L2/09-278

TO: Unicode Technical Committeee
FROM: Deborah Anderson, SEI, UC Berkeley
DATE: 5 August 2009
RE: On the proposed U+A78F LATIN LETTER MIDDLE DOT (L2/09-031R = N3567)

1. Background. In L2/09-031R (=N3567), Andrew West proposed a new Latin letter middle dot for the transliteration of PHAGS-PA LETTER SMALL A and to represent the glottal stop in Gong Hwang-cherng's phonetic representation of Tangut. The proposal author also mentioned that "it is quite probable" that it is used in transliteration and/or transcription of other languages.

Justification for a new middle dot was based on the fact that the existing middle dots are either marks of punctuation or are script-specific, and the following middle dots were cited:

00B7	MIDDLE DOT	Po	Common
1427	CANADIAN SYLLABICS FINAL MIDDLE DOT	Lo	Can Aboriginal
30FB	KATAKANA MIDDLE DOT	Po	Common
FF65	HALFWIDTH KATAKANA MIDDLE DOT	Po	Common

At the Dublin WG 2 meeting in April 2009, a concern was raised about the use of U+00B7 MIDDLE DOT for transliteration, because search engines such as Google are not able to find U+00B7.

In order to understand the problem more fully, I investigated the use of the middle dot in other languages and inquired into the search issue, and present the findings in this document.

2. Use of Middle Dot

While the use of the middle dot was raised within the context of transliteration (/transcription) of East Asian languages in L2/09-031R (N3567), the character has long been used by Americanists. Franz Boas already wrote in 1911 (*Handbook of American Indian Languages*, page 7):

> the raised period [=middle dot] is the recommended length mark to be used after a symbol to indicate that the sound is long. The colon was designated as marking excessive length, longer than that represented by <·>. For languages which make a contrast between only degrees of length, either the colon or the raised period may be used as the length mark. Where both occur, the colon represents greater length.

Examples of its use to designate length are given in figures 1–5 for Munsee, Central Sierra Miwok, Unami, Proto-Takelman, Tonkawa and Algonkian.

Hence, the middle dot is not restricted to East Asian use, and any new middle dot character will have an impact on materials for other languages, where the middle dot has been used at least

since the early twentieth century for transliteration and transcription, as well as in technical and practical orthographies. The middle dot has appeared – and continues to appear – in dictionaries, text corpora, and teaching materials for languages of the Americas.

For online documents, two characters have been identified as being used by Americanists: U+00B7 MIDDLE DOT (see Munsee and Unami references cited in the examples) and U+02D1 MODIFIER LETTER HALF TRIANGULAR COLON (see Klamath example on page 3 and Severn Ojibwe example in figure 6).

3. Search Issues for U+00B7 MIDDLE DOT

The character U+00B7 MIDDLE DOT, which has the property of Po, can be found in searches with Google, but the results are not consistent. For example, the Catalan word anul·lar (with U+00B7) is found in the search results at: http://www.google.com/search?q=anul%C2%B7lar. However, the same results occur with "anul lar" (with space between the l's) in: http://www.google.com/search?hl=en&q=anul+lar.

In order to get the proper result in searches in Google, the middle dot needs to be treated as a letter when it occurs between letters, but as a mark of punctuation when it is isolated. Mark Davis has filed a bug at Google on this, since the same treatment occurs for other marks as well, including ' and ’. (Such a correction will satisfy the problem as long as the middle dot is not next to a mark of punctuation, however, if one is looking for an exact match.) According to Peter Constable, Microsoft products also treat U+00B7 as punctuation, and most applications ignore it.

4. Options

Rather than encode another middle dot – for which the naïve user could pick from at least 23 middle dots (L2/07-258) – two characters are currently in use:

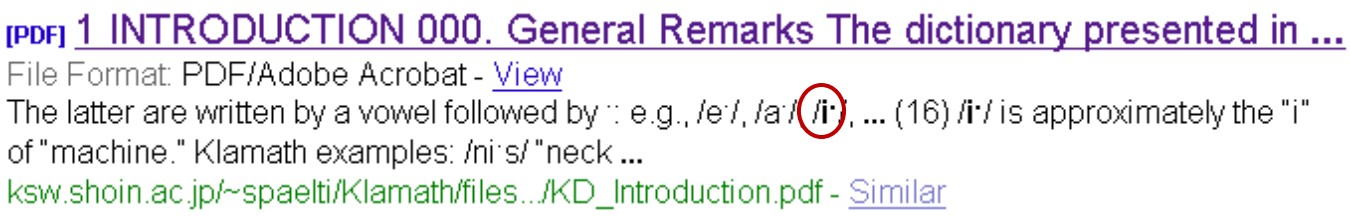
a. U+00B7 MIDDLE DOT

While this "generic" middle dot is found commonly in most fonts, it can be ignored by search engines, as noted above. However, future search engines may provide better support to find U+00B7 when it occurs between two letters, and hence improve search results. For linebreaking, U+00B7 is AI which, according to normal UAX #14 behavior, will be treated exactly as AL.

b. U+02D1 MODIFIER LETTER HALF TRIANGULAR COLON

The character U+02D1 has gc=Lm, its script is COMMON, and has the linebreaking property AL, and as a result provides the properties being requested for the proposed "LATIN LETTER MIDDLE DOT." U+02D1 was already part of Unicode 1.1, so it has well-established usage.

When doing a Google search for the combination lowercase I and U+02D1 ("iˑ"), the resulting search page can have a circular dot displayed for U+02D1, as in the following:


[PDF] 1 INTRODUCTION 000. General Remarks The dictionary presented in ...
File Format: PDF/Adobe Acrobat - View
The latter are written by a vowel followed by ˑ: e.g., /eˑ/, /aˑ/, /iˑ/, ... (16) /iˑ/ is approximately the "i" of "machine." Klamath examples: /niˑs/ "neck ...
ksw.shoin.ac.jp/~spaelti/Klamath/files.../KD_Introduction.pdf - Similar

For this example, the page it links to has the triangular shape for U+02D1:

(16) /iˑ/ is approximately the "i" of "machine."

(Source: http://ksw.shoin.ac.jp/~spaelti/Klamath/files_finished/KD_Introduction.pdf)

Note that the original text, taken from M.A.R. Barker's *Klamath Dictionary* (Berkeley and Los Angeles, 1963), had the raised circular dot:

(16) /iˑ/ is approximately the "i" of "machine."

(Source: http://ksw.shoin.ac.jp/~spaelti/Klamath/files_img/KD_Introduction.pdf)

The difference between the triangular and the circular shape was already blurred in Americanist usage, cf. the following summary drawn from Pullum and Ladusaw's *Phonetic Symbol Guide* (2nd ed., Chicago, 1986, pp. 244-5, 248-9), in which the Americanists tend to use a raised period or the colon rather than the triangular "half-length mark" and "length mark".

CHARACTER	***IPA USE***	***AMERICANIST USE***
Raised Period ˙ U+ ?	"not used", but often used as typogr. subst. for half-length mk.	preceding segment is long *Boas: preceding sound is long; use colon <:> to show extra long*
Half Length Mk ˑ U+02D1	preceding syll. is half-long	use=IPA, but employ <ˑ>
Colon : U+003A	"not used", but often used as typogr. subst. for length mark	marks length of segment *Boas: marks excessive length, or in a 2-way contrast, same as <ˑ>*
Length Mark ː U+02D0	preceding letter is long	use=IPA, colon generally substituted

5. Summary

The encoding of another middle dot for Phags-Pa is unnecessary, particularly as the middle dot is already use widely in linguistic transcription/transliteration and Americanist orthographies, and seems to be encoded on modern webpages by U+00B7 or U+02D1. The result of encoding another middle dot will be to create yet another look-alike character.

In my view, the best option for users is to use U+02D1 with a rounded glyph. This character is being used by linguists and others currently, is able to be found via search engines, and is found in both circular and triangular shapes.

6. Figures

Figure 1: Munsee

Linguistic	Practical	English
ampi·lamé·kwa·n	ambiilaméekwaan	*needle*

This shows the linguistic and the practical orthography used for Munsee, an endangered language that is a member of the Algic language family. The linguistic form uses a raised dot for the long vowel, but the practical orthography doubles the vowel to show length. This page uses U+00B7 for the raised dot.

Source: http://en.wikipedia.org/wiki/Munsee_language

Figure 2: Central Sierra Miwok

Bear Shaman[1]

1. wý?·anyk hoj?epaj šók·et·ikoŋ míw·y·ko·ŋ, wý?anyk lemè·j. lolúk·uš nét·o? šỳle·j, má?tana· palát·at·aj lé·ka·t.

1. They went out early in the morning, all the people, they went out into the hills. When he had collected a lot of fledgelings in one place, he shot a woodpecker on a white oak.

Note the use of the raised dot used in this Central Sierra Miwok text.

Source: Freeland, L.S., and Sylvia M. Broadbent, *Central Sierra Miwok Dictionary with Texts* (1960), p. 59 Accessed from: http://www.yosemite.ca.us/library/central_sierra_miwok_dictionary/page_59.html

Figure 3 Unami

Linguistic	Practical	English
kw?ˊtˑi	kwëti	*one*
níˑšˑa	nìshi	*two*

Unami is a member of the Algic language family, related to Munsee. According to the *Ethnologue*, it is extinct. This figure shows the use of the raised dot to denote consonant and vowel length in the linguistic transcription. In this document, U+00B7 is used.

Source: http://en.wikipedia.org/wiki/Delaware_languages

Figure 4 Proto-Takelman

NO. 3 PROTO-TAKELMAN 22

Sapir's Phonetics	Value	Sapir's Phonemics	Research Orthography
a	[ɑ], [ʌ]m	a	a
e	[ɛ..æ]	e	e
i	[I], [i]	i	i
o, u	[o]	o	o
û, ü	[u], [U..Ù]	u (ü)	u
û	[ʌ]	a	a
E	m[ʌ]s	—	—
w, V]u	[u]	w	w
y, V]i	[i]	y	y
a^a	[ɑ:]	aˑ	aˑ
e^e	[ɛ:]	eˑ	eˑ
i^i	[i:]	iˑ	iˑ
o^u	[ow]	oˑ	oˑ
u^u	[u:]	uˑ	uˑ
p', t', k', k'ʷ	[pʰ, tʰ, kʰ, kʷʰ]	p, t, k, kʷ	ph, th, kh, kʷh
b, d, g, gʷ	[p, t, k, kʷ]	b, d, g, gʷ	p, t, k, kʷ
p!, t!, k!, kʷ!	[ṗ, ṫ, k̇, k̇ʷ]	ṗ, ṫ, k̇, k̇ʷ	ṗ, ṫ, k̇, k̇ʷ
ɛ	[ʔ]	?	?
ts!, tsˑ!	[ċ], [ȩ]	ċ	ċ

In this figure for Proto-Takelman, the ancestor of Takelma (and member of Penutian language family), note the middle dot for vowel length in the column for "Sapir's Phonemics" and "Research Orthography". The middle dot appears also in the first column, "Sapir's Phonetics", under the last entry for a fortis sibilant [tsˑ!].

Source: Shipley, William. "Proto-Takelman." *International Journal of American Linguistics*, Vol. 35, No. 3 (Jul., 1969), pp. 226-230. Accessed from: http://www.jstor.org/stable/1264690

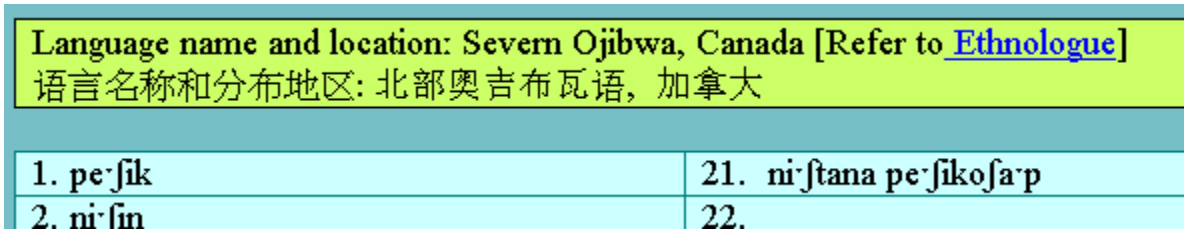
Figure 5 Tonkawa and Algonkian

HIDE (verb). PCA *kya·- (H 124). Tnk c?a·pe- (890).
HOT. PCA *kesy- (H 58). Tnk xal(al) warm, hot (762) and/or ka·le- to become hot (374).
JEALOUS. PCA *kya·- (H 124). Tnk c?ey?e- (894).

This snippet shows a comparison of "Proto-Central Algonkian" (PCA) forms and words from Tonkawa, an extinct language from the Coahuiltecan family. The middle dot is used for the long vowels in both sets of forms. (The article is by Mary Haas, a student of Edwin Sapir, who was himself mentored by Franz Boas; the use here and in figure 4 demonstrate the continuous use of the middle dot by generations of linguists.)

Source: Haas, Mary R. "Tonkawa and Algonkian." *Anthropological Linguistics*, Vol. 1, No. 2, Genetic Relationship among Languages: A Symposium Presented at the 1958 Meetings of the American Anthropological Association (Feb., 1959), pp. 1-6. Accessed from: http://www.jstor.org/stable/30022178.

Figure 6 Severn Ojibwa



This example for the Severn Ojibwa language uses U+02D1with the triangular glyph. The standard orthography renders /aa/ for /aː/, /oo/ for /oː/.

Source: http://lingweb.eva.mpg.de/numeral/Ojibwa-Severn.htm