

Proposed Principles for Character Disunifications

Peter Constable, Microsoft
2005-2-2

In the past year, UTC has approved various character disunifications – encoding new characters to create distinctions that were not previously made.¹ For implementers, these have been a mixed bag: some present no significant problems; others, however, were done in ways that leave implementers facing some significant problems that could have been avoided. To avoid such problems in the future, I propose that UTC adopt certain principles that guide how character disunifications should be handled.

Three particular disunifications are considered here: QAMATS, YERAH BEN YOMO and GLOTTAL STOP. I will describe each, explaining why the disunification of YERAH BEN YOMO and GLOTTAL STOP have resulted in problems while the disunification of QAMATS does not. By considering these three cases, some simple principles can be identified that can serve to avoid similar problems in the future.

Disunification of QAMATS

The Hebrew mark *qamats* is one of the vowel points used in pointed Hebrew text. While historically there was only one mark, it can be used to write two different vowel pronunciations. This led in recent times to publishers creating a glyph distinction in order to distinguish the two readings.

Most users do not make this distinction in texts; for them, the existing character U+05B8 HEBREW POINT QAMATS has been adequate. For those that wish to make the distinction, however, two characters are needed: *qamats gadol*, and a separate character *qamats qatan*. The latter typically differs from the former in having a longer stem.



Figure 1. Contrast between *qamats gadol* (short stem) and *qamats qatan* (long stem)

What was proposed and accepted by UTC was to leave the existing character U+05B8 HEBREW POINT QAMATS as it is, and to encode a new character U+05BA HEBREW POINT QAMATS QATAN. Because the existing character was not changed, existing implementations are unaffected, and users that do not make the distinction can continue to use it, regardless of whether the

¹ This discussion applies only to disunification of individual characters, not the disunification of entire scripts, such as the decision to encode Coptic separately from Greek.

implementation also supports the new character or not. For users that *do* make the distinction, they use the existing character, though now in fewer instances and with a more restrictive meaning.

Disunification of YERAH BEN YOMO

The Hebrew mark *yerah ben yomo* is one of the accents from the Tiberian accentual system used by Masoretic scribes to indicate textual structure within verses and to provide guidance on the correct chanting of the text. Historically, two similarly-shaped but distinct accents were used, but the distinction was at some point lost. The distinction has been rediscovered in recent years, however, and some users now want to make the distinction in encoded texts.

The existing character U+05AA HEBREW ACCENT YERAH BEN YOMO was encoded without awareness of the distinction, and it has been used in contexts where the distinction is not made. Most users do not make the distinction in texts; for them, this existing character has, thus far, been adequate. Typically, the preferred glyph for users that do not make the distinction is roughly the shape of a small v optionally with a slight vertical stem at the bottom, though the name apparently means “day-old moon”, suggesting a crescent shape.



Figure 2. Glyphs for U+05AA YERAH BEN YOMO from three existing fonts



Figure 3. Nu. 35:5:5 (right) and Ps. 1:3:3 (left) from Snaitch's edition: no contrast between historically-distinct accents (*galgal*—blue highlight—and *atnah hafukh*—red highlight)

For users that *do* wish to make the distinction, two characters are needed: *galgal*, which has roughly a crescent or semi-circular shape, and *atnah hafukh*, which roughly has the shape of a small v with a slight vertical stem.²

² I have refrained from using the name *yerah ben yomo* when describing the situation in which two accents are distinguish, using an alternate name, to avoid any predispositions about which of the two distinct accents might be represented using the existing character.



Figure 4. Nu. 35:5:5 (right) and Ps. 1:3:3 (left) from *Biblia Hebraica Leningradensia*: *galgal* (blue highlight) and *atnah hafukh* (red highlight) are distinguished

The shape of one of the two distinct accents, *atnah hafukh*, matches the representative glyph of the existing character, YERAH BEN YOMO: the small-v shape with a vertical stem. Thus, one might expect that the existing character would be used for *atnah hafukh*, and that a new character, GALGAL, would be added, having a semi-circular shape. This is not what was proposed, however: because the name *yerah ben yomo* suggests a crescent shape, the proposers apparently felt that it would be inaccurate to have a character with that name but a small-v shape while another character was added with the crescent shape.

Thus, what was proposed, and what was accepted by UTC, was to change the representative glyph for the existing character U+05AA HEBREW ACCENT YERAH BEN YOMO to a semi-circular shape, and to encode a new character U+05A2 HEBREW ACCENT ATNAH HAFUKH with the small-v shape. For users that do make the distinction, YERAH BEN YOMO must have a semi-circular shape, but for users that *do not* make the distinction, a small-v shape is required.

Disunification of LATIN LETTER GLOTTAL STOP

The character U+0294 LATIN LETTER GLOTTAL STOP was encoded to represent the phonetic symbol *glottal stop* used in linguistic transcription. In phonetic usage, the character is drawn with a cap-height glyph, but no case distinction is made. At the time it was encoded, there was no usage known that involved a case distinction. Thus, the character name does not include “small” or “capital” as would be used for cased letters. For some reason, though, this character was assigned the general-category property *lowercase letter* (Ll) rather than *letter – other* (Lo).

'ruaħ hattsa'fon, vehaf'ʃemeʃ, hitvake'ħu bene't
jo'ter. game'ru, ki ʔet hannitsa'ħon, jin'ħal, 'mi |
ʃo'ver ʔoraħ ʔet bega'dav. pa'taħ 'ruaħ hattsa'fon v

Figure 5. *Glottal stop* in phonetic transcription: cap-height glyph used (IPA 1999, p. 98)

Certain languages with Latin-based orthographies do use glottal stop as a casing character, with an uppercase and lowercase pair. In these orthographies, the capital letter is displayed with a cap-height glyph, while the small letter is displayed with a glyph of roughly x-height.

Chĩa tɬ'i k'e dawheda ts'ɿʔo nãhdɔ hɔt'e.

ʔasii wizi whenehtà nì le.

Figure 6. Bi-cameral *glottal stops* in orthographic use: lowercase (red highlight) is x-height, uppercase (blue highlight) is cap-height (from Koyina 1983)

Because the existing character has a cap-height glyph, which is what is required for phonetic transcription, it was originally proposed to change the case property of the existing character to *uppercase* and to add a new *lowercase* letter SMALL GLOTTAL STOP with an x-height glyph. Concerns were raised, however, regarding potential problems for existing implementations if the case of the existing character were changed. (E.g. it could affect indexes, file systems or other protocols that use case mapping.)

Therefore, the proposal was changed to leave the existing character as is with its originally-intended usage for phonetic transcription, and to encode *two* new characters, a casing pair, for orthographic usage. The decision of UTC, however, was to leave the existing character as is, but to encode only one new character, U+0241 LATIN CAPITAL LETTER GLOTTAL STOP.

With this UTC decision, those that want to use the existing character U+0294 LATIN LETTER GLOTTAL STOP for phonetic transcription require a cap-height glyph, which is what would be found in existing font implementations. Those that want to use the pair of characters for orthographic purposes, however, require a font that has an x-height glyph for the existing character.

Comparison of the disunifications

The three disunifications described above differ in terms of the ease with which they can be implemented: the *qamats* disunification presents no problems, while the other two disunifications present significant dilemmas for implementers. The reason for the difference is that the disunification of *qamats* left the existing character completely unchanged, while the other two disunifications did not.

The representative glyphs for the characters in question before and after the disunifications are shown in Table 1:

Character	TUS 4.0	TUS 4.1
U+05B8 HEBREW POINT QAMATS		
U+05BA HEBREW POINT QAMATS QATAN	N/A	
U+05AA HEBREW ACCENT YERAH BEN YOMO		
U+05A2 HEBREW ACCENT ATNAH HAFUKH	N/A	
U+0294 LATIN LETTER GLOTTAL STOP		
U+0241 LATIN CAPITAL LETTER GLOTTAL STOP	N/A	

Table 1. Representative glyphs in TUS 4.0 and TUS 4.1

It should be noted that the glyph for LATIN CAPITAL LETTER GLOTTAL STOP does not actually correspond to what is, in fact, used. Rather, it is an invention, created specifically to provide a capital-like contrast to the representative glyph for the existing lowercase letter.

A better comparison can be seen by considering what glyphs are required in different usage contexts: by users that do not require a two-way distinction, and by users that do. This is shown in Table 2:

Character	No distinction required	Two-way distinction required	Note
U+05B8 HEBREW POINT QAMATS			
U+05BA HEBREW POINT QAMATS QATAN	N/A		
U+05AA HEBREW ACCENT YERAH BEN YOMO			
U+05A2 HEBREW ACCENT ATNAH HAFUKH	N/A		
U+0294 LATIN LETTER GLOTTAL STOP			Cap-height glyph required for phonetic transcription; x-height glyph required for orthographic usage.
U+0241 LATIN CAPITAL LETTER GLOTTAL STOP	N/A		

Table 2. Glyphs required in different usage contexts

Consider the impact of these disunifications for font vendors or product vendors that include fonts with their products (e.g. operating systems, business-app suites). First, in the case of

qamats and *qamats qatan*, implementing support for TUS 4.1 is not a problem: the new character can be added to a font with no effect on existing documents. The revised font will be useful both for existing scenarios in which no distinction was made and also for new scenarios in which a two-way distinction is made.

In contrast, for the other two disunifications, there is no easy way for the change to be implemented.³ The glyphs for the existing characters cannot be changed in existing fonts without having potentially-damaging effects on existing documents. The new characters could be added to existing fonts, but because the glyphs for the existing characters cannot be changed, the result will be that both the existing and new characters have the same glyphs, which is not particularly useful.

Even in new fonts, which are not encumbered by legacy usage, there is no way to support both usage scenarios: in order to know what glyphs are needed for the existing characters, it must first be known whether the user does or doesn't make the two-way distinctions. The only real options are:

- create fonts that can only work for one usage scenario or the other; or
- create fonts that use the same default glyph for both existing and new characters with an alternate glyph for the existing character selectable by a font feature – but the two-way distinction will be available only in certain applications that support font-feature mechanisms.

For instance, after reviewing the disunification of *yerah ben yomo*, John Hudson (Tiro Typeworks) concluded that the best option for implementing the new character ATNAH HAFUKH was to use the same default glyph for both U+05AA YERAH BEN YOMO and U+05A2 ATNAH HAFUKH, and provide an alternate glyph for U+05AA for use when *galgal* is distinguished from *atnah hafukh*, selectable using an OpenType feature. John recently commented on this disunification on the Unicore list:⁴

“...the proposed disunification of *yerah ben yomo*... raises some problems at the display level, since in this case it is the existing character for which a glyph change would be required by users desiring to make the distinction visual... [This] is a problem we should have spotted when the new character was first proposed... But the fact that we failed to identify the problem early does not mean that the problem does not exist.

“My current inclination is to use the *etnah hafukh* glyph as default for both characters, and to handle the distinct form of *yerah ben yomo* as a glyph variant associated with a stylistic alternate feature. This is not ideal, since it requires a

³ The problem cannot be described as breaking *existing implementations*, since existing fonts can continue to be used in the same ways they were used before without any issues. Rather, the problem is that both of the post-disunification characters cannot be easily implemented, with potential for *new implementations*—revised or new fonts—to break existing documents.

⁴ Quoted from a message from John Hudson to the Unicore list, January 27, 2005, on the subject “QAMATS QATAN and HOLAM HASER FOR VAV”.

fairly sophisticated level of glyph substitution support from apps in order to handle what should be a fairly straight forward distinction between two characters.”

Some fonts are designed with specific uses in mind, and for such fonts the first option makes sense. This may be sufficient, for instance, for publishers of Hebrew religious texts who require a contrast between *galgal* and *atnah hafukh*. But this is the exceptional case: most users depend on fonts designed for general-purpose usage. Certainly for a platform vendor, such as Microsoft, fonts need to support as broad a range of uses as possible, and having to choose, for instance, between supporting phonetic transcription or the orthographies of living languages is a problem.

Avoiding the problems

It should be reasonably clear that the key factor that differentiates the *qamats* disunification from the other two is that it did not involve any change to the existing character, with only the new character requiring a different glyph. This was not the case with the other disunifications: the *yerah ben yomo* disunification involved a change in the representative glyph for U+05AA, and both resulted in a situation in which the existing character requires distinct glyphs depending on the usage.

In the case of *yerah ben yomo*, this could easily have been avoided by handling the disunification in a different way, as shown in Table 3:




Character	TUS 4.0	TUS 4.1
U+05AA HEBREW ACCENT YERAH BEN YOMO		
U+05XX HEBREW ACCENT GALGAL	N/A	

Table 3. Possible alternate disunification of *yerah ben yomo*

Reportedly, this alternative was considered by the proposers but abandoned since it would result in a less-than-ideal relationship between the name and glyph for U+05AA. End users are not the primary intended audience for character names, however, and less-than-ideal names can be mitigated by annotations or explanatory text in block descriptions. The cost of preserving the best possible name-glyph relationship has been the problems now faced by implementers, costs that will also be borne by end users.

It is too late to change the *yerah ben yomo* disunification, but the aforementioned problems associated with it can perhaps still be remedied by adding GALGAL as a second new character:





Character	TUS 4.0	TUS 5.0	Comment
U+05AA HEBREW ACCENT YERAH BEN YOMO			Used only in scenarios in which the two-way distinction is not made.
U+05A2 HEBREW ACCENT ATNAH HAFUKH	N/A		Used only in scenarios in which <i>atnah hafukh</i> is distinguished from <i>galgal</i> .
U+05xx HEBREW ACCENT GALGAL	N/A		

Table 4. Possible revised disunification of *yerah ben yomo*

This remedy to the current situation would have as a disadvantage that there would be two characters with the same glyph; in effect, one of the two characters would lose any useful purpose. That would simply have to be considered the price of having handled the initial disunification poorly. Arguably, this would be less problematic than the current situation since there are ways, at least, that the effective duplication can be dealt with in implementations, whereas there are no good ways for implementations to deal with the current situation.

The more important point, though, is that the need to create a situation in which one character becomes fully redundant could have been avoided in this case had there been a set of guiding principles for disunification in place beforehand.

The *glottal stop* disunification was a more difficult case. In terms of the glyphs needed for different usage contexts, it would have been adequate to make the existing character the capital in orthographic usage and add only one new character for the small glottal stop, but this was not a viable option because of problems related to changing the case property of the existing character. There was another alternative, though, which still remains as a possible remedy for the current problems: encode two new characters:





Character	TUS 4.0	TUS 5.0	Comment
U+0294 LATIN LETTER GLOTTAL STOP			Used only for phonetic transcription, or in orthographies without bi-cameral glottal stops.
U+0241 LATIN CAPITAL LETTER GLOTTAL STOP	N/A		Used only for orthographies that have bi-cameral glottal stops.
U+xxxx LATIN SMALL LETTER GLOTTAL STOP	N/A		

Table 5. Possible alternative / revised disunification of *glottal stop*

Again, there is a *visual* duplication of characters, though this duplication is only partial (unlike the situation that would hold for *yerah ben yomo* and *atnah hafukh*) since, in this case, the two characters would have distinct case properties. The visual duplication would be less than ideal, but it appears to be the only possible option that avoids the implementation problems described above.

In considering how these two disunifications could have been done without creating the implementation problems mentioned above, and how the disunifications can still be revised to remedy those problems, we find certain principles.

First, we do not want to disunify existing characters in a manner that entails changes to the glyphs of existing characters, since that is central to the problems that have been described.

More generally, we want to ensure that, as much as possible, viable uses of the existing character prior to disunification remain viable after the disunification.

In this regard, note that pre-disunification use of U+05AA specifically for *atnah hafukh*, with a private-use code point for the contrasting character *galgal*, would be viable, and this use of U+05AA would remain viable if a new character GALGAL were added. Thus, the possible disunification shown in Table 4 would have been workable. On the other hand, a pre-disunification use of U+0294 for an orthographic capital letter, with a private-use character for the lowercase counterpart, would not really have been viable because of the case property of U+0294; thus, it should not be essential to preserve that usage in a disunification of *glottal stop*.

Another principle we find is that, if it is not possible to add only one new character for the *distinct* letterform that is needed, then two new characters should be added, as that is the only way to create a distinction without creating implementation problems in relation to the existing character. This results in a visual or complete duplication of characters, and should be avoided if possible, but it must be recognized that there may be situations in which implementation problems cannot be avoided without such duplication. For instance, the existing character LATIN LETTER GLOTTAL STOP could not be used as the capital of a casing pair because its existing case property is *lowercase*, hence it was not possible to add only one new character for the x-height SMALL GLOTTAL STOP. The only way to create the distinction, then, without creating implementation problems in relation to the existing character is to add two new characters.

A further principle to note is that we must not give higher priority to a desire to have appropriate names for characters than we give to the impact on implementations. For instance, using the name YERAH BEN YOMO for *atnah hafukh* certainly would not be ideal, but that would be a much less serious concern than the problems now presented to implementers in how to support the existing and new characters.

Proposed principles on character disunification

In light of the preceding discussion, I propose that UTC adopt the following principles to be applied whenever a character disunification is being considered. These would be applied as general guidelines that could be overridden, but only after careful consideration.

When disunifying an existing character in the UCS, the following principles will be observed:

1. As much as possible, viable uses of existing characters prior to disunification will be preserved after disunification.
2. The normative properties of existing characters will not be changed.

3. The representative glyphs of existing characters will not be changed, and the range of glyphs expected for existing characters will not increase as a result of disunification.
4. If a character disunification cannot be achieved by adding one new character without requiring a change in normative properties of the existing character and without changing the representative glyph or range of expected glyphs for the existing character, then new characters will be added for each of the distinct, specific letterforms required; the existing character will not be intended for use in scenarios in which the distinct, specific letterforms are used. This may result in visually-duplicate characters, which in general should be avoided if possible, but may be necessary under the aforementioned conditions.
5. While it is desirable that a character name be fully appropriate to the given character and its representative glyph, concern over less-than-ideal names will not provide a sufficient basis for overriding principles 1 to 4, above.

Exceptions to these principles will be permitted only after careful consideration and on the basis of substantial rationale.

If these principles are accepted by UTC, I would further recommend that they be proposed for inclusion in WG2's *Principles and Procedures* document.

References

- Dotan, Aron, ed. 2001. *Biblia Hebraica Leningradensia*. Peabody, MA: Hendrickson Publishers.
- International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.
- Koyina, Laiza. 1983. *Dq weda goðle xè Teèt'o si. (The Blind Man and the Loon.)* Yellowknife, NWT, Canada: Northwest Territories Department of Education.
- Snaith, Norman Henry. 1982. *תורה נביאים וכתובים. (Hebrew Old Testament.)* London: British and Foreign Bible Society.