ISO/IEC JTC1/SC2/WG2 N4339

September 21, 2012

Title: **Examples of Collation Tailoring for U+00B7 MIDDLE DOT**

Source: Ken Whistler

Status: Expert Contribution

Action: For consideration by WG2

One of the arguments proffered for the need to encode a new character for **LATIN LETTER MIDDLE DOT** is that the character properties for U+00B7 MIDDLE DOT are inappropriate for that use, even though the glyph for U+00B7 is indistinguishable from the proposed middle dot functioning as a letter.

Although properties for characters interact in many ways with various algorithms, the main problem being referred to here is that because U+00B7 MIDDLE DOT is a punctuation character (i.e., has a character property assignment of General Category=Po), it cannot be found in searches. This is a result of the fact that most searches, unless they are tailored in some way, tend to distinguish "content" characters (roughly, letters and numbers) but to ignore "non-content" characters (roughly, punctuation, spaces and format characters). [Note that symbols fall in between and are sometimes treated as content and sometimes ignored as non-content, depending on search criteria and implementations.]

Related to the problem of failing searches is the allegation that U+00B7 will not sort correctly in lists which treat it as a letter. This follows from the fact that for other than binary order collations, punctuation (and format) characters also tend to be ignored in sorting. For multi-level collation algorithms, punctuation characters are often either ignored completely, or they get weights at a quaternary level which makes them far less significant for ordering than the primary collation weights typically given to letters and numbers. For an orthography using a middle dot as a letter, then, the sorting results will be odd and "wrong", because the middle dot isn't getting handling for the ordering as if it were another letter of the orthography.

Although this kind of default behavior is annoying for end users who expect their characters to "just work", the issue for U+00B7 MIDDLE DOT is just one of the spectrum of behavior contrary to expectations for searching and sorting which can result from arbitrary choices of particular characters to be used as part of orthographies. It is essentially the same problem which results from choices of other punctuation or symbol characters (e.g. "@", "&", ":", ".", "º" and so forth) to serve as significant elements of an orthography, rather than as punctuation or symbols.

Furthermore, as regards searching and sorting, in particular, there are mechanisms available for getting the expected behavior for particular corpuses and orthographies. This is a result of the fact that searching and sorting is typically based on mechanisms which allow tailoring of one sort or another. And there are ways to define tailored collators which, with a little effort, can produce the expected results.

This document gives an example of how such tailoring works, using a short sample of nonsense disyllabic “words” constructed to illustrate both the potential use of U+00B7 MIDDLE DOT to represent a consonant letter (i.e. a glottal stop), as seen in some East Asian transliteration schemes, and the use of U+00B7 to represent a length mark, as seen in many orthographies based on Americanist phonetic practice.

Table 1 shows the input data, in no particular order, as it was created.

Table 1: Input Data

baba	bebe	bibi	·aba	·ebe
·ibi	ba·a	be·e	bi·i	ba·ba
be·be	bi·bi	baba·	bebe·	bibi·
ba·ba·	be·be·	bi·bi·	baʔa	beʔe
biʔi	ba·ʔa	be·ʔe	bi·ʔi	baʔa·
beʔe·	biʔi·	ba·ʔa·	be·ʔe·	bi·ʔi·
ʔaba	ʔebe	ʔibi	ʔa·ba	ʔe·be
ʔi·bi	ʔaba·	ʔebe·	ʔibi·	ʔa·ba·
ʔe·be·	ʔi·bi·	ʔaʔa	ʔeʔe	ʔiʔi
ʔa·ʔa	ʔe·ʔe	ʔi·ʔi	ʔaʔa·	ʔeʔe·
ʔiʔi·	ʔa·ʔa·	ʔe·ʔe·	ʔi·ʔi·	ʔaʔa
ʔeʔe	ʔiʔi	ʔa·ʔa	ʔe·ʔe	ʔi·ʔi
ʔaʔa·	ʔeʔe·	ʔiʔi·	ʔa·ʔa·	ʔe·ʔe·
ʔi·ʔi·	ʔaca	ʔece	ʔici	ʔa·ca
ʔe·ce	ʔi·ci	ʔaca·	ʔece·	ʔici·
ʔa·ca·	ʔe·ce·	ʔi·ci·	ʔacʰa	ʔecʰe
ʔicʰi	ʔa·cʰa	ʔe·cʰe	ʔi·cʰi	ʔacʰa·
ʔecʰe·	ʔicʰi·	ʔa·cʰa·	ʔe·cʰe·	ʔi·cʰi·
ʔac’a	ʔec’e	ʔic’i	ʔa·c’a	ʔe·c’e
ʔi·c’i	ʔac’a·	ʔec’e·	ʔic’i·	ʔa·c’a·
ʔe·c’e·	ʔi·c’i·	ʔača	ʔeče	ʔiči
ʔa·ča	ʔe·če	ʔi·či	ʔača·	ʔeče·
ʔiči·	ʔa·ča·	ʔe·če·	ʔi·či·	ʔačʰa
ʔečʰe	ʔičʰi	ʔa·čʰa	ʔe·čʰe	ʔi·čʰi
ʔačʰa·	ʔečʰe·	ʔičʰi·	ʔa·čʰa·	ʔe·čʰe·
ʔi·čʰi·	ʔač’a	ʔeč’e	ʔič’i	ʔa·č’a
ʔe·č’e	ʔi·č’i	ʔač’a·	ʔeč’e·	ʔič’i·
ʔa·č’a·	ʔe·č’e·	ʔi·č’i·		

In addition to many middle dots, this data also makes use of U+0294 LATIN LETTER GLOTTAL STOP (ʔ), to illustrate how various tailoring options might treat it in comparison to the middle dot. It also contains multiple instances of U+02B0 MODIFIER LETTER SMALL H and U+02BC MODIFIER LETTER APOSTROPHE, which are commonly used in Americanist orthographies to represent aspiration and glottalization, respectively.

Table 2 shows the results of processing the data in Table 1 through an indexing utility program, using the default settings for that program. An indexing utility combines both a lexing operation, which has to search through text, segmenting units according to some criteria, and then a sorting operation, which puts all of the segmented units into some defined order. In this case, the default settings for the program make use of some generic criteria for lexing (including breaking elements at punctuation characters), and the DUCET default table values for the Unicode Collation Algorithm for its ordering. In Table 2, the first column represents the number of times a particular unit was found in the input data, and the second column shows the text of the actual unit which was parsed out.

Table 2: Vanilla Lexing and UCA Default Ordering

1	aba
9	ba
2	baba
2	baʔa
9	be
2	bebe
2	beʔe
9	bi
2	bibi
2	biʔi
2	ca
2	ča
2	ce
2	če
2	cʰa
2	čʰa
2	cʰe
2	čʰe
2	cʰi
2	čʰi
2	ci
2	či
2	cʼa
2	čʼa
2	cʼe
2	čʼe
2	cʼi
2	čʼi
1	e
1	ebe
1	i
1	ibi
24	ʔa
2	ʔaba
2	ʔaca
2	ʔača
2	ʔacʰa
2	ʔačʰa
2	ʔacʼa
2	ʔačʼa
4	ʔaʔa
24	ʔe
2	ʔebe
2	ʔece
2	ʔeče
2	ʔecʰe
2	ʔečʰe
2	ʔecʼe
2	ʔečʼe
4	ʔeʔe
24	ʔi
2	ʔibi
2	ʔicʰi
2	ʔičʰi
2	ʔici
2	ʔiči
2	ʔicʼi
2	ʔičʼi
4	ʔiʔi

The first obvious thing to note about Table 2 is that all of the middle dots in the original data have disappeared! This is a result of the generic criteria for lexing, which treats the middle dot as punctuation, and as thus constituting the edge of words to be indexed, rather than being included in them. It is this kind of behavior which has led to the requirement to encode a separate Latin **letter** middle dot, which would have different properties and not “disappear” like this.

Now, however, consider a somewhat modified processing of the same input data with the same indexing utility program. In this case, using parameters available to the program, the lexing module is instructed not to treat middle dot as a segmentation boundary, but rather to include it in words. Separately, the sorting module is instructed to use a tailored version of DUCET which treats middle dot as having a primary weight (like a letter), rather than being ignored for collation, and which in fact gives the middle dot the exact same primary weight as the letter glottal stop. The results of this processing can be seen in Table 3.

Table 3: Middle Dot as Letter Weighted Equivalent to Glottal Stop

1	baba
1	baba·
2	ba·a
1	baʔa·
1	ba·ba
1	ba·ba·
1	ba·ʔa
1	ba·ʔa·
1	bebe
1	bebe·
1	be·be
1	be·be·
2	be·e
1	beʔe·
1	be·ʔe
1	be·ʔe·
1	bibi
1	bibi·
1	bi·bi
1	bi·bi·
2	bi·i
1	biʔi·
1	bi·ʔi
1	bi·ʔi·
2	·aba
1	ʔaba·
1	ʔaca
1	ʔača
1	ʔaca·
1	ʔača·
1	ʔacʰa
1	ʔačʰa
1	ʔacʰa·
1	ʔačʰa·
1	ʔac’a
1	ʔač’a
1	ʔac’a·
1	ʔač’a·
2	ʔaʔa
2	ʔaʔa·
1	ʔa·ba
1	ʔa·ba·
1	ʔa·ca
1	ʔa·ča
1	ʔa·ca·
1	ʔa·ča·
1	ʔa·cʰa
1	ʔa·čʰa
1	ʔa·cʰa·
1	ʔa·čʰa·
1	ʔa·c’a
1	ʔa·č’a
1	ʔa·c’a·
1	ʔa·č’a·
2	ʔa·ʔa
2	ʔa·ʔa·
2	·ebe
1	ʔebe·
1	ʔece
1	ʔeče
1	ʔece·
1	ʔeče·
1	ʔecʰe
1	ʔečʰe
1	ʔecʰe·
1	ʔečʰe·
1	ʔec’e
1	ʔeč’e
1	ʔec’e·

1	ʔeč’e·
1	ʔe·be
1	ʔe·be·
1	ʔe·ce
1	ʔe·če
1	ʔe·ce·
1	ʔe·če·
1	ʔe·cʰe
1	ʔe·čʰe
1	ʔe·cʰe·
1	ʔe·čʰe·
1	ʔe·c’e
1	ʔe·č’e
1	ʔe·c’e·
1	ʔe·č’e·
2	ʔeʔe
2	ʔeʔe·
2	ʔe·ʔe
2	ʔe·ʔe·
2	·ibi
1	ʔibi·
1	ʔicʰi
1	ʔičʰi
1	ʔicʰi·
1	ʔičʰi·
1	ʔici
1	ʔiči
1	ʔici·
1	ʔiči·
1	ʔic’i
1	ʔič’i
1	ʔic’i·
1	ʔič’i·
1	ʔi·bi
1	ʔi·bi·
1	ʔi·cʰi
1	ʔi·čʰi
1	ʔi·cʰi·
1	ʔi·čʰi·
1	ʔi·ci
1	ʔi·či
1	ʔi·ci·
1	ʔi·či·
1	ʔi·c’i
1	ʔi·č’i
1	ʔi·c’i·
1	ʔi·č’i·
2	ʔiʔi
2	ʔiʔi·
2	ʔi·ʔi
2	ʔi·ʔi·

With those two small changes to the processing, the middle dot has now “magically” reappeared in the output, incorporated in the words identified as units, and is being treated as a full letter. Furthermore, it is fully interfiled with the other explicit glottal stops, which demonstrates that the ordering here has correctly followed the tailoring rules giving it the same weight as a glottal stop. (Of course, a particular transliteration system using middle dot for glottal stop wouldn’t use both characters at the same time – this data is just being used as proof of concept to show that the U+00B7 here has effectively become a letter for this data, and is in all ways being treated exactly as if it were a glottal stop letter.)

Finally, to demonstrate the flexibility of this approach, the same indexing utility program, using the same input data, was given a different set of tailoring instructions. This time, instead of weighting the middle dot as equivalent to a glottal stop, the tailoring treats it as a modifier letter representing length, which is the usual convention for Americanist orthographies. The tailoring also modified the weight for the modifier letters for aspiration and glottalization. In particular, U+00B7 MIDDLE DOT was tailored to weight exactly equivalent to a combining macron, giving it a secondary weight equivalent to that other common accentual representation of vowel length. U+02BC MODIFIER LETTER APOSTROPHE was tailored to weight exactly equivalent to a combining comma above. And U+02B0 MODIFER LETTER SMALL H was tailored to weight as a generic accent, so that it no longer is treated as a variant of the letter h. (There is no specific combining mark in Americanist orthography for aspiration – the point here is to treat it, for ordering purposes, in parallel with glottalization.) Additionally, the glottal stop was tailored to come at the start of the order of Latin letters, rather than the near the end.

Those tailoring options, taken together, closely approximate the conventions typically seen in dictionary ordering of linguistic data using the Americanist orthographies. The result of indexing the input data using these options can be seen in Table 4.

Table 4: Middle Dot as Modifier Letter Weighted Equivalent to Macron

2	ʔaʔa
2	ʔaʔa·
2	ʔa·ʔa
2	ʔa·ʔa·
1	ʔaba
1	ʔaba·
1	ʔa·ba
1	ʔa·ba·
1	ʔaca
1	ʔaca·
1	ʔacʰa
1	ʔacʰa·
1	ʔac’a
1	ʔac’a·
1	ʔača
1	ʔača·
1	ʔačʰa
1	ʔačʰa·
1	ʔač’a
1	ʔač’a·
1	ʔa·ca
1	ʔa·ca·
1	ʔa·cʰa
1	ʔa·cʰa·
1	ʔa·c’a
1	ʔa·c’a·
1	ʔa·ča
1	ʔa·ča·
1	ʔa·čʰa
1	ʔa·čʰa·
1	ʔa·č’a
1	ʔa·č’a·
2	ʔeʔe
2	ʔeʔe·
2	ʔe·ʔe
2	ʔe·ʔe·
1	ʔebe
1	ʔebe·
1	ʔe·be
1	ʔe·be·
1	ʔece
1	ʔece·
1	ʔecʰe
1	ʔecʰe·
1	ʔec’e
1	ʔec’e·
1	ʔeče
1	ʔeče·
1	ʔečʰe
1	ʔečʰe·
1	ʔeč’e
1	ʔeč’e·
1	ʔe·ce
1	ʔe·ce·
1	ʔe·cʰe
1	ʔe·cʰe·
1	ʔe·c’e
1	ʔe·c’e·
1	ʔe·če
1	ʔe·če·
1	ʔe·čʰe
1	ʔe·čʰe·
1	ʔe·č’e
1	ʔe·č’e·
2	ʔiʔi
2	ʔiʔi·
2	ʔi·ʔi
2	ʔi·ʔi·
1	ʔibi
1	ʔibi·
1	ʔi·bi
1	ʔi·bi·
1	ʔici
1	ʔici·
1	ʔicʰi
1	ʔicʰi·
1	ʔic’i
1	ʔic’i·
1	ʔiči
1	ʔiči·
1	ʔičʰi
1	ʔičʰi·
1	ʔič’i
1	ʔič’i·
1	ʔi·ci
1	ʔi·ci·
1	ʔi·cʰi
1	ʔi·cʰi·
1	ʔi·c’i
1	ʔi·c’i·
1	ʔi·či
1	ʔi·či·
1	ʔi·čʰi

1	ʔi·čʰi·
1	ʔi·č’i
1	ʔi·č’i·
1	·aba
1	baʔa
1	baʔa·
1	ba·ʔa
1	ba·ʔa·
1	ba·a
1	baba
1	baba·

1	ba·ba
1	ba·ba·
1	beʔe
1	beʔe·
1	be·ʔe
1	be·ʔe·
1	bebe
1	bebe·
1	be·be
1	be·be·
1	be·e

1	biʔi
1	biʔi·
1	bi·ʔi
1	bi·ʔi·
1	bibi
1	bibi·
1	bi·bi
1	bi·bi·
1	bi·i
1	·ebe
1	·ibi

In this ordering, it is easy to see that the middle dot is treated as a part of the words and as significant for ordering, but is no longer equivalent to the glottal stop in any way. Instead it makes a secondary difference for vowel length in this orthographic convention.

Now it is true that the particular indexing utility I have used here was custom software that I built myself, using an implementation of the Unicode Collation Algorithm that I also built myself. I am not expecting that anyone wanting to use characters in data like this would have to go off and write full implementations from scratch. But this is not exactly “rocket science”, when it comes to currently available software. Anybody with a modest command of Perl, for example, could duplicate the lexing aspects of this behavior, treating U+00B7 as a special case for the purposes of searching, matching, and text segmentation for any particular set of data. And a full, open source implementation of the Unicode Collation Algorithm is available in ICU, with a large number of preconfigured standard tailorings and a full syntax for defining any number of additional arbitrary tailorings for collation. It is a relatively straightforward task for someone using ICU to define a custom collator which could replicate the kinds of custom ordering I have demonstrated above.