Proposed changes for line breaking on orthographic syllables

To: UTC
From: Robin Leroy, PAG
Date: 2023-03-25

This proposal is the result of the review by the Properties and Algorithms Group of proposal L2/22-080R2, *Line breaking at orthographic syllable boundaries*. A rationale for the changes proposed herein may be found in that document, or in L2/22-086 *Specification updates for orthographic syllables and line breaking*; only deviations from the proposed text and property value assignments in L2/22-080R2 are motivated in this document.

> **PAG amendment:** The changes with respect to L2/22-080R2 are explained in boxes like this one.

Additions ~~and deletions~~ with respect to the current proposed update (revision 50, draft 3, dated 2022-12-07) are highlighted in cyan. Additions ~~and deletions~~ with respect to Unicode Version 15.0 that are already part of the proposed update are highlighted in yellow, as in the proposed update.

Contents

A Changes to properties

The proposed changes to the unicodetools repository with respect to the unicodetools repository at f6b0188da3853de772dc4f1a060f65256b24d2a3 (current main), including all UCD files—in particular the source files LineBreak.txt and PropertyValueAliases.txt—as well as the invariants tests, are given as a patch file attached to this document.

They may also be seen online by comparing these files between:

- 25c51bd1451b0ff1da00245052e3407d7f013fee (current state of unicode-org/unicodetools#422), and
- f6b0188da3853de772dc4f1a060f65256b24d2a3 (current main).

> **PAG amendment:** The change from lb=AL to lb=AK for U+25CC DOTTED CIRCLE is not included. See the discussion under LB 28b in this document.
>
> The line breaking class for U+1107F BRAHMI NUMBER JOINER is changed from CM to GL rather than WJ: this is because classes WJ and ZW are reserved for explicit overrides to the line breaking algorithm. Characters which prohibit line breaks because of shaping are instead given class GL; examples include the Mongolian vowel separator, the combining grapheme joiner, some format controls for Egyptian hieroglyphs, etc.

B Changes to UAX #29

[Only tables 3 and 4 are changed. The remainder of the annex is omitted.]

Table 3. Word_Break Property Values

Value	Summary List of Characters
[15 unchanged rows omitted.]	
Numeric	Line_Break = Numeric *or ~~any of the following:~~* General_Category = Decimal_Number ~~U+FF10 (0) FULLWIDTH DIGIT ZERO~~ ~~..U+FF19 (9) FULLWIDTH DIGIT NINE~~ *and not* U+066C (٬) ARABIC THOUSANDS SEPARATOR

Table 4. Sentence_Break Property Values

Value	Summary List of Characters
[9 unchanged rows omitted.]	
Numeric	Line_Break = Numeric *or ~~any of the following:~~* General_Category = Decimal_Number ~~U+FF10 (0) FULLWIDTH DIGIT ZERO~~ ~~..U+FF19 (9) FULLWIDTH DIGIT NINE~~

PAG amendment: L2/22-080R2 does not change UAX #29. Since numbers are being removed from lb=NU, the Word_Break and Sentence_Break property assignments would change if the Unicode 15.0 derivation were retained. This was not the intent of the proposal for line breaking at orthographic syllable boundary.

The changes to UAX #29 leave the property assignments invariant.

C Changes to UAX #14

2 Definitions

[Unchanged text omitted.]

Table 1. Line Breaking Classes

Class	Descriptive Name	Examples	Behavior
		[4 sections omitted.]	
		Other Characters	
AI	*Ambiguous (Alphabetic or Ideographic)*	Characters with Ambiguous East Asian Width	Act like AL when the resolved EAW is N; otherwise, act as ID
AK	*Aksara*	Consonants	Form orthographic syllables in Brahmic scripts
AL	*Alphabetic*	Alphabets and regular symbols	Are alphabetic characters or symbols that are used with alphabetic characters
AP	*Aksara Pre-Base*	Pre-base repha	Form orthographic syllables in Brahmic scripts
AS	*Aksara Start*	Independent vowels	Form orthographic syllables in Brahmic scripts
CJ	*Conditional Japanese Starter*	Small kana	Treat as NS or ID for strict or normal breaking.
EB	*Emoji Base*	All emoji allowing modifiers	Do not break from following Emoji Modifier
EM	*Emoji Modifier*	Skin tone modifiers	Do not break from preceding Emoji Base
H2	*Hangul LV Syllable*	Hangul	Form Korean syllable blocks
H3	*Hangul LVT Syllable*	Hangul	Form Korean syllable blocks

HL	*Hebrew Letter*	Hebrew	Do not break around a following hyphen; otherwise act as Alphabetic
ID	*Ideographic*	Ideographs	Break before or after, except in some numeric context
JL	*Hangul L Jamo*	Conjoining jamo	Form Korean syllable blocks
JV	*Hangul V Jamo*	Conjoining jamo	Form Korean syllable blocks
JT	*Hangul T Jamo*	Conjoining jamo	Form Korean syllable blocks
RI	*Regional Indicator*	REGIONAL INDICATOR SYMBOL LETTER A .. Z	Keep pairs together. For pairs, break before and after other classes
SA	*Complex Context Dependent (South East Asian)*	South East Asian: Thai, Lao, Khmer	Provide a line break opportunity contingent on additional, language-specific context analysis
VF	*Virama Final*	Viramas for final consonants	Form orthographic syllables in Brahmic scripts
VI	*Virama*	Conjoining viramas	Form orthographic syllables in Brahmic scripts
XX	*Unknown*	Most unassigned, private-use	Have as yet unknown line breaking behavior or unassigned code positions

3 Introduction

[Unchanged text omitted.]

3.1 Determining Line Break Opportunities

~~Three~~ Four principal styles of context analysis determine line break opportunities.

1. *Western:* spaces and hyphens are used to determine breaks
2. *East Asian:* lines can break anywhere, unless prohibited
3. *South East Asian:* line breaks require morphological analysis
4. *Brahmic:* line breaks can occur at the boundaries of any orthographic syllable

The Western style is commonly used for scripts employing the space character. Hyphenation is often used with space-based line breaking to provide additional line break opportunities—however, it requires knowledge of the language and it may need user interaction or overrides.

The second style of context analysis is used with East Asian ideographic and syllabic scripts. In these scripts, lines can break anywhere, except before or after certain characters. The precise set of prohibited line breaks may depend on user preference or local custom and is commonly tailorable.

Korean makes use of both styles of line break. When Korean text is justified, the second style is commonly used, even for interspersed Latin letters. But when ragged margins are used, the Western style (relying on spaces) is commonly used instead, even for ideographs.

The third style is used for scripts such as Thai, which allow line breaks only at word boundaries, but do not mark word boundaries in any way, so that the determination of line break opportunities requires language dependent text analysis.~~do not use spaces, but which restrict word breaks to syllable boundaries, whose determination requires knowledge of the language comparable to that required by a hyphenation algorithm. Such an algorithm is~~ Algorithms and data for such analysis are beyond the scope of the Unicode Standard.

The fourth style is used for Brahmic scripts that allow line breaks to occur at the boundaries of any orthographic syllable, without restricting them to word boundaries. This style is only supported for scripts that encode orthographic syllables in primarily phonetic order.

For multilingual text, the Western ~~and~~, East Asian, and Brahmic styles can be unified into a single set of specifications, based on the information in this annex. Unicode characters have explicit line breaking properties assigned to them. These properties can be utilized to implement the effect of both of these two styles of context analysis for line break opportunities. Customization for user preferences or document style can then be achieved by tailoring that specification.

In bidirectional text, line breaks are determined before applying rule L1 of the Unicode Bidirectional Algorithm [UAX9]. However, line breaking is strictly independent of directional properties of the characters or of any auxiliary information determined by the application of rules of that algorithm.

[Unchanged section 4 omitted.]

5 Line Breaking Properties

[Unchanged text omitted.]

5.1 Description of Line Breaking Properties

Line breaking classes are listed alphabetically. Each line breaking class is marked with an annotation in parentheses with the following meanings:

Label	Meaning for the Class
(A)	It **allows** a break opportunity **after** in specified contexts.
(XA)	It **prevents** a break opportunity **after** in specified contexts.
(B)	It **allows** a break opportunity **before** in specified contexts.
(XB)	It **prevents** a break opportunity **before** in specified contexts.

(P)	It **allows** a break opportunity for a **pair** of same characters.
(XP)	It **prevents** a break opportunity for a **pair** of same characters.

Note: The use of the letters **B** and **A** in these annotations marks the position of the break opportunity relative to the character. It is not to be confused with the use of the same letters in the other parts of this annex, where they indicate the positions of the characters relative to the break opportunity.

[Descriptions unchanged with respect to the proposed update are omitted. The descriptions remain in alphabetical order of short names.]

AK: Aksara (XB/XA)

The AK line break class is used for scripts that use the Brahmic style of context analysis and have a virama of Indic syllabic category Virama or Invisible_Stacker. It contains characters that can occur as the bases of orthographic syllables and can also follow a virama of Indic syllabic category Virama or Invisible_Stacker within the same orthographic syllable. Depending on the script, this may include characters with the Indic syllabic categories Consonant, Vowel_Independent, or Number.

1B05..1B33	BALINESE LETTER AKARA..BALINESE LETTER HA
1B45..1B4C	BALINESE LETTER KAF SASAK..BALINESE LETTER ARCHAIC JNYA
A984..A9B2	JAVANESE LETTER A..JAVANESE LETTER HA
11005..11037	BRAHMI LETTER A..BRAHMI LETTER OLD TAMIL NNNA
11071..11072	BRAHMI LETTER OLD TAMIL SHORT E..BRAHMI LETTER OLD TAMIL SHORT O
11075	BRAHMI LETTER OLD TAMIL LLA
11305..1130C	GRANTHA LETTER A..GRANTHA LETTER VOCALIC L
1130F..11310	GRANTHA LETTER EE..GRANTHA LETTER AI
11313..11328	GRANTHA LETTER OO..GRANTHA LETTER NA
1132A..11330	GRANTHA LETTER PA..GRANTHA LETTER RA
11332..11333	GRANTHA LETTER LA..GRANTHA LETTER LLA
11335..11339	GRANTHA LETTER VA..GRANTHA LETTER HA
11360..11361	GRANTHA LETTER VOCALIC RR..GRANTHA LETTER VOCALIC LL

11F04..11F10	KAWI LETTER A..KAWI LETTER O
11F12..11F33	KAWI LETTER KA..KAWI LETTER JNYA

AP: Aksara Pre-Base (B/XA)

The AP line break class is only used for scripts that use the Brahmic style of context analysis. It contains the characters of such scripts that are part of an orthographic syllable but in logical order precede the base or any half-forms. This includes characters with the Indic syllabic categories Consonant_Preceding_Repha, Consonant_With_Stacker, and Consonant_Prefixed.

11003..11004	BRAHMI SIGN JIHVAMULIYA..BRAHMI SIGN UPADHMANIYA
11F02	KAWI SIGN REPHA

AS: Aksara Start (XB/XA)

The AS line break class is only used for scripts that use the Brahmic style of context analysis. It contains characters that can occur as the bases of orthographic syllables, but can not follow a virama of Indic syllabic category Virama or Invisible_Stacker within the same orthographic syllable. Depending on the script, this may include characters with the Indic syllabic categories Consonant, Vowel_Independent, Number, and several others.

1BC0..1BE5	BATAK LETTER A..BATAK LETTER U
AA00..AA28	CHAM LETTER A..CHAM LETTER HA
11066..1106F	BRAHMI DIGIT ZERO..BRAHMI DIGIT NINE
11350	GRANTHA OM
1135E..1135F	GRANTHA LETTER VEDIC ANUSVARA..GRANTHA LETTER VEDIC DOUBLE ANUSVARA
11EE0..11EF1	MAKASAR LETTER KA..MAKASAR LETTER A
11F50..11F59	KAWI DIGIT ZERO..KAWI DIGIT NINE

CM: Combining Mark (XB) (Non-tailorable)

Combining Characters

Combining character sequences are treated as units for the purpose of line breaking. The line breaking behavior of the sequence is that of the base character.

The preferred base character for showing combining marks in isolation is U+00A0 NO-BREAK SPACE. If a line break before or after the combining sequence is desired, U+200B ZERO WIDTH SPACE can be used. The use of U+0020 SPACE as a base character is deprecated.

For most purposes, combining characters take on the properties of their base characters, and that is how the CM class is treated in rule LB9 of this specification. As a result, if the sequence <0021, 20E4> is used to represent a triangle enclosing an exclamation point, it is effectively treated as EX, the line break class of the exclamation mark. If U+26A0 WARNING SIGN had been used, which also looks like an exclamation point inside a triangle, it would have the line break class of AL. Only the latter corresponds to the line breaking behavior expected by users for this symbol. To avoid surprising behavior, always use a base character that is a symbol or letter (Line Break AL) when using enclosing combining marks (General_Category Me).

The CM line break class includes all combining characters with General_Category Mc, Me, and Mn, unless listed explicitly elsewhere. This includes *viramas* that don't have line break class VI or VF.

NU: Numeric (XP)

These characters behave like ordinary characters (AL) in the context of most characters but activate the prefix and postfix behavior of prefix and postfix characters.

Numeric characters consist of decimal digits (all characters of General_Category Nd), except:

1. those with East_Asian_Width F (Fullwidth)
2. those from scripts that use the Brahmic style of context analysis,

plus these characters:

066B	ARABIC DECIMAL SEPARATOR
066C	ARABIC THOUSANDS SEPARATOR

PAG amendment: L2/22-080R2 proposes removing the numbers from the affected scripts from lb=NU (giving them lb=AS or lb=ID instead), but omits the update to this description.

VF: Virama Final (XB/A)

The VF line break class is only used for scripts that use the Brahmic style of context analysis. It contains the viramas of Indic syllabic category Pure_Killer in scripts where the final consonant of a phonological syllable is expressed as a sequence of a consonant and such a virama, and the final consonant needs to be kept together with the preceding orthographic syllable. This includes:

1BF2..1BF3	BATAK PANGOLAT..BATAK PANONGONAN

Viramas of Indic syllabic category Pure_Killer that don't meet the conditions for line break class VF use the line break class CM.

VI: Virama (XB/XA)

The VI line break class is only used for scripts that use the Brahmic style of context analysis. It contains the viramas of Indic syllabic categories Virama and Invisible_Stacker of such scripts.

1B44	BALINESE ADEG ADEG
A9C0	JAVANESE PANGKON
11046	BRAHMI VIRAMA
1134D	GRANTHA SIGN VIRAMA
11F42	KAWI CONJOINER

6 Line Breaking Algorithm

[Text unchanged with respect to the proposed update omitted. LB28b is inserted after LB28.]

***LB28b** Do not break inside the orthographic syllables of Brahmic scripts.*

AP × (AK | ◌ | AS)

(AK | ◌ | AS) × (VF | VI)

(AK | ◌ | AS) VI × (AK | ◌)

(AK | ◌ | AS) × (AK | ◌ | AS) VF

PAG amendment: L2/22-080R2 assigns lb=AK to the dotted circle ◌. This would induce breaks in sequences such as e÷◌ॅ÷◌ु or a÷◌ॅ. These sequences are used to describe combining sequences, *e.g.*, in the comments in http://www.unicode.org/Public/UCD/latest/ucd/NormalizationTest.txt.

However, retaining lb=AL would introduce breaks even in sequences (dotted circle, virama), formerly AL × CM, now AL ÷ VI. Such sequences are clearly not degenerate in the sense of UAX #29, and we should try to make them work well. The PAG considered multiple options to avoid these breaks, such as using AK | AL instead of AK in this rule. However, in order to support subjoined dotted circles (AK × VI × ◌), but (AK × VI ÷ AL which may legitimately occur in multilingual text), the dotted circle needed to be special cased.

Instead of creating a new class lb=DC and replacing all usages of AL by AL|DC throughout the rules (as was done with class HL), the PAG favoured explicit reference to the exceptional character. There is now ample precedent for the rules using more general Unicode regular expressions: they make usage of the properties General_Category, Extended_Pictographic, and East_Asian_Width.

[Text unchanged with respect to the proposed update omitted.]

8 Customization

[Text unchanged with respect to the proposed update omitted.]

8.1 Types of Tailoring

[Text unchanged with respect to the proposed update omitted.]

8.2 Examples of Customization

[Examples 1 through 7 are unchanged with respect to the proposed update and omitted.]

~~***Example 8.*** For some implementations it may be difficult to implement LB9 due to the added complexity of its indefinite length context. Because combining marks are most commonly applied to characters of class AL, rule LB10 by itself generally produces acceptable results for such implementations, but such an approximation is not a conformant tailoring.~~

Example 8. Some scripts that traditionally follow the Brahmic style of context analysis are nowadays occasionally written with spaces, and word-based line breaking might be desired in that case. This can be accomplished by remapping the line break classes AK, AP, and AS to AL; and VI or VF to CM. In some cases other word-forming characters, such as U+A9CF JAVANESE PANGRANGKEP, also need to be remapped to AL. Digits, which may have line break class AS or ID in such scripts, need to be remapped to NU. Punctuation, which may have line break class ID in such scripts, need to be remapped to AL or BA.