L2/16-336

Title: **Candidate Characters for Word_Break = ALetter**
Source: Mark Davis (Google, Inc.), Laurențiu Iancu (Microsoft Corporation), and Ken Whistler
Status: Individual contribution
Action: For consideration by the Unicode Technical Committee
Date: 2016-11-03

Proposal

The proposal is to assign the property value Word_Break = ALetter to a set of characters identified in this document, to prevent word boundaries between those characters and other alphabetic letters. The set consists of 35 characters as of Unicode 9.0, listed in Table 1.

Background

At the UTC meeting #148, the following feedback report [L2/16-205] was discussed:

> Date/Time: Thu Jun 2 08:46:28 CDT 2016
> Name: David Corbett
> Report Type: Error Report
> Opt Subject: Word_Break of U+02D7
>
> U+02D7 MODIFIER LETTER MINUS SIGN's Word_Break is MidLetter but should be ALetter. Like other IPA modifier letters, it follows the letter it modifies; it does not need a letter after it. Maybe its General_Category should be changed to Modifier_Letter to match U+02D0 MODIFIER LETTER TRIANGULAR COLON.

And an action item was assigned:

> [148-A41a] Action Item for Laurențiu Iancu: Propose which gc = Sk characters to include in ALetter (or to add to Alphabetic).

This document analyzes the set of characters with General_Category = Modifier_Symbol (gc = Sk) and proposes a subset of candidates to be assigned the property value Word_Break = ALetter (WB = LE), to prevent word boundaries between those characters and adjacent alphabetic letters.

Character selection criteria

Several criteria were used to rationalize the selection of candidates from the 121 characters with General_Category = Modifier_Symbol as of Unicode 9.0. These criteria are outlined below and are exclusion criteria, so the characters that meet them are deemed unsuitable to be absorbed in the set of Word_Break = ALetter.

a. Spacing clones of diacritics, including common, Greek specific, and fullwidth clones.
b. Symbols used with CJK and other nonalphabetic scripts, such as script specific tone marks used with Bopomofo.
c. Arabic pedagogical symbols.
d. Spacing kana voicing marks and emoji modifiers. These have the Word_Break property values Katakana and, respectively, E_Modifier as of Unicode 9.0.
e. Contour tone marks. These are listed separately, to facilitate the examination of this criterion by the UTC.

Using these criteria, the set of 121 characters with General_Category = Modifier_Symbol as of Unicode 9.0 can be partitioned into three subsets as follows:

1. Characters that qualify for being included in the set Word_Break = ALetter. These are diacritic-like spacing modifier symbols that occur almost exclusively in phonetic transcription systems such as IPA or UPA. There are 35 proposed characters, listed in Table 1.
2. Characters excluded per criteria (a) through (d) outlined above. These are more clearly excluded due to their nature or specialized use. There are 58 symbols in this set, listed in Table 2.
3. Contour tone marks. These are also proposed to be excluded, per criterion (e), but are listed as a separate category in Table 3, pending UTC discussion. There are 28 characters in this set.

Proposed property change

The 35 characters identified in Table 1 as qualifying for a property change all have Word_Break = Other as of Unicode 9.0, except for U+02D7 MODIFIER LETTER MINUS SIGN, which is Word_Break = MidLetter. Their other properties, as of Unicode 9.0, are General_Category = Modifier_Symbol, Alphabetic = No, Ideographic = No, Script = Common, and Sentence_Break = Other. They are also all Line_Break = Alphabetic, with the exception of U+02DF MODIFIER LETTER CROSS ACCENT, which is Line_Break = Break_Before.

Given the derivation expression for the property value Word_Break = ALetter [UAX #29],

> Alphabetic = Yes, or
> U+05F3 HEBREW PUNCTUATION GERESH
> and Ideographic = No
> and Word_Break ≠ Katakana
> and Line_Break ≠ Complex_Context (SA)
> and Script ≠ Hiragana
> and Word_Break ≠ Extend
> and Word_Break ≠ Hebrew_Letter,

there are three ways in which the proposed characters can become ALetter:

1. Directly assign them Word_Break = ALetter, while keeping them Alphabetic = No, similar to the explicit handling of U+05F3 HEBREW PUNCTUATION GERESH.
2. Assign them the contributory property Other_Alphabetic = Yes. This causes them to become Alphabetic, which by derivation changes their status for Word_Break. It also affects other properties and algorithms that depend upon Alphabetic (including Sentence_Break).
3. Change their General_Category property value from Modifier_Symbol (gc = Sk) to Modifier_Letter (gc = Lm). This change also causes them to become Alphabetic. It also affects other properties and algorithms that are sensitive to the difference between gc = Lm and gc = Sk.

For option #3, a change in General_Category can be potentially disruptive, because it affects identifiers and collation (unless those are also modified to compensate).

For option #2, there are no Modifier_Symbol characters in the Alphabetic class as of Unicode 9.0. This may or may not be a concern, but is noted here for creating a new partial overlap between two sets of characters that used to be disjoint.

Therefore, it seems prudent to accomplish a Word_Break property value of ALetter directly by addition to Word_Break, using option #1, rather than indirectly through a change in Alphabetic, either via Other_Alphabetic (option #2) or via General_Category (option #3).

For these reasons, the proposed approach, pending review and recommendations by the UTC, is to include the characters listed in Table 1 directly into the definition expression for Word_Break = ALetter.

Regarding the effects on sentence segmentation, assigning Word_Break = ALetter to the proposed characters (without changing their Alphabetic classification) would not affect the relation between Word_Break and Sentence_Break. That is because all 121 characters with General_Category = Modifier_Symbol are, and remain, Sentence_Break = Other, so sentence boundaries are prohibited at those characters. Therefore, the additional prevention of word boundaries would not result in the formation of word segments straddling a sentence boundary (which would reintroduce a "Mr.Hamster" problem [L2/15-068].)

Examples

To illustrate the proposed change with an example using the character in the original feedback report, U+02D7 (˗) MODIFIER LETTER MINUS SIGN, which is Word_Break = MidLetter in Unicode 9.0, the character sequence "˗a b˗c d˗ ˗" <02D7 0061 0020 0062 02D7 0063 0020 0064 02D7 0020 02D7> would have the following word boundaries:

Before the change (as of Unicode 9.0): |˗|a| |b ˗ c| |d|˗| |˗|
After the proposed change: |˗ a| |b ˗ c| |d ˗| |˗|

The word boundaries that are eliminated by the proposed change are shown in red. The word boundaries that are not affected are shown in blue.

Similarly, using one of the proposed characters which have Word_Break = Other as of Unicode 9.0, such as U+02D6 (˖) MODIFIER LETTER PLUS SIGN, the sequence "˖a b˖c d˖ ˖" <02D6 0061 0020 0062 02D6 0063 0020 0064 02D6 0020 02D6> would have the following word boundaries:

Before the change (as of Unicode 9.0): |˖|a| |b|˖|c| |d|˖| |˖|
After the proposed change: |˖ a| |b ˖ c| |d ˖| |˖|

Character sets

Table 1 below lists the characters proposed for inclusion in the set Word_Break = ALetter. Table 2 and Table 3 list the remaining characters from the set General_Category = Modifier_Symbol.

Table 1: Modifier_Symbol characters proposed for inclusion in the set Word_Break = ALetter

Code	Character name	Notes
02C2	MODIFIER LETTER LEFT ARROWHEAD	Phonetic modifiers
02C3	MODIFIER LETTER RIGHT ARROWHEAD	
02C4	MODIFIER LETTER UP ARROWHEAD	
02C5	MODIFIER LETTER DOWN ARROWHEAD	
02D2	MODIFIER LETTER CENTRED RIGHT HALF RING	
02D3	MODIFIER LETTER CENTRED LEFT HALF RING	
02D4	MODIFIER LETTER UP TACK	
02D5	MODIFIER LETTER DOWN TACK	
02D6	MODIFIER LETTER PLUS SIGN	
02D7	MODIFIER LETTER MINUS SIGN	Phonetic modifier: Change WB = MidLetter to WB = ALetter
02DE	MODIFIER LETTER RHOTIC HOOK	Phonetic modifier
02DF	MODIFIER LETTER CROSS ACCENT	Phonetic modifier, Line_Break = Break_Before
02ED	MODIFIER LETTER UNASPIRATED	Phonetic modifiers
02EF	MODIFIER LETTER LOW DOWN ARROWHEAD	
02F0	MODIFIER LETTER LOW UP ARROWHEAD	
02F1	MODIFIER LETTER LOW LEFT ARROWHEAD	
02F2	MODIFIER LETTER LOW RIGHT ARROWHEAD	
02F3	MODIFIER LETTER LOW RING	
02F4	MODIFIER LETTER MIDDLE GRAVE ACCENT	
02F5	MODIFIER LETTER MIDDLE DOUBLE GRAVE ACCENT	
02F6	MODIFIER LETTER MIDDLE DOUBLE ACUTE ACCENT	
02F7	MODIFIER LETTER LOW TILDE	
02F8	MODIFIER LETTER RAISED COLON	
02F9	MODIFIER LETTER BEGIN HIGH TONE	
02FA	MODIFIER LETTER END HIGH TONE	
02FB	MODIFIER LETTER BEGIN LOW TONE	
02FC	MODIFIER LETTER END LOW TONE	
02FD	MODIFIER LETTER SHELF	
02FE	MODIFIER LETTER OPEN SHELF	
02FF	MODIFIER LETTER LOW LEFT ARROW	
A720	MODIFIER LETTER STRESS AND HIGH TONE	Phonetic modifiers
A721	MODIFIER LETTER STRESS AND LOW TONE	
A789	MODIFIER LETTER COLON	Modifier letters [L2/06-259R]
A78A	MODIFIER LETTER SHORT EQUALS SIGN	

Code	Character name	Notes
AB5B	MODIFIER BREVE WITH INVERTED BREVE	Modifier letter [L2/11-202]

Table 2: Modifier_Symbol characters excluded from the proposed Word_Break reclassification, per the exclusion criteria (a) through (d)

Code	Character name	Notes
005E	CIRCUMFLEX ACCENT	Spacing clones of dia-critics
0060	GRAVE ACCENT	
00A8	DIAERESIS	
00AF	MACRON	
00B4	ACUTE ACCENT	
00B8	CEDILLA	
02D8	BREVE	Spacing clones of dia-critics; 02D9 also used with Bopomofo
02D9	DOT ABOVE	
02DA	RING ABOVE	
02DB	OGONEK	
02DC	SMALL TILDE	
02DD	DOUBLE ACUTE ACCENT	
02EA	MODIFIER LETTER YIN DEPARTING TONE MARK	Script specific marks used with Bopomofo
02EB	MODIFIER LETTER YANG DEPARTING TONE MARK	
0375	GREEK LOWER NUMERAL SIGN	Specialized use
0384	GREEK TONOS	Spacing clones of Greek diacritics
0385	GREEK DIALYTIKA TONOS	
1FBD	GREEK KORONIS	
1FBF	GREEK PSILI	
1FC0	GREEK PERISPOMENI	
1FC1	GREEK DIALYTIKA AND PERISPOMENI	
1FCD	GREEK PSILI AND VARIA	
1FCE	GREEK PSILI AND OXIA	
1FCF	GREEK PSILI AND PERISPOMENI	
1FDD	GREEK DASIA AND VARIA	
1FDE	GREEK DASIA AND OXIA	
1FDF	GREEK DASIA AND PERISPOMENI	
1FED	GREEK DIALYTIKA AND VARIA	
1FEE	GREEK DIALYTIKA AND OXIA	
1FEF	GREEK VARIA	
1FFD	GREEK OXIA	
1FFE	GREEK DASIA	
309B	KATAKANA-HIRAGANA VOICED SOUND MARK	Spacing, script specific, WB = Katakana
309C	KATAKANA-HIRAGANA SEMI-VOICED SOUND MARK	
FBB2	ARABIC SYMBOL DOT ABOVE	Arabic pedagogical symbols
FBB3	ARABIC SYMBOL DOT BELOW	

Code	Character name	Notes
FBB4	ARABIC SYMBOL TWO DOTS ABOVE	
FBB5	ARABIC SYMBOL TWO DOTS BELOW	
FBB6	ARABIC SYMBOL THREE DOTS ABOVE	
FBB7	ARABIC SYMBOL THREE DOTS BELOW	
FBB8	ARABIC SYMBOL THREE DOTS POINTING DOWNWARDS ABOVE	
FBB9	ARABIC SYMBOL THREE DOTS POINTING DOWNWARDS BELOW	
FBBA	ARABIC SYMBOL FOUR DOTS ABOVE	
FBBB	ARABIC SYMBOL FOUR DOTS BELOW	
FBBC	ARABIC SYMBOL DOUBLE VERTICAL BAR BELOW	
FBBD	ARABIC SYMBOL TWO DOTS VERTICALLY ABOVE	
FBBE	ARABIC SYMBOL TWO DOTS VERTICALLY BELOW	
FBBF	ARABIC SYMBOL RING	
FBC0	ARABIC SYMBOL SMALL TAH ABOVE	
FBC1	ARABIC SYMBOL SMALL TAH BELOW	
FF3E	FULLWIDTH CIRCUMFLEX ACCENT	Fullwidth spacing clones of diacritics
FF40	FULLWIDTH GRAVE ACCENT	
FFE3	FULLWIDTH MACRON	
1F3FB	EMOJI MODIFIER FITZPATRICK TYPE-1-2	WB = E_Modifier as of Unicode 9.0.
1F3FC	EMOJI MODIFIER FITZPATRICK TYPE-3	
1F3FD	EMOJI MODIFIER FITZPATRICK TYPE-4	
1F3FE	EMOJI MODIFIER FITZPATRICK TYPE-5	
1F3FF	EMOJI MODIFIER FITZPATRICK TYPE-6	

Table 3: Tone marks, excluded from the proposed Word_Break reclassification, per criterion (e), pending UTC discussion

Code	Character name	Notes
02E5	MODIFIER LETTER EXTRA-HIGH TONE BAR	Contour tone letters
02E6	MODIFIER LETTER HIGH TONE BAR	
02E7	MODIFIER LETTER MID TONE BAR	
02E8	MODIFIER LETTER LOW TONE BAR	
02E9	MODIFIER LETTER EXTRA-LOW TONE BAR	
A700	MODIFIER LETTER CHINESE TONE YIN PING	Tone marks, used with either IPA letters or CJK ideographs [L2/04-107]
A701	MODIFIER LETTER CHINESE TONE YANG PING	
A702	MODIFIER LETTER CHINESE TONE YIN SHANG	
A703	MODIFIER LETTER CHINESE TONE YANG SHANG	
A704	MODIFIER LETTER CHINESE TONE YIN QU	
A705	MODIFIER LETTER CHINESE TONE YANG QU	
A706	MODIFIER LETTER CHINESE TONE YIN RU	
A707	MODIFIER LETTER CHINESE TONE YANG RU	
A708	MODIFIER LETTER EXTRA-HIGH DOTTED TONE BAR	Tone letters [L2/04-107]
A709	MODIFIER LETTER HIGH DOTTED TONE BAR	
A70A	MODIFIER LETTER MID DOTTED TONE BAR	
A70B	MODIFIER LETTER LOW DOTTED TONE BAR	
A70C	MODIFIER LETTER EXTRA-LOW DOTTED TONE BAR	
A70D	MODIFIER LETTER EXTRA-HIGH DOTTED LEFT-STEM TONE BAR	
A70E	MODIFIER LETTER HIGH DOTTED LEFT-STEM TONE BAR	
A70F	MODIFIER LETTER MID DOTTED LEFT-STEM TONE BAR	
A710	MODIFIER LETTER LOW DOTTED LEFT-STEM TONE BAR	
A711	MODIFIER LETTER EXTRA-LOW DOTTED LEFT-STEM TONE BAR	
A712	MODIFIER LETTER EXTRA-HIGH LEFT-STEM TONE BAR	
A713	MODIFIER LETTER HIGH LEFT-STEM TONE BAR	
A714	MODIFIER LETTER MID LEFT-STEM TONE BAR	
A715	MODIFIER LETTER LOW LEFT-STEM TONE BAR	
A716	MODIFIER LETTER EXTRA-LOW LEFT-STEM TONE BAR	