1996)

- to be superseded by

[3a] *The Unicode Standard, Version 3.0,* (Reading, Massachusetts: Addison-Wesley 1999)

[4] Donald E. Knuth and Michael F. Plass, *Breaking Lines into Paragraphs*, , republished in *Digital Typography*, CSLI 78, (Stanford, California: CLSI Publications1997)

[5] Donald E. Knuth, *T_EX, the Program*, Volume B of *Computers & Typesetting*, (Reading, Massachusetts: Addison-Wesley 1986)

9.0 Acknowledgments

The initial assignments of properties are based on input by Michel Suignard. Mark Davis provided algorithmic verification and formulation of the rules. Ken Whistler, Rick McGowan and other members of the editorial committee provided valuable feedback. Tim Partridge enlarged the information on dictionary usage.

10.0 Changes from previous revision:

Extensive changes to all sections as result of review in the Unicode Technical Committee.

11.0 Copyright

Unicode Home Page: http://www.unicode.org

Unicode Technical Reports: http://www.unicode.org/unicode/reports