



# Required conjunct forms in extended grapheme clusters

Norbert Lindenberg, 2023-07-05

## Proposal

This document proposes to update the definition of extended grapheme clusters in [UAX 29, Unicode Text Segmentation](#), to include required conjunct forms that Unicode represents by a sequence of a virama-like character followed by a consonant or independent vowel, and to tie Myanmar *kinzi* glyphs to the base characters they belong to.

Grapheme cluster breaks should be prevented within the following character sequences:

- Balinese: U+1B44  BALINESE ADEG ADEG followed by any Balinese consonant or independent vowel.
- Chakma: U+11133 CHAKMA VIRAMA followed by any Chakma consonant.
- Dives Akuru: U+1193E DIVES AKURU VIRAMA followed by any Dives Akuru consonant.
- Javanese: U+A9C0  JAVANESE PANGKON followed by any Javanese consonant or independent vowel.
- Kawi: U+11F42 KAWI CONJOINER followed by any Kawi consonant or independent vowel.
- Kharoshthi: U+10A3F KHAROSHTHI VIRAMA followed by any Kharoshthi consonant.
- Khmer: U+17D2 KHMER SIGN COENG, optionally followed by U+200D ZERO WIDTH JOINER, followed by any Khmer consonant or independent vowel.
- Myanmar: U+1039 MYANMAR SIGN VIRAMA followed by any Myanmar consonant.
- Soyombo: U+11A99 SOYOMBO SUBJOINER followed by any Soyombo consonant.
- Sundanese: U+1BAB SUNDANESE SIGN VIRAMA followed by any Sundanese consonant.
- Tai Tham: U+1A60 TAI THAM SIGN SAKOT followed by any Tai Tham consonant.
- Tulu-Tigalari (accepted for Unicode 16.0): U+11D30 TULU-TIGALARI CONJOINER followed by any Tulu-Tigalari consonant.
- Zanabazar Square: U+11A47 ZANABAZAR SQUARE SUBJOINER followed by any Zanabazar Square consonant.

“Consonant” and “independent vowel” in this context means any character with an Indic syllabic category (InSC) value of Consonant or Vowel\_Independent. The character U+25CC ○ DOTTED CIRCLE, which can be used as a consonant placeholder in any script, should be included. “Virama-like character” means any character with an Indic syllabic category value of Virama or Invisible\_Stacker; characters with InSC= Pure\_Killer are excluded.

## Required conjunct forms

This proposal addresses conjunct forms that are “required” in the sense that rendering the underlying character sequence as a conjunct form is the standard behavior, depending neither on context nor on choices made by a type designer. For most of the scripts covered by this proposal, the Unicode Standard invented a special character whose only purpose is be used in the representation of conjunct forms; such characters have Indic syllabic category Invisible\_Stacker, which implies that they should never be shown by themselves. In Balinese and Javanese, the Unicode Standard uses a character that can either combine with the following character to a conjunct form or remain visible; such characters have Indic syllabic category Virama.

Users often perceive conjunct forms themselves as base-level characters, and have names for them, such as *coeng ta* for the conjunct form ࣚ of ត *ta* in Khmer, or *gantungan ka* for the conjunct form ࣚ of ក *ka* in Balinese. The following table compares the user-perceived characters with extended grapheme clusters according to today’s specification and the proposed version. Note in particular the “្រ” in Khmer, a character that Unicode invented and that does not exist in normal written Khmer.

	Khmer	Balinese
Text	ស្រី	ក្រ
User-perceived base-level characters	ស ្រ ី	ក ្រ ី
Extended grapheme clusters today	ស្រ ្រី	ក្រ ្រី
Extended grapheme clusters proposed	ស្រី	ក្រ

Conjunct forms in Brahmic scripts are generally derived from consonants; in some scripts, however, conjunct forms also exist for some characters classified as independent vowels, especially vocalic liquids. For such scripts, this proposal allows all independent vowels, as the precise set for which conjunct forms exist is not always known.

In Khmer, the character U+200D ZERO WIDTH JOINER has been proposed as a marker for final *coengs* (see [L2/22-290](#)) and is therefore allowed within the character sequence.

For this proposal only required conjunct forms are considered where the virama-like character combines with a subsequent character (as opposed to a preceding one) to create a conjunct form, as for these combinations the current definition of extended grapheme clusters breaks the conjunct forms.

The virama-like characters covered in this proposal were selected as follows:

- For characters with InSC=Invisible\_Stacker, it was assumed that conjunct form creation is always required, as these characters are not meant to be displayed by themselves. If such a character occurs without a character with which it can combine, the text is incorrect, and the result of segmentation may be arbitrary. The only question is whether the character must combine with a subsequent character (as opposed to only with a preceding character). The block descriptions in the Unicode Standard or script proposals make clear that this is always the case for the scripts proposed, except for the Myanmar *kinzi*, which is discussed below. In the Masaram Gondi and Gunjala Gondi scripts, the virama produces a half form of the preceding consonant, so it is not required that it be kept together with the subsequent consonant. No adequate information could be found for Meetei Mayek.
- For characters with InSC=Virama, a script-by-script analysis is necessary to determine whether conjunct form creation is required, contextual, or discretionary, whether the virama combines with subsequent or only with preceding characters, and which role zero width joiners might play. The proposal therefore includes only the two scripts with which the author is sufficiently familiar to perform this analysis, Balinese and Javanese. Experts in other scripts are invited to submit their own proposals.

A Myanmar *kinzi* is a special conjunct form that is encoded as a three-character sequence <Consonant, Pure\_Killer, Invisible\_Stacker> before the consonant on top of which it should be displayed. In this case the Invisible\_Stacker combines with two preceding characters. However, it should not be separated from the base consonant on top of which it is displayed, so preventing grapheme cluster breaks between the *kinzi* and the subsequent consonant is still correct.

## Compatibility with normalization

UAX 29 states as a goal that segmentation of text in normalization form D should produce the same results as segmentation of canonically equivalent unnormalized text. Unfortunately, normalization can in some cases produce results that are technically canonical equivalent, but not actually equivalent in the eyes of the user because they break conjunct formation.

One such case occurs in Khmer when a final *coeng* (conjunct form) occurs after the diacritic U+17DD ្ណ KHMER SIGN ATTHACAN, which has canonical combining class 230. The *coeng* is encoded as a sequence of U+17D2 ្គ KHMER SIGN COENG followed by a consonant or independent vowel. U+17D2 as a virama-like character has ccc=9, so when normalization encounters a sequence <17DD, 17D2>, it reorders to <17D2, 17DD>. This breaks the two-character sequence representing the *coeng*, so the SIGN COENG (which is not supposed to be rendered) and the subsequent consonant or independent vowel are rendered separately. Instead of ្ណ្គ, the user sees ្គ្ណ, unless the font compensates for the reordering.

This problem can occur in Brahmic scripts in which viramas combine with subsequent consonants to create conjunct forms, that have combining marks with canonical combining classes above 9, and in which such a combining mark is allowed to occur immediately before the virama. It is then possible that normalization reorders a virama away from the consonant or independent vowel with which it is supposed to combine, so that the conjunct form is no longer created.<sup>1</sup>

This problem can not be fixed within the current normalization forms because of the [Normalization Stability](#) policy.

To maintain equivalent results between unnormalized and normalized input in cluster segmentation, we have to account for characters that can occur before the virama in unnormalized text and can be reordered after the virama as part of normalization. Let's call these characters *swappable*.

In order to be swappable, a character must have a canonical combining class above 9. The following scripts in this proposal have characters with such combining classes: Balinese, Chakma, Kharoshthi, Khmer, Myanmar, and Tai Tham. However, the ones in Balinese are intended only for use with musical symbols, not in orthographic syllables, so we can ignore them.

---

<sup>1</sup> Additional cases might occur in scripts in which viramas combine with preceding consonants to create conjunct forms and that have combining marks with ccc values below 9. I'm less familiar with these scripts, and such scripts are not relevant to this proposal, so I'll leave it at "might".

In order to be swappable, a character also must be allowed to occur directly before the virama according to the encoding order for orthographic syllables in its script. Unfortunately, the Unicode Standard does not define complete and coherent encoding orders for any of these scripts, so we have to look at other sources.

Chakma, Kharoshthi, and Tai Tham are supported by the [OpenType Universal Shaping Engine](#), which provides a generic encoding order that covers the first two, but not Tai Tham. The encoding order for Chakma does not allow the combining marks with `ccc>9` to occur before the virama, so they're not swappable. For Kharoshthi, two combining marks with `ccc>9` can occur before the virama and are therefore swappable: U+10A38 𑀭 KHAROSHTHI SIGN BAR ABOVE and U+10A3A 𑀮 KHAROSHTHI SIGN DOT BELOW. For Tai Tham, because there is no agreed-upon encoding order yet, and in particular because the script has final (post-vowel) conjunct forms, we have to assume that all nine combining marks with `ccc>9` are swappable.

For Khmer, the encoding order documented in section 16.4 Khmer of the Unicode Standard 15.0 makes U+17DD 𑀓 KHMER SIGN ATTHACAN swappable, as described above. (The encoding order developed in [L2/22-290](#) avoids this problem by changing the encoding of final *coengs*, but is not a standard yet.)

Myanmar has only one combining mark with `ccc>9`, and this character cannot occur before the virama according to the standard reference for Myanmar encoding order, [UTN 11](#).

The characters identified as swappable are listed in the **Swappable** macro in the next section.

## Integration into UAX 29

The [current proposed update to UAX 29](#) adds support for conjunct forms in six Brahmic scripts. It uses character class definitions based on properties other than `Grapheme_Cluster_Break`, adding flexibility to the existing specification mechanism. At its core is the new rule GB9c:

LinkingConsonant ExtCccZwj\* ConjunctLinker ExtCccZwj\* × LinkingConsonant

In this rule, `LinkingConsonant` means the consonants and `ConjunctLinker` the viramas of the six supported Brahmic scripts. `ExtCccZwj` is a set of characters that may occur in between; my doubts about this set are recorded in [feedback on PRI 469](#).

The need for the `LinkingConsonant` and `ExtCccZwj` before `ConjunctLinker` isn't explained in the proposed update; they may reflect that in North-Indian scripts viramas tend to create half-forms of

the preceding consonant before including subsequent consonants into conjuncts. For the scripts covered in this proposal, there's no need for them. The ExtCccZwj between ConjunctLinker and LinkingConsonant allows too many characters for the scripts covered in this proposal; no character other than ZWJ or swappables should ever occur here. On the other hand, ConjunctLinker needs to allow for InSC= Invisible\_Stacker, and LinkingConsonant for InSC=Vowel\_Independent for some scripts.

Rather than trying to merge the scripts covered in this proposal into rule GB9c, it is therefore safer to support them with a new rule GB9d:

### **ConjunctLinker2 SwappableZWJ\* × LinkingBase2**

where:

- **ConjunctLinkingScripts2C**=[\p{sc=Cakm}\p{sc=Diak}\p{sc=Khar}\p{sc=Mymr}\p{sc=Soyo}\p{sc=Sund}\p{sc=Lana}\p{sc=Tuti}\p{sc=Zanb}]
- **ConjunctLinkingScripts2CV**=[\p{sc=Bali}\p{sc=Java}\p{sc=Kawi}\p{sc=Khmr}]
- **ConjunctLinkingScripts2**=[**ConjunctLinkingScripts2C**||**ConjunctLinkingScripts2CV**]
- **ConjunctLinker2**=[**ConjunctLinkingScripts2**&&[\p{Indic\_Syllabic\_Category=Virama}\p{Indic\_Syllabic\_Category=Invisible\_Stacker}]]
- **ZWJ**=[\u200D]
- **Swappable**=[\u{10A38}\u{10A3A}\u{17DD}\u{1A75}-\u{1A7C}\u{1A7F}]
- **SwappableZWJ**=[**Swappable**||**ZWJ**]
- **LinkingBase2C**=[**ConjunctLinkingScripts2C**&&[\p{Indic\_Syllabic\_Category=Consonant}]]
- **LinkingBase2CV**=[**ConjunctLinkingScripts2CV**&&[\p{Indic\_Syllabic\_Category=Consonant}\p{Indic\_Syllabic\_Category=Vowel\_Independent}]]
- **LinkingBase2**=[**LinkingBase2C**||**LinkingBase2CV**][[○]]

Note that the above assumes that the Tulu-Tigalari script will get “Tuti” as its ISO 15924 script code – this may have to be corrected after a script code has been assigned.

As in the proposed update, character sets from different scripts are lumped together, which is permissible because UAX 29, section 1.2 Rule Constraints, does not require breaking at script boundaries.

## Feedback on earlier proposals

This is not the first time better support for conjunct forms or complete orthographic syllables has been proposed. An early proposal covering the Myanmar and Khmer scripts was made in 2005 in [L2/05-352](#). In 2017, a more comprehensive proposal evolved in [L2/17-167R](#), [L2/17-258](#), and [L2/17-222](#), but was delegated to CLDR after [feedback on PRI 355](#), and [L2/18-055](#). Several issues were noted in that process, of which I only cite those that might apply to the scripts proposed here:

- [L2/17-258](#) excluded U+1B44 ꦱ BALINESE ADEG ADEG, U+A9C0 ꦲ JAVANESE PANGKON and other characters with general category Spacing\_Mark from the list of supported viramas because they “would introduce complications”. However, it did not explain the nature of these “complications”.
- In [PRI 355 feedback](#), one commenter noted that in some scripts characters with InSC=Virama are more often shown as visible viramas than forming conjuncts. While this may be true for some scripts, it is not true for Balinese and Javanese.
- The same commenter argued that grapheme clusters should end after U+200D ZERO WIDTH JOINER because “ZWJ is generally used in Indic as an invisible letter”. He did not explain what is meant by “used as an invisible letter”. In general, the purpose of ZWJ varies between scripts; in the use proposed for Khmer it certainly should not disrupt conjunct formation, and so should not end the grapheme cluster.
- The same commenter argued that all breaks should be prevented after characters with InSC=Invisible\_Stacker. This would add Gunjala Gondi, Masaram Gondi, and Meetei Mayek to the set of supported scripts.
- Another commenter notes that preventing breaks after viramas results in longer grapheme clusters, making editing them more cumbersome. This is true; however, the appropriate solution for this is not to break within the character sequence encoding a required conjunct forms, but to find other places to break within the cluster, such as before spacing marks. But first it needs to be decided whether editing or safe rendering is the primary use case for default grapheme clusters (see [L2/23-140](#)).

Neither of the issues raised should prevent this proposal from moving forward.

## Acknowledgments

I’d like to thank Martin Hosken for feedback on drafts of this proposal.