L2/12-361

ISO/IEC JTC1/SC2/WG2 N4340

September 28, 2012

Title: **Comments in Response to Irish Comments on Middle Dot**

Source: Ken Whistler

Status: Expert Contribution

Action: For consideration by WG2

In the Irish comments on ISO/IEC PDAM 2 10646:2012, included in SC2 N4233, in comment T5, there is a statement reiterating Ireland's support for encoding A78F LATIN LETTER MIDDLE DOT. That statement is followed by an extensive number of comments referring to the US National Body comments on this and prior amendment ballots objecting to the encoding of that character.

I would like to respond to the Irish comments in several particulars, to clarify the objections that people have been raising with regard to this proposed encoding.

Transliteration is Not the Issue

The argument made in the Irish comments cites Andrew West's stated requirement, "... to have a character with the properties of a *letter* to represent in transliteration the *letter* of another script." But there are a couple of implied premises here which are not correct, in my opinion.

First, the middle dot in question is used in a 'Phags-Pa transliteration system, as noted, but that is simply a use made of a convention in sinology of the middle dot for transcription of a glottal stop (or somewhat indifferently for phonological representation of a glottal stop). The point here is that there is no obvious requirement that all transcriptional conventions end up represented by **letters** (gc=L) in Unicode. And when such a convention also gets applied to a particular trans**literation** system, that use doesn't automatically require matching Unicode properties, even when the target of the transliteration is itself a letter. The Irish comments recognize that the middle dot is also used in Tangut transcription, and indeed the original usage by Karlgren was also for ancient Chinese transcription, not transliteration. But it keeps coming back to focus on the use for 'Phags-Pa transliteration, as if that fact automatically made the case stronger for requiring letter properties for the character. It doesn't.

Second, even when used in a transliteration system, there is no self-evident requirement that a character exactly match the properties of the target it transliterates. There are a number of reasons why this might be the case, but the clearest and easiest to describe would be instances where transliterations make use of arbitrary symbols, such as numbers or string labels, to represent elements which may yet be undeciphered in some script.

Furthermore, for example, if Karlgren had chosen to use a "2" or a "*" or a ")" or a "-" to represent the "laryngeal explosive" in his ancient Chinese reconstruction instead of a dot and that convention had

caught on, would we be having this same argument about how a LATIN LETTER RIGHT PARENTHESIS or a LATIN LETTER HYPHEN needed to be encoded because it was being used in sinology to represent a glottal stop "letter"?

Examples of non-letters used in transliteration

Romanization Systems and Roman-Script Spelling Conventions (1994), by the U.S. Board of Geographic Names, recommends the use of middle dot in transliteration to differentiate consonant sequences from digraphic transliterations. E.g. "k ·h" versus "kh", "z ·h" versus "zh" and so forth for Arabic, and "y·", "·y", "·e", "t·s" and so forth for Russian in certain cases. True, this middle dot is not being used to transliterate a letter per se, but it *is* being used to make distinctions in these transliterations internal to words. So this convention has the same issues for searchability and collation as the use of a middle dot for a letter (or phone) in Sinological conventions.

There are other instances of the use of odd, arbitrary conventions for letters or punctuation to represent word elements in transliteration. Some of these even make it into ISO transliteration standards. For example, the ISO 11940 standard for transliteration of Thai recommends U+01C2 LATIN LETTER ALVEOLAR CLICK to transliterate U+0E2F THAI CHARACTER PAIYANNOI (gc=Lo). In that case, the phonetic semantics of the letter U+01C2 has nothing whatsoever to do with the function of paiyannoi in Thai. The same standard also recommends the punctuation character U+00AB LEFT-POINTING DOUBLE ANGLE QUOTATION MARK to transliterate U+0E46 THAI CHARACTER MAIYAMOK (gc=Lm). Once again, there is no connection whatsoever between that double guillemet punctuation character and the semantics of maiyamok, which indicates repetition. Furthermore, this is a clear case of a punctuation mark being used to transliterate a modifier letter in another script.

As regards glottal stop, in particular, there are many examples of representation of a glottal stop by various characters, both in transliteration and in more-or-less standard orthographies. Many of these simply borrowed de novo from whatever was available to make a distinction, given the printing technology at the time. In addition to the middle dot for the "laryngeal explosive" which Karlgren popularized among sinologists[1], there are all the hamza and alef forms appearing in the Semitic traditions. The latter have mostly made it into Unicode with separate characters, for a variety of reasons, but that precedent doesn't necessarily mean that the conventional use of some shape to represent a glottal stop is a sufficient reason to justify separate encoding. For example, there are the classic typewriter substitutions of "?" or "7" for a glottal stop. These are widespread in language materials in the mid-20th century, particularly in the Americas. But the use of such punctuation or number substitutions for a glottal stop, no matter how widespread, has not been taken as requiring separate encoding of a LATIN LETTER QUESTION MARK or a LATIN LETTER SEVEN to represent them in Unicode.

[1] The Irish ballot comments cite the use of the dot for glottal stop as "going back to Sinologists Dragonov in the 1930s and Karlgren in the 1940s." Actually, this use by Bernhard Karlgren predates his publication of *Grammata Serica* in 1940, and can be traced back even further to his seminal article, "The Reconstruction of Ancient Chinese," published in 1922 in *T`oung Pao*, Vol. XXI. See p. 12 and p. 24 for examples of Karlgren's use of a middle dot to represent what he called a "laryngeal explosive" in that article.

One would probably not have to look too far to find other instances of linguists from the mid-20th century making other typewriter key substitutions, using available symbols such as “@”, “*”, or “+”, for example, to represent sounds such as schwa, which were also not easily available for representation in typescript material.

The point here is that simply pointing to a transliteration convention (or orthographic convention in a non-transcriptional context) which makes use of a punctuation mark, symbol, or number to represent a letter (or to represent some other non-ignorable phonological or semantic distinction in words) is not sufficient, by itself, to justify separate encoding of a Latin letter character version of that punctuation mark, symbol, or number.

Middle Dot Versus IPA Length Mark

Much of the argumentation in the Irish ballot comments is making the case that U+02D1 MODIFIER LETTER HALF TRIANGULAR COLON would be an inappropriate character to represent the middle dot used in the ‘Phags-Pa transliteration convention for glottal stop. This argumentation is addressed at earlier U.S. NB ballot comments and associated documentation (N3678), which suggested use of either U+00B7 MIDDLE DOT or U+02D1 as an alternative to encoding a new character for this purpose.

I actually agree with the Irish ballot comments that U+02D1 is not particularly appropriate for representation of the ‘Phags-Pa dot for glottal stop. On the other hand, U+02D1 is a perfectly reasonable choice for the representation of length by means of a midline single dot in widespread Americanist practice. In cases where the users of such an orthography care less about the exact visual presentation of the data, but expect the length mark to behave by default as a letter-extending modifier mark signifying length, then making use of the IPA modifier letter with exactly that semantic is an obvious choice.

On the other hand, if users of such an Americanist orthography expect the length mark dot to be presented reliably as a rounded dot in all contexts, instead of a triangular dot shape, then U+00B7 is a reasonable choice to make. If that choice is made, then of course there are consequences for processing of data. Because of the ambiguous usage of U+00B7 and its default character properties, a user of data including U+00B7 as a length mark has to put up with the requirement for tailoring of certain processes, if they expect searching and sorting and segmentation to work as expected. Fortunately, there are ways to do precisely the kind of tailoring needed for specialized processing of corpus data. In my separate contribution I have provided examples which show the result of such tailoring of software processes, and how it can be used to adjust searching, sorting, and segmentation to produce the desired effects. On the other hand, for many applications using U+00B7 as a length mark, the main issue is simply correct appearance in presentation, rather than systematic corpus processing or searching and sorting. In such contexts, as for example, the preparation of a text for publication, use of U+00B7 requires no tailoring or adjustments at all.

Shape of U+02D1

As regards the presentation of U+02D1, the Irish ballot comments state that "It is not in the purview of the author of N3678, or of the US National Body, to alter by fiat the shape of the character 02D0 or 02D1 which exist distinct from MIDDLE DOT and from COLON to support the *explicitly-triangular* character used by the International Phonetic Association." And they cite John C. Wells as "mak[ing] it clear that the idea that the 02D1 MODIFIER LETTER HALF TRIANGULAR COLON could have any other shape than triangular was quite out of the question."

First of all, it is simply wrong to imply that the author[s] of N3678 or the U.S. National Body is attempting to "alter by fiat" the shape of U+02D1 (or U+02D0). Nobody is suggesting any change to the glyph in the code charts of ISO/IEC 10646 or the Unicode Standard, nor any change to accepted IPA use and practice.

On the other hand, U+02D1 is not owned by the IPA, any more than U+0074 LATIN SMALL LETTER T, which is also used in IPA transcription, is owned by the IPA. It is certainly the case that the original purpose of encoding U+02D1 was to represent the triangular-shaped half-colon length mark in IPA, so that there would be a character usable for that in IPA which was not afflicted with the multiple functions ascribed to U+00B7, and so that IPA fonts could be constructed which would reliably display the IPA length mark which the shape expected in IPA transcription. However, as the Irish national body is also fond of reminding us, once a character is encoded in the UCS, it is available to all users, and may be put to use in contexts and with meanings not originally envisioned and outside of the ability of SC2 (or the Unicode Consortium) to constrain or control. This is one of the reasons, for example, that the Irish national body spends so much effort in ensuring that pictographic symbols added to the standard are not encumbered with overly precise glyphs and semantic definitions on the assumption that they could somehow forever be constrained to use only in some narrowly conceived original mapping context.

In the case of U+02D1, and indeed for any character used in IPA transcription, if such a character is used outside of a context narrowly defined as formal IPA transcription, it is not self-evident that a user is then constrained to display it precisely as shown in the Handbook of the International Phonetic Association or else not use it at all. As for most characters in the standard, there is ample latitude for font design and glyph variation for presentation – such issues are simply outside the scope of 10646. The text of ISO/IEC 10646, for good reasons, is silent about what constraints there are how much glyph variation is allowable for any particular character. In fact, what it *does* say in Clause 13 is:

> Graphic symbols[2] are to be regarded as typical visual representations of the corresponding graphic characters. This International Standard does not attempt to prescribe the exact shape or colour of each character. The shape is affected by the design of the font or other representation method employed, which is out of scope.

In my opinion, the appropriateness of particular glyph design choices for U+02D1 (or any other character) depends both on the aesthetics of font design and the intent of use for the application. The

[2] "Graphic symbol" is synonymous with the term "glyph" as used in the Unicode Standard.

choice of a rounded dot glyph for U+02D1 in a font would be completely inappropriate for a chart font, for a general-purpose IPA font, or for an application doing OCR of IPA transcriptional data. To do so would only be confusing and error-prone. On the other hand, a linguist doing large corpus work on Miwokan or Yokutsan data – language families which typically have very common use of a middle dot as a length mark (both for vowels and for consonants) – may be fully justified in representing that length mark with U+02D1 and then designing a special-purpose font for publication of his or her data and analysis using a rounded dot display glyph for U+02D1, so that the data appears as traditionally shown for those languages.

These kinds of concerns for glyph design are a matter of judgment. As I see it, what one would want to avoid would be deliberate misrepresentation or confusion about the data. The extreme case sometimes cited would be displaying a “b” with the glyph for an “a” and vice versa, which could only serve to misrepresent the text content. But using an adjusted glyph for a length mark as in the example I just cited for Miwokan or Yokutsan data, is *not* a misrepresentation of the data – it is, rather, an appropriate use of the standard, in my opinion.

I fully expect the Irish national body to continue to disagree with this take on the issue for use and presentation of characters such as U+02D1 – or indeed, for punctuation marks in general. But I consider the ongoing efforts to continue slicing and dicing characters such as the middle dot in 10646, cloning more and more versions based on functional, property, or small glyph differences, to be rather damaging to the standard in the long run. This isn’t anything new, of course – there are many such instances of characters of identical appearance cloned in the standard on functional or property grounds, dating all the way back to the original repertoire of Unicode 1.0. The problem I see is that the continued accumulation of further such distinctions in encoding, often on dubious grounds, incrementally damages the standard further by sowing confusion about the use of common characters, particularly when proposed disunifications and new encodings happen long after data has been generated using already existing but potentially ambiguous characters.

Middle Dot and the Sinological Glottal Stop Dot

I mentioned in the above discussion that I agree with the Irish ballot comments contention that U+02D1 is not appropriate for representation of the glottal stop dot seen in Karlgren’s reconstruction of ancient Chinese and also in the ‘Phags-Pa transliteration convention cited by Andrew West. However, I do not agree with the further claim that U+00B7 is not appropriate. In fact, U+00B7 MIDDLE DOT is precisely what I would recommend for this purpose.

Furthermore, I don’t think that either N3678 or the various U.S. national body comments should be interpreted as having “accepted West’s requirement for a letter (a character with a letter property) rather than a punctuation character for the purpose of transliterating ‘Phags-pa.” As a consequence, a lot of the criticism then addressed at the use of U+02D1 is rather beside the point. U+02D1 is an acceptable alternative for the representation of a middle dot length mark, under certain circumstances, as just discussed.

However, the argument for the need for a newly encoded LATIN LETTER MIDDLE DOT for ‘Phags-pa transliteration hinges, rather, almost entirely on the claim that the character properties for the existing U+00B7 won’t work for this function – a claim that I have addressed in the discussion of transliteration above.

If the character property argument for separate encoding of a LATIN LETTER MIDDLE DOT does not convince, then the case for encoding it effectively boils down to a claim that *this* middle dot doesn’t look like *that* middle dot – ours is bigger and fatter, in fact “50% larger”. This argument might convince if there was systematic evidence provided of regular contrastive use between a larger middle dot for transliterating glottal stop and a smaller middle dot for segmentation punctuation, but to date I haven’t seen such evidence. I have yet to find such evidence in Karlgren’s publications, for example. And in its absence, this claim, based mostly on the interpretation of non-uniform typography from the early 20th century, strikes me as an ex post facto attempt to make the encoding of a LATIN LETTER MIDDLE DOT palatable by saying, see, we’ll make it less confusable with the existing U+00B7 MIDDLE DOT by making the glyph fatter and putting in an annotation saying it should be larger. In other words, this is more of an argument to defuse potential national body objections to the encoding, rather than an attempt to demonstrate an actual contrastive usage in plain text requiring separate encoding in the first place.

Indeed, if size of the dot is the distinguishing criterion, then I see no really good reason not to just proceed to the suggestion to use U+2022 BULLET instead for this larger middle dot, because the proposed larger glyph for LATIN LETTER MIDDLE DOT would not be systematically distinguishable from the glyph for *that* character.

But What’s the Harm, Anyway?

The argument of last resort for LATIN LETTER MIDDLE DOT then ends up being stated more or less along these lines: “What’s the harm?? We require this character. You don’t have to use it if you don’t like it. So because we require it, it must be encoded.”

I see this argument – often heard in WG2 in one form or another – as a reasonable one in many cases. It is appropriate to make such an argument, for example, when some particular expert, government, or national body objects to the whole idea of adding characters for an entire script, or subsets of characters for historical use or minority language use, or the encoding of a limited use script or a constructed script, or sets of pictographic symbols used on cell phones, etc., etc. In such cases there really is a clear requirement and no alternative, and in most cases the person or entity objecting can simply ignore the additions and not be harmed.

But the middle dot argument is of another type. In this case, the objection to LATIN LETTER MIDDLE DOT is couched not as a matter of “We don’t like that thing you’re trying to represent, and don’t think characters should be encoded for it,” but rather as a claim that what is to be represented is already encoded, and the requirement is one which amounts to duplicate encoding.

The reason why I, in particular, am passionate about this middle dot case, is that the middle dots in the standard are an *especially* egregious case where the standard is already confusing to use and interpret,

and where addition of more middle dots, no matter how well-intentioned, can really only serve to increase the confusion, and thereby marginally *decrease* the usefulness of the standard for consistent, well-understood representation of text content.

For example, an Americanist linguist attempting to use Unicode/10646 to represent the middle dot length marks in Miwokan data is already faced with a dilemma: should I use U+00B7 MIDDLE DOT, which looks right, but has the "wrong properties" and doesn't say anything about its being a length mark, or should I use U+02D1 MODIFIER LETTER HALF TRIANGULAR COLON, which has the length mark semantics I want, but which "looks wrong"?

Now, let's suppose that SC2 goes ahead and approves and publishes U+A78F LATIN LETTER MIDDLE DOT. How are things any better for the hypothetical linguist with Miwokan data to represent? At this point there is additional choice: U+A78F has the right properties, but still doesn't say anything about being a length mark – in fact, it claims to be a glottal stop used for transliteration for 'Phags-pa and transcription for Tangut, which I don't know anything about and clearly have nothing to do with Miwokan – and the glyph is better than U+02D1, but is unacceptable, because it is too "fat", and doesn't look like U+00B7.

Is the Miwokanist any better off? Nah. The level of confusion has simply been ratcheted up another notch, because now there are three possible choices, none of which seems to fit the Goldilocks bill. Each one has *something* wrong with it.

The inevitable conclusion is that the Miwokanist could come back to WG2 and claim a requirement to encode *another* middle dot character for the Miwokan/Yokutsan length mark character: LATIN MODIFIER LETTER MIDDLE DOT LENGTH MARK. It would need to **look like** U+00B7, have **the same script as** U+A78F, but otherwise **the character properties of** U+02D1. And the justification for that requirement would be exactly the same as that proffered for the LATIN LETTER MIDDLE DOT in the first place: a claim that a special character is needed which works exactly right for *my* middle dot, which is different from *your* middle dot.

This scenario is not just an idle speculation. The proposal for U+A78F LATIN LETTER MIDDLE DOT is not the first time this has happened for middle dot, in particular. The last round was the argumentation which led to the encoding of U+2E31 WORD SEPARATOR MIDDLE DOT "for Avestan, Samaritan, ...". The claim then was that although U+00B7 was indeed a middle dot punctuation mark, it did not function systematically to separate words, which is what the middle dot in Avestan (and similar scripts) does. So another middle dot punctuation was required for encoding which had specific segmentation properties distinct from those for U+00B7. As a result, the standard is now permanently stuck with two middle dot punctuation marks with subtly different segmentation behavior, and with little capability for end users to know how to choose between them.

That, in turn, was a reprise of the earlier separate encoding of an otherwise identical looking U+2027 HYPHENATION POINT, which also looks just like U+00B7 and is also a punctuation mark, but which is used to indicate syllabification, and which does *not* break words. So actually *three* middle dot punctuation marks with subtly different segmentation behavior. Does anybody other than an ICU

implementation tester really claim that they clearly “understand” the implications of the following suite of character property values which in effect were used to justify the separate encoding of U+00B7, U+2027, and U+2E31?

	Line_Break	Word_Break	Extender
U+00B7	AI	MidLetter	Yes
U+2027	BA	MidLetter	No
U+2E31	BA	Other	No

This is another example of why many folks don’t think that separating middle dots by function and semantics is actually helping anyone. It just leads to confusion regarding when to use one and when to use another.

A Bridge Too Far

U+A78F LATIN LETTER MIDDLE DOT is the bridge too far, in my opinion. Continuing to ratchet up the number of middle dots, ostensibly distinguished clearly by their character properties, and basing the argumentation on precedent for other cases, isn’t really going to help anyone. In this case, further muddying the water for all the middle dot characters does represent a harm for everybody using the standard, and a statement that “We require it – you don’t need to use it” simply doesn’t cut any mustard here.