

More ALetters: A Report on UTC 162-A2

Roozbeh Pournader

2026-09-01

During UTC meeting #162, I was assigned an action item to "Provide a document on potential addition of more modifier letters to "ALetter" in Table 3 of UAX #29."

I researched the topic, and here propose some more characters to be added to ALetter. Here is the proposal:

- Change the Word_Break property of the following characters from Other to ALetter:
 - U+02D9 DOT ABOVE
 - U+0888 ARABIC RAISED ROUND DOT
 - U+1807 MONGOLIAN SIBE SYLLABLE BOUNDARY MARKER
 - U+180A MONGOLIAN NIRUGU
 - U+2E32 TURNED COMMA
 - U+2E35..U+2E39 TURNED SEMICOLON..TOP HALF SECTION SIGN
 - U+A700..U+A707 MODIFIER LETTER CHINESE TONE YIN PING..MODIFIER LETTER CHINESE TONE YANG RU
 - U+AB6A MODIFIER LETTER LEFT TACK and U+AB6B MODIFIER LETTER RIGHT TACK
 - U+10EC9 ARABIC SMALL BASELINE FATHA and U+10ECA ARABIC SMALL BASELINE DOTLESS HEAD OF KHAH

The earlier additions to ALetter were selected primarily because they occur adjacent to letters in IPA, UPA, and other phonetic transcription systems. [L2/16-336](#) explicitly describes them as modifier symbols used chiefly in phonetic transcription and recommends changing only Word_Break, avoiding collateral changes to Alphabetic, identifiers, and other properties.

What prompted this action

The immediate answer intended by UTC action item 162-A2 was almost certainly U+A700..U+A707. The PRI #396 feedback says explicitly that these characters are used

like the other Chinese tone letters and cites examples with Latin/IPA letters as well as Han characters.

The research

The criterion I used in making the recommendations was “Does the character function as part of a word or transcription, such that the default word-boundary algorithm should not break between it and adjacent letters?” which is similar to what [L2/16-336](#) used. An audit of the current repertoire produced the following results:

- Almost every character with General_Category=Lm is already ALetter, indirectly through Alphabetic=Yes, as UAX #29 explains.
- In Unicode 18, the only characters whose names begin with MODIFIER LETTER and which still have Word_Break=Other are:
 - U+A700..U+A707: Traditional Chinese corner tone marks used in linguistic and dialectological transcription. The encoding proposal, [L2/04-107](#), shows them with IPA letters, other phonetic marks, and Han characters. They are directly analogous to U+A708..U+A716, which are already ALetter. Assigning ALetter to these will improve their behavior with Latin and IPA letters. It will not affect them when they are next to Han characters, because the default UAX #29 segmentation breaks around ideographs. (A decision about these was deferred in 2016, pending further research.)
 - U+AB6A and U+AB6B: Postposed symbols for fronted and backed vowel realizations in Scots phonetic transcription, attested in the *Scottish National Dictionary* and *Linguistic Atlas of Scotland*. They were proposed in [L2/19-075R](#) and were encoded after the L2/16-336 repertoire review.
- One additional gc=Sk character, U+02D9 DOT ABOVE, has essentially the same orthographic behavior but does not have “MODIFIER LETTER” in its name. This is the Bopomofo neutral-tone mark. The other four Bopomofo tone marks are gc=Lm and already ALetter. Taiwan’s official manual for Bopomofo (https://language.moe.gov.tw/001/Upload/files/site_content/M0001/juyin/html_en/index.html) places the neutral-tone dot before the Bopomofo syllable in both horizontal and vertical text, so the present value of Word_Break=Other produces a word break between the tone

mark and its syllable. (This is the right code point for the tone mark, as described in [Table 18-9 of Core Spec 17.0](#).)

The remaining gc=Lm characters outside ALetter are intentional special cases. They should retain their existing treatment.

Additional candidates beyond the 2016 audit

The 2016 review in [L2/16-336](#) examined only General_Category=Sk. I looked at some more characters too. I'm also proposing the following characters to be made ALetter:

- U+2E32 TURNED COMMA and U+2E35..U+2E39 (Palaeotype marks): These are spacing phonetic modifiers, used for nasalization, glottal closure, retraction, advancement, retroflexion, and unilateral pronunciation. The proposal, [L2/09-425](#), gives examples such as a,, bɿ, dɿ, and dɿ̣. They were encoded as gc=Po, so L2/16-336 never considered them.
- U+0888 ARABIC RAISED ROUND DOT: It always follows an alef-like character and basically acts like a spacing combining mark in some Quranic orthographies (see the proposal, [L2/19-306](#)). Its baseline counterpart U+0887 is a letter and consequently already receives ALetter; U+0888 is currently an accidental word-break outlier because it was encoded as gc=Sk.
- U+10EC9 ARABIC SMALL BASELINE FATHA and U+10ECA ARABIC SMALL BASELINE DOTLESS HEAD OF KHAH: These Unicode 18 characters were originally proposed as gc=Lo, which would have made them Alphabetic and automatically ALetter. We assigned them gc=Sk instead because we believed their properties should parallel U+0888. They should become ALetter like U+0888 should.
- U+180A MONGOLIAN NIRUGU: Its main use is to extend the stem connecting letters, basically a Mongolian *tatweel*. Its principal function therefore should not produce a word boundary.
- U+1807 MONGOLIAN SIBE SYLLABLE BOUNDARY MARKER: It explicitly marks a syllable boundary within a word. This could theoretically be a MidLetter, but MidLetter appears restricted to characters that become punctuation at beginning or end of words, and there's no evidence that this character is used in such places.

Editorial corrections

The current UAX #29 categorically says that modifier symbols (gc=Sk) are not ALetter, even though Table 3 already lists many Sk exceptions. I suggest replacing that passage with:

Modifier letters (General_Category=Lm) are almost all included in ALetter by virtue of their Alphabetic property. Most modifier symbols (General_Category=Sk) are not ALetter; However, some modifier symbols that function as components of orthographic words or phonetic transcriptions are explicitly assigned ALetter.

The corresponding blanket statement in Core Specification Section 7.8, "General Category", should also be qualified in the same way. In Core Spec 17.0, the text reads:

Modifier letters (gc = Lm) also have the word break property value (wb = ALetter), while modifier symbols (gc = Sk) do not. This means that for default determination of word break boundaries, modifier symbols will cause a word break, while modifier letters proper will not.

It should be replaced with something like:

Most modifier letters (gc = Lm) have the word break property value wb = ALetter, while most modifier symbols (gc = Sk) do not. However, selected modifier symbols that function as components of orthographic words or phonetic transcriptions are explicitly assigned wb = ALetter, and some modifier letters have other Word_Break values. See Unicode Standard Annex #29, "Unicode Text Segmentation".