L2/22-080R

Line breaking at orthographic syllable boundaries

Norbert Lindenberg with contributions from Elika Etemad and Vaishnavi Murthy Yerkadithaya
2022-04-15

Proposal

This document proposes:

- To introduce a new style of context analysis in line breaking for certain Brahmic scripts, which breaks lines at the boundaries of orthographic syllables and uses the new Line_Break property values AK (Aksara), AP (Aksara Pre-Base), AS (Aksara Start), VF (Virama Final), and VI (Virama), as specified below in the section Specification of line breaking at orthographic syllable boundaries.
- To update the Line_Break property to use this new style of line breaking for the scripts Balinese, Batak, Brahmi, (Eastern) Cham, Grantha, Javanese, Kawi, Makasar, and Tulu Tigalari, as specified below in the sections about these scripts.
- To update the Line_Break property value of the character U+25CC DOTTED CIRCLE from AL to AK to enable its use as a placeholder for subjoined consonants, as specified below in the section Enabling the use of dotted circle as a placeholder for subjoined consonants.

A gray background in this document indicates proposed specification text for Unicode Standard Annex #14: Unicode Line Breaking Algorithm, or proposed Line_Break property data for the Unicode data files LineBreak.txt and PropertyValueAliases.txt.

This document relies on a definition of the term "orthographic syllable" and on updates to the line breaking information for the Balinese, Batak, Brahmi, and Grantha scripts in the Unicode Standard proposed in L2/22-086 Specification updates for orthographic syllables and line breaking, as well as on the Indic syllabic categories assigned to characters in Brahmic scripts in the IndicSyllabicCategory.txt file.

Error report

This proposal takes up an error report that Elika Etemad sent to the Unicode Consortium on 2019-11-05:

> **Overview:**
>
> UAX14 and Unicode Chapter 17.4 disagree on line-breaking in Javanese.
>
> **Details:**

Unicode Chapter 17.4 says that Javanese breaks between orthographic syllables, and defines a BNF pattern for these syllables.

UAX14 says Javanese is treated as AL, which does not allow breaks between units.

These requirements conflict.

Proposal:

In UAX14, recategorize Javanese as SA, which is defined to determine breakpoint based on lexical analysis.

Links:

https://www.unicode.org/versions/Unicode12.0.0/ch17.pdf
http://unicode.org/reports/tr14/#AL "no line breaks are allowed between pairs"
http://unicode.org/reports/tr14/#SA "require morphological analysis to determine break opportunities"

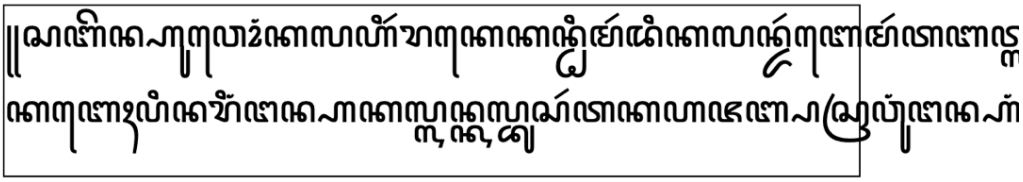
Current Unicode line breaking

Unicode Standard Annex #14: Unicode Line Breaking Algorithm and the associated Unicode data file LineBreak.txt specify a standard algorithm for line breaking. It identifies three principal styles of context analysis in determining line break opportunities:

- *Western*: spaces and hyphens are used to determine breaks.
- *East Asian*: lines can break anywhere, unless prohibited.
- *South East Asian*: line breaks require morphological analysis.

The line breaking style used for a script is determined by the Line_Break property values of its main letters: AL for Western style; ID for East Asian style; SA for South East Asian style. The main Javanese letters, consonants and independent vowels, are currently set to AL, resulting in the Western style.

The Western style requires the use of spaces, hyphens, or similar characters to identify possible line breaks. Since most Javanese text does not contain spaces or hyphens, no line breaks can be found except for those after punctuation. As the bug report states, class AL is incompatible with Javanese line breaking requirements. In current browsers, Javanese text overflows the width of the paragraph it's supposed to fit in (here indicated by the black rectangle).



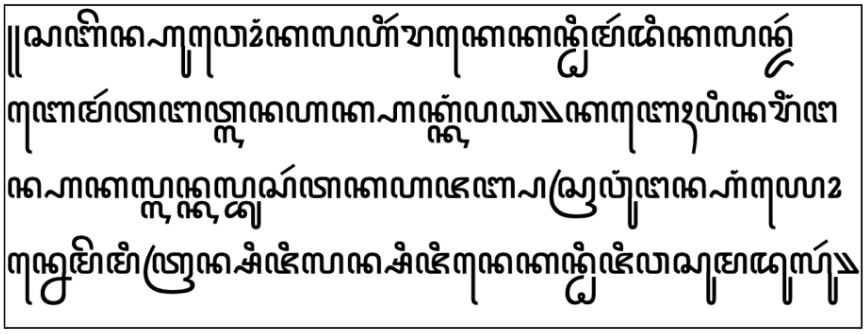
The bug report proposes to recategorize Javanese as SA (Complex Context Dependent – South East Asian). This would move a solution for Javanese line breaking out of the scope of the Unicode Line Breaking Algorithm, as it requires external algorithms for determining line breaks for scripts using the South East Asian style. It would also indicate to potential implementors that it's a hard problem, as the scripts for which SA was originally created (Thai, Lao, Khmer, and Myanmar) require the creation and application of language specific dictionaries. In the absence of an external line breaking algorithm for Javanese, the Unicode Line Breaking Algorithm falls back to Western style, so that the end result wouldn't change.

In reality, the complexity implied by the South East Asian style is not required for Javanese, nor for several other Brahmic scripts, as line breaking at orthographic syllable boundaries doesn't actually require morphological or otherwise complex analysis.

Goals and non-goals for this proposal

Goals for this proposal:

- Introduce line breaking at orthographic syllable boundaries, as specified in the Unicode Standard for Javanese and other scripts, as the fourth principal style of context analysis in determining line break opportunities.
- Enable rendering of the Javanese text above in a more sensible way, at a level that users would consider a reasonable default.



- Do the same for several other Brahmic scripts.

Non-goals for this proposal:

- Coverage of all Brahmic scripts that might need this style of line breaking. The Unicode Standard has no information on line breaking for the majority of Brahmic scripts. This proposal includes the ones for which the standard states that line breaks occur at orthographic syllable boundaries, or for which I've otherwise been able to obtain such information. For other scripts, experts are invited to provide information on line breaking so that this style of line breaking can be enabled where appropriate.
- Perfection in line break data. For most of the included scripts, not enough information is available on which base characters have conjunct forms, or on the line breaking behavior around punctuation.

The proposed data should result in reasonable line breaking behavior (far more reasonable than the current situation), but may need to be fine-tuned later.

- Use of this style of line breaking as a second level for other styles. According to Martin Hosken and Muthu Nedumaran, line breaking at orthographic syllable boundaries is commonly used as a second level to fit long words into short lines for scripts that primarily use the Western style (e.g., Tamil) or the South East Asian style (e.g., Thai, Myanmar), similar to the use of hyphenation in some other languages. An extension of the Unicode line breaking algorithm to support multiple levels is beyond the scope of this proposal.
- Support for scripts with visual encoding order. Four Brahmic scripts in Unicode use a primarily visual encoding order: Thai, Lao, New Tai Lue, and Tai Viet. These four scripts use the South East Asian style of context analysis, so breaking at orthographic syllable boundaries for them has not been considered. For Thai in particular, it is known that syllable boundaries cannot be easily determined.
- Breaking within orthographic syllables. Some writing systems allow line breaks within orthographic syllables. In some scripts this is the normal convention, as documented for Batak and Tulu-Tigalari below. In other scripts, such as Kawi and Javanese, it is seen occasionally when a writer runs out of space on a palm leaf or other writing surface on which already written characters can't be erased. In rare cases, such line breaks have also been used in typeset materials. In Unicode-based rendering, this is difficult to reproduce, as breaking within an orthographic syllable will in most cases cause the font rendering system to insert a dotted circle before the marks that were moved to the new line. The Unicode line breaking algorithm should therefore not break within an orthographic syllable. Developers of enhanced line breaking algorithms that are designed for tighter integration with rendering algorithms may choose to implement this feature.
- Fixes to the definition of grapheme clusters. The specification for grapheme clusters in UAX 29 is broken for many Brahmic scripts, as it separates viramas from subsequent consonants with which they would otherwise create conjunct forms. For viramas of Indic syllabic category Virama that is at least surprising to users; for viramas of Indic syllabic category Invisible_Stacker it is an obvious failure. Class definitions similar to the ones used in this proposal might be used to fix the specification for grapheme clusters. Doing so, however, is outside the scope of this proposal.

Specification of line breaking at orthographic syllable boundaries

The orthographic syllables of Brahmic scripts that encode them in primarily phonetic order are easy to identify with a regular expression. For line breaking, it's not necessary to distinguish between all the different kinds of marks that can attach to a base, or to watch the sequencing of these marks, as would be necessary for validation and rendering. Only a few classes of characters are needed, primarily base characters (consonants, independent vowels, and some others), conjoining virama (those with Indic syllabic category Virama or Invisible_Stacker), visible virama (those with Indic syllabic category Virama or Pure_Killer), and other marks. Most orthographic syllables could be recognized by this:

Base Mark* (Conjoining_Virama Base Mark*)* Visible_Virama?

Note that this regular expression does not identify which of the base characters is the actual base of the orthographic syllable – in some cases the first (or first several) base characters may combine with viramas to create half-forms or repha forms, in which case a later base character would be the actual base. For line breaking, this is irrelevant.

In most Brahmic scripts, not all characters that can be the base of an orthographic syllable can also be attached to such a base using a virama. For example, independent vowels generally can be bases, but only some of them can be attached to a base. In some scripts, numbers can be bases, but they can't be attached to a base. In Tamil, only a small set of consonants have conjunct forms. To allow line breaks between a conjoining virama and a base with which it can't conjoin, we use separate classes for base characters that can both precede and follow a virama within the same orthographic syllable (AK – Aksara), and characters that can only precede a virama (AS – Aksara Start).

A few Brahmic scripts have characters that are encoded before the base character and, where allowed, half-forms, of the orthographic syllable they belong to. This includes characters of the Indic syllabic categories Consonant_Preceding_Repha, Consonant_With_Stacker, and Consonant_Prefixed. We're adding class AP (Aksara Pre-Base) for such characters.

In some Brahmic scripts, final consonants expressed as a consonant and a virama of Indic syllabic category Pure_Killer need to be kept together with the preceding orthographic syllable. For such scripts, this virama gets line break class VF (Virama Final). For scripts where such a combination can be separated from the preceding orthographic syllable, the virama uses the existing line break class CM. If at some point in the future line breaking at orthographic syllable boundaries were implemented for the Myanmar script, the repha-like kinzi marks would require special attention, as they are encoded as sequences Consonant-Pure_Killer-Virama, but must not be kept together with preceding orthographic syllables.

As is already the case for letters in the current specification of the Unicode Line Breaking Algorithm, arbitrary sequences of characters of class CM and ZWJ are allowed after any character of classes AK, AP, AS, VI, or VF.

The complete regular expression for orthographic syllables, omitting CM and ZWJ, then becomes:

AP? (AS | AK) (VI AK)* (VI | VF | (AS | AK) VF)?

The remainder of this section describes the changes to Unicode Standard Annex #14: Unicode Line Breaking Algorithm needed to add line breaking at orthographic syllable boundaries.

In Table 1, Line Breaking Classes, section Other Characters, add:

Class	Descriptive Name	Examples	Behavior
AK	*Aksara*	Consonants	Form orthographic syllables in Brahmic scripts
AP	*Aksara Pre-Base*	Pre-base repha	Form orthographic syllables in Brahmic scripts
AS	*Aksara Start*	Independent vowels	Form orthographic syllables in Brahmic scripts
VF	*Virama Final*	Viramas for final consonants	Form orthographic syllables in Brahmic scripts
VI	*Virama*	Conjoining viramas	Form orthographic syllables in Brahmic scripts

The new line breaking classes need formal short and long names to be defined in the Unicode data file PropertyValueAliases.txt:

```
lb ; AK                             ; Aksara
lb ; AP                             ; Aksara_Prebase
lb ; AS                             ; Aksara_Start
lb ; VF                             ; Virama_Final
lb ; VI                             ; Virama
```

Back in UAX #14, section 3.1, Determining Line Break Opportunities, add:

> 4. *Brahmic:* line breaks can occur at the boundaries of any orthographic syllable
>
> ...
>
> 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.

In the same section, change:

- "Three" to "Four".
- "the Western and East Asian styles" to "the Western, East Asian, and Brahmic styles".

In section 5.1, Description of Line Breaking Properties, add:

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. As a special case, U+25CC DOTTED CIRCLE is included.

1B05..1B33	BALINESE LETTER AKARA..BALINESE LETTER HA
1B45..1B4C	BALINESE LETTER KAF SASAK..BALINESE LETTER ARCHAIC JNYA
25CC	DOTTED CIRCLE
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
11392..113B5	TULU-TIGALARI LETTER KA..TULU-TIGALARI LETTER LLLA
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
113D1	TULU-TIGALARI REPHA
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
11380..11389	TULU-TIGALARI LETTER A..TULU-TIGALARI LETTER VOCALIC LL
1138B	TULU-TIGALARI LETTER EE
1138E	TULU-TIGALARI LETTER AI
11390..11391	TULU-TIGALARI LETTER OO..TULU-TIGALARI LETTER AU
11EE0..11EF1	MAKASAR LETTER KA..MAKASAR LETTER A
11F50..11F59	KAWI DIGIT ZERO..KAWI DIGIT NINE

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
113D0	TULU-TIGALARI CONJOINER
11F42	KAWI CONJOINER

Also in section 5.1, Description of Line Breaking Properties, in the subsection Combining Characters, change "This includes viramas" to "This includes viramas that don't have line break class VI or VF".

In section 6.2, Tailorable Line Breaking Rules, add the following rule, which keeps orthographic syllables together. As the line break classes used in orthographic syllables are new and not handled in any other rule, any sequences with characters with the new classes that are not handled in this rule fall through to the default rule LB31, which breaks on both sides of the orthographic syllable. Combining marks, ZERO WIDTH JOINER, and ZERO WIDTH NON-JOINER do not need to be considered here, as rule LB9 has already made them invisible to subsequent rules.

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

$$\mathrm{AP} \times (\mathrm{AK} \mid \mathrm{AS})$$
$$(\mathrm{AK} \mid \mathrm{AS}) \times (\mathrm{VF} \mid \mathrm{VI})$$
$$\mathrm{VI} \times \mathrm{AK}$$
$$(\mathrm{AK} \mid \mathrm{AS}) \times (\mathrm{AK} \mid \mathrm{AS})\ \mathrm{VF}$$

Examples showing the application of these rules are provided in the next section.

In section 8.2, Examples of Customization, delete example 8 because its premise "combining marks are most commonly applied to characters of class AL" is no longer correct.

In the same section, add:

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.

Examples for line breaking at orthographic syllable boundaries

This section details the interaction between the new line break classes and new and old line breaking rules with a few examples. Each example shows:

- A sequence of characters, with gaps between them that will be filled in with line breaking decisions.
- The code points for the characters.
- The Line_Break property value for each character, as specified in the script-specific sections below.
- Several rows showing the impact of line breaking rules that modify the sequence of line break classes or make decisions that allow or prohibit line breaks. Commonly applicable rules are LB9, which subsumes combining marks and ZERO WIDTH JOINER into their base characters; the new rule LB28b specified above; and the final rule LB31, which breaks "everywhere else".
- The final result with orthographic syllables separated by line breaking opportunities.

In the columns for line breaking decisions, "÷" means a line break opportunity has been identified; "×" means no line break is allowed. Gray text indicates that this cell has no change and is not involved in the rules applied in this step.

The first example, using Kawi text, shows three of the four components of LB28b in action. Note the pre-base repha 𑼾, the conjoiner 𑼿, which pulls the consonants before and after into an orthographic syllable, and the vowel killer 𑽁, which is simply treated as a combining mark.

Characters	𑼕		𑼾		𑼲		𑼕		𑼿		𑼕		𑼵		𑽁
Code points: 11F-	-26		-02		-2D		-26		-42		-26		-31		-41
Line_Break classes	AK		AP		AK		AK		VI		AK		AK		CM
LB9	AK		AP		AK		AK		VI		AK		AK		
LB28b: AP × (AK | AS)	AK		AP	×	AK		AK		VI		AK		AK		
LB28b: (AK | AS) × (VF | VI)	AK		AP	×	AK		AK	×	VI		AK		AK		
LB28b: VI × AK	AK		AP	×	AK		AK	×	VI	×	AK		AK		
LB31	AK	÷	AP	×	AK	÷	AK	×	VI	×	AK	÷	AK		
Result	𑼕	÷	𑼲𑼾			÷	𑼕𑼿𑼵					÷	𑽁		

The next example, using Batak text, shows how the remaining of the four components of LB28b keeps a final consonant written as a consonant-virama combination together with the previous orthographic syllable, which is necessary to enable the Batak-specific reordering behavior.

Characters	ᯖ	◌᯲	ᯒ	◌᯳	ᯐ	◌᯦	ᯔ	◌᯲	ᯐ	◌᯦
Code points: 1B-	-D7	-EC	-D2	-EA	-C9	-F3	-C2	-E7	-C9	-F3
Line_Break classes	AS	CM	AS	CM	AS	VF	AS	CM	AS	VF
LB9	AS		AS		AS	VF	AS		AS	VF
LB28b: (AK | AS) × (VF | VI)	AS		AS		AS	× VF	AS		AS	× VF
LB28b: (AK | AS) × (AK | AS) VF	AS		AS	×	AS	× VF	AS	×	AS	× VF
LB31	AS	÷	AS	×	AS	× VF	÷	AS	×	AS	× VF
Result	ᯖ᯲	÷	ᯒᯐ◌᯦				÷	ᯔᯐ◌᯦		

The final example, using Balinese text, includes one missed line break opportunity: The user had inserted a ZERO WIDTH NON-JOINER to prevent a virama from combining with the following consonant into a conjunct form. However, the Unicode line breaking algorithm treats ZWNJ as a combining mark, so that it gets subsumed into the preceding character that is not treated as a combining mark, which here is the virama. This lets the rule LB28b to see a VI-AK sequence, which prevents a line break.

It might be possible to work around this by adding a rule before LB9 that adds a line break opportunity after a virama-ZWNJ sequence. However, ZWNJ is used in a large variety of ways, some of which might conflict with such a rule. No such rule is therefore proposed. It is better to miss a line break opportunity than to break in the wrong place. Users can avoid the issue by using ZERO WIDTH SPACE to make the virama visible and enable a line break at the same time.

Characters	ᬓ	ᬳ	◌᭄	ZWNJ	ᬬ	◌᭄	ᬫ	◌᭄	ᬢ	◌᭄
Code points: 1B-	-18	-27	-44	200C	-2B	-38	-31	-44	-1D	-36
Line_Break classes	AK	AK	VI	CM	AK	CM	AK	VI	AK	CM
LB9	AK	AK	VI		AK		AK	VI	AK	
LB28b: (AK | AS) × (VF | VI)	AK	AK ×	VI		AK		AK ×	VI	AK	
LB28b: VI × AK	AK	AK ×	VI	×	AK		AK ×	VI ×	AK	
LB31	AK ÷	AK ×	VI	×	AK	÷	AK ×	VI ×	AK	
Result	ᬓ ÷	ᬳ᭄ᬬ				÷	ᬫ᭄ᬢ			

Specifying Line_Break property data

The documentation for the new line break classes AK, AP, AS, VF, and VI proposed above should be sufficient to select the appropriate classes for characters that form orthographic syllables.

For other characters, a general assumption is that in scripts that allow line breaks at the boundaries of orthographic syllables line breaks can also occur before and after all other spacing characters. An easy way to accomplish this is to give them line break class ID. In some cases, however, conventions may require exceptions:

- Some punctuation may need to be attached to the end of orthographic syllables. Line break class BA works for this.
- Some decorative combinations of punctuation may need to stay grouped together. This can be accomplished by inserting U+2060 WORD JOINER between them.
- Some punctuation characters may not be allowed at the beginning or end of lines. The line break classes CL and CP are designed for this purpose.

Where characters are shared between scripts using the Script_Extensions property, the impact of changing their line break classes needs to be evaluated across scripts, as is done for Grantha and Javanese below.

Line breaking for Balinese

Comparison of current and proposed line breaking.

Based on updated information on line breaking for Balinese in L2/22-086 Specification updates for orthographic syllables and line breaking, this section proposes to enable line breaking at orthographic syllable boundaries for Balinese.

The changes to the Line_Break property shown in red are proposed:

```
1B00..1B03;CM     # Mn     [4] BALINESE SIGN ULU RICEM..BALINESE SIGN SURANG
1B04;CM           # Mc         BALINESE SIGN BISAH
1B05..1B33;AL→AK  # Lo    [47] BALINESE LETTER AKARA..BALINESE LETTER HA
1B34;CM           # Mn         BALINESE SIGN REREKAN
1B35;CM           # Mc         BALINESE VOWEL SIGN TEDUNG
1B36..1B3A;CM     # Mn     [5] BALINESE VOWEL SIGN ULU..BALINESE VOWEL SIGN RA REPA
1B3B;CM           # Mc         BALINESE VOWEL SIGN RA REPA TEDUNG
1B3C;CM           # Mn         BALINESE VOWEL SIGN LA LENGA
1B3D..1B41;CM     # Mc     [5] BALINESE VOWEL SIGN LA LENGA TEDUNG..BALINESE VOWEL SIGN TALING REPA TEDU
1B42;CM           # Mn         BALINESE VOWEL SIGN PEPET
1B43;CM           # Mc         BALINESE VOWEL SIGN PEPET TEDUNG
1B44;CM→VI        # Mc         BALINESE ADEG ADEG
1B45..1B4C;AL→AK  # Lo     [8] BALINESE LETTER KAF SASAK..BALINESE LETTER ARCHAIC JNYA
1B50..1B59;NU→ID  # Nd    [10] BALINESE DIGIT ZERO..BALINESE DIGIT NINE
1B5A..1B5B;BA     # Po     [2] BALINESE PANTI..BALINESE PAMADA
1B5C;AL→ID        # Po         BALINESE WINDU
1B5D..1B60;BA     # Po     [4] BALINESE CARIK PAMUNGKAH..BALINESE PAMENENG
1B61..1B6A;AL→ID  # So    [10] BALINESE MUSICAL SYMBOL DONG..BALINESE MUSICAL SYMBOL DANG GEDE
1B6B..1B73;CM     # Mn     [9] BALINESE MUSICAL SYMBOL COMBINING TEGEH..BALINESE MUSICAL SYMBOL COMBININ
1B74..1B7C;AL→ID  # So     [9] BALINESE MUSICAL SYMBOL RIGHT-HAND OPEN DUG..BALINESE MUSICAL SYMBOL LEFT
1B7D..1B7E;BA     # Po     [2] BALINESE PANTI LANTANG..BALINESE PAMADA LANTANG
```

Line breaking for Batak

Comparison of current and proposed line breaking.

Based on updated information on line breaking for Batak in L2/22-086 Specification updates for orthographic syllables and line breaking, this section proposes to enable line breaking at orthographic syllable boundaries for Batak.

The changes to the Line_Break property shown in red are proposed:

```
1BC0..1BE5;AL→AS  # Lo    [38] BATAK LETTER A..BATAK LETTER U
1BE6;CM           # Mn         BATAK SIGN TOMPI
1BE7;CM           # Mc         BATAK VOWEL SIGN E
1BE8..1BE9;CM     # Mn     [2] BATAK VOWEL SIGN PAKPAK E..BATAK VOWEL SIGN EE
1BEA..1BEC;CM     # Mc     [3] BATAK VOWEL SIGN I..BATAK VOWEL SIGN O
1BED;CM           # Mn         BATAK VOWEL SIGN KARO O
1BEE;CM           # Mc         BATAK VOWEL SIGN U
1BEF..1BF1;CM     # Mn     [3] BATAK VOWEL SIGN U FOR SIMALUNGUN SA..BATAK CONSONANT SIGN H
1BF2..1BF3;CM→VF  # Mc     [2] BATAK PANGOLAT..BATAK PANONGONAN
1BFC..1BFF;AL     # Po     [4] BATAK SYMBOL BINDU NA METEK..BATAK SYMBOL BINDU PANGOLAT
```

Line breaking for Brahmi

Based on updated information on line breaking for Brahmi in L2/22-086 Specification updates for orthographic syllables and line breaking, this section proposes to enable line breaking at orthographic syllable boundaries for Brahmi.

A special situation exists with the non-decimal numbers in the Brahmi script: In general, a line break can occur before and after every character; however, the character U+1107F BRAHMI NUMBER JOINER causes a required ligature between the two surrounding number characters. These numbers are not part of orthographic syllables, and so treating the number joiner as a virama is not appropriate. Instead, we treat it as a word joiner.

The changes to the Line_Break property shown in red are proposed:

```
11000;CM          # Mc         BRAHMI SIGN CANDRABINDU
11001;CM          # Mn         BRAHMI SIGN ANUSVARA
11002;CM          # Mc         BRAHMI SIGN VISARGA
11003..11004;AL→AP# Lo     [2] BRAHMI SIGN JIHVAMULIYA..BRAHMI SIGN UPADHMANIYA
11005..11037;AL→AK# Lo    [51] BRAHMI LETTER A..BRAHMI LETTER OLD TAMIL NNNA
11038..11045;CM   # Mn    [14] BRAHMI VOWEL SIGN AA..BRAHMI VOWEL SIGN AU
11046;CM→VI       # Mn         BRAHMI VIRAMA
11047..11048;BA   # Po     [2] BRAHMI DANDA..BRAHMI DOUBLE DANDA
11049..1104D;AL→ID# Po     [5] BRAHMI PUNCTUATION DOT..BRAHMI PUNCTUATION LOTUS
11052..11065;AL→ID# No    [20] BRAHMI NUMBER ONE..BRAHMI NUMBER ONE THOUSAND
11066..1106F;NU→AS# Nd    [10] BRAHMI DIGIT ZERO..BRAHMI DIGIT NINE
11070;CM          # Mn         BRAHMI SIGN OLD TAMIL VIRAMA
11071..11072;AL→AK# Lo     [2] BRAHMI LETTER OLD TAMIL SHORT E..BRAHMI LETTER OLD TAMIL SHORT O
11073..11074;CM   # Mn     [2] BRAHMI VOWEL SIGN OLD TAMIL SHORT E..BRAHMI VOWEL SIGN OLD TAMIL SHORT O
11075;AL→AK       # Lo         BRAHMI LETTER OLD TAMIL LLA
1107F;CM→WJ       # Mn         BRAHMI NUMBER JOINER
```

Line breaking for (Eastern) Cham

The Unicode Standard's "Cham" script represents only Eastern Cham, as it turned out that Western Cham needs to be encoded separately.

The Unicode Standard, section 16.10, Cham, says "Opportunities for line breaks occur after any full orthographic syllable in Cham." It also describes the separately encoded final consonants as the final components of orthographic syllables. This can be accomplished by using line break class CM or BA. For those with general category Lo, line break class BA seems more appropriate.

The changes to the Line_Break property shown in red are proposed:

```
AA00..AA28;AL→AS  # Lo    [41] CHAM LETTER A..CHAM LETTER HA
AA29..AA2E;CM     # Mn     [6] CHAM VOWEL SIGN AA..CHAM VOWEL SIGN OE
AA2F..AA30;CM     # Mc     [2] CHAM VOWEL SIGN O..CHAM VOWEL SIGN AI
AA31..AA32;CM     # Mn     [2] CHAM VOWEL SIGN AU..CHAM VOWEL SIGN UE
AA33..AA34;CM     # Mc     [2] CHAM CONSONANT SIGN YA..CHAM CONSONANT SIGN RA
AA35..AA36;CM     # Mn     [2] CHAM CONSONANT SIGN LA..CHAM CONSONANT SIGN WA
AA40..AA42;AL→BA  # Lo     [3] CHAM LETTER FINAL K..CHAM LETTER FINAL NG
AA43;CM           # Mn         CHAM CONSONANT SIGN FINAL NG
AA44..AA4B;AL→BA  # Lo     [8] CHAM LETTER FINAL CH..CHAM LETTER FINAL SS
AA4C;CM           # Mn         CHAM CONSONANT SIGN FINAL M
AA4D;CM           # Mc         CHAM CONSONANT SIGN FINAL H
AA50..AA59;AL→ID  # Nd    [10] CHAM DIGIT ZERO..CHAM DIGIT NINE
AA5C;AL→ID        # Po         CHAM PUNCTUATION SPIRAL
AA5D..AA5F;BA     # Po     [3] CHAM PUNCTUATION DANDA..CHAM PUNCTUATION TRIPLE DANDA
```

Line breaking for Grantha

Comparison of current and proposed line breaking.

Based on updated information on line breaking for Grantha in L2/22-086 Specification updates for orthographic syllables and line breaking, this section proposes to enable line breaking at orthographic syllable boundaries for Grantha.

The Grantha script shares several characters with other scripts via the Script_Extensions property. Most of these characters are combining marks and therefore not affected by the proposed changes in the Line_Break property. The remaining characters are:

- 1CD3;AL VEDIC SIGN NIHSHVASA
- 1CF2..1CF3;AL VEDIC SIGN ARDHAVISARGA..VEDIC SIGN ROTATED ARDHAVISARGA
- 0BE6..0BEF;NU TAMIL DIGIT ZERO..TAMIL DIGIT NINE
- 0BF0..0BF2;AL TAMIL NUMBER TEN..TAMIL NUMBER ONE THOUSAND
- 0BF3;AL TAMIL DAY SIGN
- 11FD0..11FD1;AL TAMIL FRACTION ONE QUARTER..TAMIL FRACTION ONE HALF-1
- 11FD3;AL TAMIL FRACTION THREE QUARTERS
- 0964..0965;BA DEVANAGARI DANDA..DEVANAGARI DOUBLE DANDA

Of these, 0964..0965 could attach to Grantha orthographic syllables; others could be separated from them and could instead form “words” within their own categories. This is unlikely to cause significant problems in line breaking.

The changes to the Line_Break property shown in red are proposed:

```
11300..11301;CM   # Mn     [2] GRANTHA SIGN COMBINING ANUSVARA ABOVE..GRANTHA SIGN CANDRABINDU
11302..11303;CM   # Mc     [2] GRANTHA SIGN ANUSVARA..GRANTHA SIGN VISARGA
11305..1130C;AL→AK# Lo     [8] GRANTHA LETTER A..GRANTHA LETTER VOCALIC L
1130F..11310;AL→AK# Lo     [2] GRANTHA LETTER EE..GRANTHA LETTER AI
11313..11328;AL→AK# Lo    [22] GRANTHA LETTER OO..GRANTHA LETTER NA
1132A..11330;AL→AK# Lo     [7] GRANTHA LETTER PA..GRANTHA LETTER RA
11332..11333;AL→AK# Lo     [2] GRANTHA LETTER LA..GRANTHA LETTER LLA
11335..11339;AL→AK# Lo     [5] GRANTHA LETTER VA..GRANTHA LETTER HA
1133B..1133C;CM   # Mn     [2] COMBINING BINDU BELOW..GRANTHA SIGN NUKTA
1133D;AL→BA       # Lo         GRANTHA SIGN AVAGRAHA
1133E..1133F;CM   # Mc     [2] GRANTHA VOWEL SIGN AA..GRANTHA VOWEL SIGN I
11340;CM          # Mn         GRANTHA VOWEL SIGN II
11341..11344;CM   # Mc     [4] GRANTHA VOWEL SIGN U..GRANTHA VOWEL SIGN VOCALIC RR
11347..11348;CM   # Mc     [2] GRANTHA VOWEL SIGN EE..GRANTHA VOWEL SIGN AI
1134B..1134C;CM   # Mc     [2] GRANTHA VOWEL SIGN OO..GRANTHA SIGN AU
1134D;CM→VI       # Mc         GRANTHA SIGN VIRAMA
11350;AL→AS       # Lo         GRANTHA OM
11357;CM          # Mc         GRANTHA AU LENGTH MARK
1135D;AL→BA       # Lo         GRANTHA SIGN PLUTA
1135E..1135F;AL→AS# Lo     [2] GRANTHA LETTER VEDIC ANUSVARA..GRANTHA LETTER VEDIC DOUBLE ANUSVARA
11360..11361;AL→AK# Lo     [2] GRANTHA LETTER VOCALIC RR..GRANTHA LETTER VOCALIC LL
11362..11363;CM   # Mc     [2] GRANTHA VOWEL SIGN VOCALIC L..GRANTHA VOWEL SIGN VOCALIC LL
11366..1136C;CM   # Mn     [7] COMBINING GRANTHA DIGIT ZERO..COMBINING GRANTHA DIGIT SIX
11370..11374;CM   # Mn     [5] COMBINING GRANTHA LETTER A..COMBINING GRANTHA LETTER PA
```

Line breaking for Javanese

Comparison of current and proposed line breaking.

The Unicode Standard, section 17.4, Javanese, says “Opportunities for line breaking occur after any full orthographic syllable. Hyphens are not used.” It also discusses a repetition of U+A9BA JAVANESE VOWEL SIGN TALING that occurs at line breaks in some documents. This repetition is not a requirement, however, and would have to be implemented at a level above simple line breaking, similar to the insertion of hyphenation marks.

In traditional writing, line breaks within orthographic syllables can be found occasionally; however, as noted above, these shouldn’t be supported in the Unicode Line Breaking Algorithm. In modern signage, spaces are commonly inserted between words; however, there’s no standard recommending or prescribing the use of spaces and how lines should be broken in their presence. As line breaks in signage, if any, would be done manually anyway, this can be ignored for the Unicode Line Breaking Algorithm.

The Javanese script shares one character with the Buginese (Lontara’) script via the Script_Extensions property: U+A9CF JAVANESE PANGRANGKEP. The proposed change could cause a line break after this character. According to the Unicode Standard, section 17.2, PANGRANGKEP is used in Buginese as a word duplicator, in which case breaking after it would be fine. However, the information on the use of PANGRANGKEP in Buginese as well as on line breaking in that script appear to be incomplete and require more research.

The changes to the Line_Break property shown in red are proposed:

```
A980..A982;CM     # Mn     [3] JAVANESE SIGN PANYANGGA..JAVANESE SIGN LAYAR
A983;CM           # Mc         JAVANESE SIGN WIGNYAN
A984..A9B2;AL→AK  # Lo    [47] JAVANESE LETTER A..JAVANESE LETTER HA
A9B3;CM           # Mn         JAVANESE SIGN CECAK TELU
A9B4..A9B5;CM     # Mc     [2] JAVANESE VOWEL SIGN TARUNG..JAVANESE VOWEL SIGN TOLONG
A9B6..A9B9;CM     # Mn     [4] JAVANESE VOWEL SIGN WULU..JAVANESE VOWEL SIGN SUKU MENDUT
A9BA..A9BB;CM     # Mc     [2] JAVANESE VOWEL SIGN TALING..JAVANESE VOWEL SIGN DIRGA MURE
A9BC..A9BD;CM     # Mn     [2] JAVANESE VOWEL SIGN PEPET..JAVANESE CONSONANT SIGN KERET
A9BE..A9BF;CM     # Mc     [2] JAVANESE CONSONANT SIGN PENGKAL..JAVANESE CONSONANT SIGN CAKRA
A9C0;CM→VI        # Mc         JAVANESE PANGKON
A9C1..A9C6;AL→ID  # Po     [6] JAVANESE LEFT RERENGGAN..JAVANESE PADA WINDU
A9C7..A9C9;BA     # Po     [3] JAVANESE PADA PANGKAT..JAVANESE PADA LUNGSI
A9CA..A9CD;AL→ID  # Po     [4] JAVANESE PADA ADEG..JAVANESE TURNED PADA PISELEH
A9CF;AL→BA        # Lm         JAVANESE PANGRANGKEP
A9D0..A9D9;NU→ID  # Nd    [10] JAVANESE DIGIT ZERO..JAVANESE DIGIT NINE
A9DE..A9DF;AL→ID  # Po     [2] JAVANESE PADA TIRTA TUMETES..JAVANESE PADA ISEN-ISEN
```

Line breaking for Kawi

The Kawi script is new in Unicode 15.0. The Proposal to encode Kawi specifies that “line breaks may occur after every orthographic syllable”.

The changes to the Line_Break property shown in red are proposed:

```
11F00..11F01;CM   # Mn     [2] KAWI SIGN CANDRABINDU..KAWI SIGN ANUSVARA
11F02;AL→AP       # Lo         KAWI SIGN REPHA
11F03;CM          # Mc         KAWI SIGN VISARGA
11F04..11F10;AL→AK# Lo    [13] KAWI LETTER A..KAWI LETTER O
11F12..11F33;AL→AK# Lo    [34] KAWI LETTER KA..KAWI LETTER JNYA
11F34..11F35;CM   # Mc     [2] KAWI VOWEL SIGN AA..KAWI VOWEL SIGN ALTERNATE AA
11F36..11F3A;CM   # Mn     [5] KAWI VOWEL SIGN I..KAWI VOWEL SIGN VOCALIC R
11F3E..11F3F;CM   # Mc     [2] KAWI VOWEL SIGN E..KAWI VOWEL SIGN AI
11F40;CM          # Mn         KAWI VOWEL SIGN EU
11F41;CM          # Mc         KAWI SIGN KILLER
11F42;CM→VI       # Mn         KAWI CONJOINER
11F43..11F44;BA   # Po     [2] KAWI DANDA..KAWI DOUBLE DANDA
11F45..11F4F;ID   # Po    [11] KAWI PUNCTUATION SECTION MARKER..KAWI PUNCTUATION CLOSING SPIRAL
11F50..11F59;NU→AS# Nd    [10] KAWI DIGIT ZERO..KAWI DIGIT NINE
```

Line breaking for Makasar

The Unicode Standard, section 17.8, Makasar, says “Line breaks normally appear after syllable boundaries. Hyphens or other marks indicating continuance are not used.”

The changes to the Line_Break property shown in red are proposed:

```
11EE0..11EF1;AL→AS# Lo    [18] MAKASAR LETTER KA..MAKASAR LETTER A
11EF2;AL→BA       # Lo         MAKASAR ANGKA
11EF3..11EF4;CM   # Mn     [2] MAKASAR VOWEL SIGN I..MAKASAR VOWEL SIGN U
11EF5..11EF6;CM   # Mc     [2] MAKASAR VOWEL SIGN E..MAKASAR VOWEL SIGN O
11EF7..11EF8;AL→BA# Po     [2] MAKASAR PASSIMBANG..MAKASAR END OF SECTION
```

Line breaking for Tulu-Tigalari

The Tulu-Tigalari script has been accepted for encoding in a future version of the Unicode Standard. The Updated proposal to encode the Tulu-Tigalari script in Unicode is silent on line breaking. According to Vaishnavi Murthy Yerkadithaya, in traditional writing line breaks can occur even within orthographic syllables: between the pre-base vowels *-ee* or *-ai* and orthographic syllable cores, between orthographic syllable cores and post-base vowels *-aa* or *-au length mark*, except when they're part of canonical decompositions of independent vowels, and before bindus and visargas. Hyphens are not used in these cases. Line breaks can also occur before or after punctuation marks such as dandas. In modern writing, line breaks may occur at orthographic syllable boundaries or at word boundaries; hyphens may be used. Word spaces were not used in palm leaf manuscripts and stone inscriptions, but have appeared in later paper manuscripts.

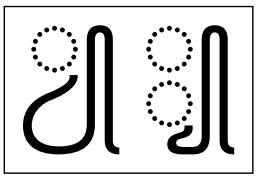
As discussed above, this proposal avoids breaking within orthographic syllables; it also does not insert hyphens.

The following Line_Break property data is proposed:

```
11380..11389;AS   # Lo    [10] TULU-TIGALARI LETTER A..TULU-TIGALARI LETTER VOCALIC LL
1138B;AS          # Lo         TULU-TIGALARI LETTER EE
1138E;AS          # Lo         TULU-TIGALARI LETTER AI
11390..11391;AS   # Lo     [2] TULU-TIGALARI LETTER OO..TULU-TIGALARI LETTER AU
11392..113B5;AK   # Lo    [36] TULU-TIGALARI LETTER KA..TULU-TIGALARI LETTER LLLA
113B7;ID          # Lo         TULU-TIGALARI SIGN AVAGRAHA
113B8..113BA;CM   # Mc     [3] TULU-TIGALARI VOWEL SIGN AA..TULU-TIGALARI VOWEL SIGN II
113BB..113C0;CM   # Mn     [6] TULU-TIGALARI VOWEL SIGN U..TULU-TIGALARI VOWEL SIGN VOCALIC LL
113C2;CM          # Mc         TULU-TIGALARI VOWEL SIGN EE
113C5;CM          # Mc         TULU-TIGALARI VOWEL SIGN EE
113C7..113CA;CM   # Mc     [4] TULU-TIGALARI VOWEL SIGN OO..TULU-TIGALARI SIGN CANDRA ANUNASIKA
113CC..113CD;CM   # Mc     [2] TULU-TIGALARI SIGN ANUSVARA..TULU-TIGALARI SIGN VISARGA
113CE;CM          # Mn         TULU-TIGALARI SIGN VIRAMA
113CF;CM          # Mc         TULU-TIGALARI SIGN LOOPED VIRAMA
113D0;VI          # Mn         TULU-TIGALARI CONJOINER
113D1;AP          # Lo         TULU-TIGALARI REPHA
113D4..113D5;ID   # Po     [2] TULU-TIGALARI DANDA..TULU-TIGALARI DOUBLE DANDA
113D7..113D8;ID   # Po     [2] TULU-TIGALARI SIGN OM PUSHPIKA..TULU-TIGALARI SIGN SHRII PUSHPIKA
113E1..113E2;CM   # Mn     [2] TULU-TIGALARI VEDIC TONE SVARITA..TULU-TIGALARI VEDIC TONE ANUDATTA
```

Enabling the use of dotted circle as a placeholder for subjoined consonants

The character DOTTED CIRCLE, U+25CC, is currently classified as AL (Alphabetic). There doesn't seem to be a strong reason for this classification, as DOTTED CIRCLE is not used as part of a word. Instead, DOTTED CIRCLE is commonly used as a base character placeholder to show combining marks or conjunct forms. It is also occasionally used as a placeholder for subjoined consonants to show how combining marks or conjunct forms change when attached to subjoined consonants (the first use of this may have been in Ida Bagus Adi Sudewa: The Balinese Alphabet). Reviewers have called out that DOTTED CIRCLE should not be separated from surrounding quotation marks; however, this is already assured by quotation marks using line break class QU and rule LB19, which prevents breaks before and after characters with that class.



To support this use, and prevent accidental line breaks within such arrangements, DOTTED CIRCLE should be reclassified as AK (Aksara).

```
25CC;AL→AK          # So          DOTTED CIRCLE
25CD;AL             # So          CIRCLE WITH VERTICAL FILL
```

Note that Microsoft updated the Universal Shaping Engine in 2019 to treat DOTTED CIRCLE as a consonant in order to support this use.

Changes from previous version

The following significant changes have been made since the original version L2/22-080 of 2022-03-28:

- Split out the proposed changes to The Unicode Standard, Core Specification, as L2/22-086 Specification updates for orthographic syllables and line breaking.
- Separated conjoining and visible viramas in the first draft regular expression for orthographic syllables.
- Removed support for Myanmar kinzi marks from the second draft regular expression and the rule LB28b, as Myanmar is not one of the scripts targeted with this proposal.
- Corrected handling of final viramas in the second draft regular expression.
- Added short and long name definitions in PropertyValueAliases.txt for the new line break classes.
- Completed lists of characters in the new line break classes.
- Clarified that the rule LB28b doesn't need to mention combining marks because of rule LB9.

- Added examples demonstrating the operation of the Unicode line breaking algorithm to detect orthographic syllable boundaries.
- Added a note that the impact of the Script_Extensions property needs to be considered when specifying line break data, and discussed sharing of characters between scripts in the sections on Grantha and Javanese.
- Added a reference to a comparison of current and proposed line breaking in Grantha.
- Showed changes in proposed Kawi line break data relative to that proposed in PRI 442 feedback for Unicode 15.
- Added discussion of quotation marks around dotted circle.
- Clarified text and added links to referenced Unicode data files.

Acknowledgments

I'd like to thank Aditya Bayu Perdana, Andrew Cunningham, Andy Heninger, Ben Yang, Elika Etemad, Ken Whistler, Martin Hosken, Muthu Nedumaran, Richard Ishida, Vaishnavi Murthy Yerkadithaya, and members of the Unicode Script Ad Hoc for providing feedback on drafts of this proposal. 𑍐