L2/03-195R

Proposal to Encode Alternative Characters for Biblical Hebrew

Date: 2003-06-09

Author: Peter Constable, SIL International

Address: 7500 W. Camp Wisdom Rd.
Dallas, TX 75236
USA

Tel: +1 972 708 7485

Email: peter_constable@sil.org

A. Administrative

1.	Title	Proposal to Encode Alternative Characters for Biblical Hebrew
2.	Requester's name	Peter Constable, SIL International; John Hudson, Tiro Typeworks; Eli Evans, Logos Research Systems; Kent Richards, Society of Biblical Literature; Paul Nelson, Microsoft; Ralph Hancock; Kirk Lowery, Westminster Hebrew Institute
3.	Requester type	Expert contribution
4.	Submission date	2003-06-09
5.	Requester's reference	
6a.	Completion	This is a complete proposal. (Other proposals for characters needed for Biblical Hebrew will be forthcoming, but this proposal is complete in and of itself.)
6b.	More information to be provided?	Only as required for clarification.

B. Technical—General

1a.	New Script? Name?	No
1b.	Addition of characters to existing block? Name?	These can be included in the existing Hebrew block. Alternately, a distinct block might be considered in order to distinguish these characters, which would be intended for Biblical Hebrew only.
2.	Number of characters in proposal	14
3.	Proposed category	D
4.	Proposed level of implementation and rationale	2 (proposal includes combining characters, but not any that are listed in B.2 of ISO 10646)
5a.	Character names included in proposal?	Yes
5b.	Character names in accordance with guidelines?	Yes
5c.	Character shapes reviewable?	Yes
6a.	Who will provide computerized font?	Either SIL International or Tiro Typeworks can provide a font, if needed.
6b.	Font currently available?	Yes
6c.	Font format?	TrueType

7a.	Are references (to other character sets, dictionaries, descriptive texts, etc.) provided?	Yes
7b.	Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?	Yes
8.	Does the proposal address other aspects of character data processing?	Yes, suggested character properties are included (see section E).

C. Technical—Justification

1.	Has this proposal for addition of character(s) been submitted before?	No
2a.	Has contact been made to members of the user community?	Yes
2b.	With whom?	Biblical Hebrew researchers, content providers of materials related to Biblical Hebrew
3.	Information on the user community for the proposed characters is included?	These characters would be used by scholars in the field of Biblical Hebrew studies.
4.	The context of use for the proposed characters	Corpora for Biblical Hebrew text research, software products providing Biblical Hebrew text content, scholarly publications (commentaries, journals, etc.)
5.	Are the proposed characters in current use by the user community?	Yes
6a.	Must the proposed characters be entirely in the BMP?	Preferably, though not necessarily.
6b.	Rationale?	These could be kept with existing Hebrew characters (though there may be benefits in having them in a distinct block for Biblical Hebrew).
7.	Should the proposed characters be kept together in a contiguous range?	Yes
8a.	Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	No
8b.	Rationale for inclusion?	n/a
9a.	Can any of the proposed characters be considered to be similar (in appearance or function) to an existing character?	Yes
9b.	Rationale for inclusion?	See discussion in section F.
10.	Does the proposal include the use of combining characters and/or use of composite sequences?	Yes
11.	Does the proposal contain characters with any special properties?	No

D. SC2/WG2 Administrative

1. Relevant SC2/WG2 document numbers
2. Status (list of meeting number and corresponding action or disposition)
3. Additional contact to user communities, liaison organizations, etc.
4. Assigned category and assigned priority/time frame

Other comments

E. Proposed Characters

A code chart and list of character names are shown on a new page. Code positions within the existing Hebrew block are suggested. Existing characters are included in the chart for reference, shown in pale blue.

E.1 Code Chart

	05E	05F
0	נ	ׯ
1	ס	װ
2	ע	ױ
3	ף	׳
4	פ	״
5	ץ	◌ֵ
6	צ	◌ֶ
7	ק	◌ַ
8	ר	◌ָ
9	ש	◌ֹ
A	ת	◌ֺ
B	◌ִ	◌ֻ
C	◌ּ	◌ֽ
D	◌ׁ	◌ֿ
E		◌ׂ
F		◌ׄ

E.2 Character Names

05EB BIBLICAL HEBREW POINT METEG-SILLUQ
05EC BIBLICAL HEBREW POINT SHIN DOT
05ED BIBLICAL HEBREW POINT SIN DOT

05F5 BIBLICAL HEBREW VOWEL SCHWA
05F6 BIBLICAL HEBREW VOWEL HATAF SEGOL
05F7 BIBLICAL HEBREW VOWEL HATAF PATAH
05F8 BIBLICAL HEBREW VOWEL HATAF QAMATS
05F9 BIBLICAL HEBREW VOWEL HIRIQ
05FA BIBLICAL HEBREW VOWEL TSERE
05FB BIBLICAL HEBREW VOWEL SEGOL
05FC BIBLICAL HEBREW VOWEL PATAH
05FD BIBLICAL HEBREW VOWEL QAMATS
05FE BIBLICAL HEBREW VOWEL HOLAM
05FF BIBLICAL HEBREW VOWEL QUBUTS

E.3 Unicode Character Properties

All of the proposed characters should have a general category of Mn. The canonical combining classes should be as follows:

Character		Canonical combining class
05EB	BIBLICAL HEBREW POINT METEG-SILLUQ	220
05EC	BIBLICAL HEBREW POINT SHIN DOT	10 (some value less than that of dagesh)
05ED	BIBLICAL HEBREW POINT SIN DOT	11 (some value less than that of dagesh)
05F5	BIBLICAL HEBREW VOWEL SCHWA	220
05F6	BIBLICAL HEBREW VOWEL HATAF SEGOL	220
05F7	BIBLICAL HEBREW VOWEL HATAF PATAH	220
05F8	BIBLICAL HEBREW VOWEL HATAF QAMATS	220
05F9	BIBLICAL HEBREW VOWEL HIRIQ	220
05FA	BIBLICAL HEBREW VOWEL TSERE	220
05FB	BIBLICAL HEBREW VOWEL SEGOL	220
05FC	BIBLICAL HEBREW VOWEL PATAH	220
05FD	BIBLICAL HEBREW VOWEL QAMATS	220
05FE	BIBLICAL HEBREW VOWEL HOLAM	27 (some value greater than that of rafe)
05FF	BIBLICAL HEBREW VOWEL QUBUTS	220

Table 1. Canonical combining classes of proposed characters.

The proposed classes for BIBLICAL HEBREW POINT SHIN DOT, BIBLICAL HEBREW POINT SIN DOT and BIBLICAL HEBREW VOWEL HOLAM are intended only to indicate the intended relative ordering of Biblical Hebrew characters. There are existing characters with these fixed position classes, and there is no intent to suggest that the characters proposed here belong in the same fixed position class with other existing characters. It may be necessary to adjust the numeric value of fixed position classes (maintaining the order for existing characters) to create gaps into which these new characters can be placed.

All other properties should match those of other similar characters, such as U+0591 HEBREW ACCENT ETNAHTA.

F. Other Information

The proposed characters duplicate existing characters in the Hebrew block in order to overcome inadequacies in relation to encoding of Biblical Hebrew text. It is proposed that the existing characters would continue to be used for modern Hebrew and Yiddish, and that existing mappings from industry legacy Hebrew character sets would remain as presently defined; but that the new characters would be used for Biblical Hebrew.[1] The domain of usage for the two groups of characters would therefore be distinct.

The existing Hebrew characters are considered inadequate for encoding of Biblical Hebrew. This is due to the canonical combining classes to which they are assigned (each of the existing characters is assigned to a unique fixed position class). Because of the defined combining classes, any sequence involving some combination of these characters is canonically equivalent to every ordering permutation of the same characters. For instance, the sequence < U+05B7, U+05B4 > is canonically equivalent to the sequence < U+05B4, U+05B7 >, and the sequence

[1] Mappings from legacy encoding systems for Biblical Hebrew that have been developed within the academic sector would use the new characters.

< U+05B5, U+05BD > is canonically equivalent to the sequence < U+05BD, U+05B5 >. As a result, different orderings of these characters effectively cannot be represented in Unicode in that they cannot be reliably preserved in data interchange. It is a essential requirement for Biblical Hebrew text, however, to be able to represent specific orderings of vowel combinations, or combinations of vowels with meteg.

F.1 Vowel combinations

Normally, combinations of vowel marks on a single consonant would not occur in Hebrew script. Combinations of vowels can occur in Biblical Hebrew text, however, as a result of phonological changes over time combined with a strict policy of not changing the consonantal framework of the text:

> These examples show that the Masoretes added their vowels to a consonantal framework which they did not allow themselves to alter. This is also shown by the constant spelling of יְרוּשָׁלִַם (in the printed editions: יְרוּשָׁלִַם, e.g., Josh 10:1), reflecting as it were *yᵉrušālaim*. This vocalization indicates that in their manuscripts the Masoretes found the ancient form ירושלם (= יְרוּשָׁלִֵם, ***yᵉrušālēm***) and that they added the *hireq* between the *lamed* and the final *mem* because they could not change the consonantal text by adding a *yod*. The addition was meant to accomodate the pronunciation *yᵉrušālayim* which had become

Figure 1. Vowel combinations arising from changing vocalization and a constant consonantal framework (Tov 1992, p. 43).

It is necessary, therefore, to encode sequences such as the following:

< lamed, qamets, hiriq, final mem >

Using existing characters, this would be represented as follows:

< U+05DC HEBREW LETTER LAMED,
U+05B8 POINT QAMETS,
U+05B4 POINT HIRIQ,
U+05DD HEBREW LETTER FINAL MEM >

Note, however, that this sequence is transformed under canonical ordering and normalization to the following:

< U+05DC HEBREW LETTER LAMED,
U+05B4 POINT HIRIQ,
U+05B8 POINT QAMETS,
U+05DD HEBREW LETTER FINAL MEM >

Thus, the specific ordering of the vowels cannot be preserved using existing characters. Moreover, for the particular combinations of vowels that do occur in Biblical Hebrew text, such as qamets + hiriq, the order that is produced under canonical ordering and normalization is exactly the opposite of the order that is required.

A reliable means of encoding particular vowel combinations in particular orders is required in order to provide adequate representation of the text, both for research purposes, and also to facilitate publishing of texts.

F.2 Combinations of vowels with meteg

The Hebrew character *meteg* (also called *gaʿyah*, literally "raising" of the voice) is part of the Tiberian accentuation system developed by Masoretic scribes. Its function has been described variably as denoting "secondary stress" (Tov 1992, p. 68), or that "the reading of the syllable on which it is marked is to be slowed down, and not slurred over" (Yeivin 1980, p. 242).

Meteg has the same visual form as *silluq*, which is also part of the Tiberian accent system and is used to mark the end of the second of two major subdivisions of each verse. Since the meteg and silluq are visually identical and have complementary distribution (they occur in distinct parts of the verse and never co-occur on the same syllable), the same encoded character can be used to represent both graphemes. Where appropriate, I will continue to discuss meteg and silluq individually, but otherwise will hereafter refer to them together as *meteg-silluq*.

Both meteg and silluq very frequently co-occur together with vowel marks, though they can also occur without a vowel. When they do co-occur with a vowel, they are usually written to the left of the vowel; in the case of hataf vowels (hataf segol, hataf patah, hataf qamets), at least in the *Biblia Hebraica Stuttgartensia* (BHS), they are usually written between the two components of the hataf vowel.

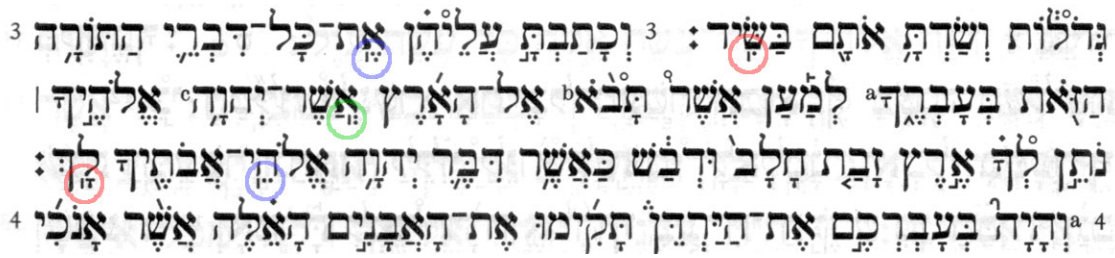

Figure 2. Silluq left of vowel (red highlight); meteg left of vowel (blue) and medial in hataf vowel (green); from Deut. 27:2–4a (BHS).

This relative ordering of meteg-silluq and vowels is not always maintained, however. Yeivin (1980) discusses the matter of positioning of meteg (ga‘yah) briefly:

> 314. *Ga^{c}ya* is generally written to the left of a vowel sign marked under the same letter. In some MSS, such as A and L, this convention is carefully maintained, with very few exceptions--and those usually due to correction, or to lack of space in the regular position. In other MSS, such as C, S, S^{1}, *ga^{c}ya* is often written to the right of the vowel sign, without any particular reason. With *shewa* also, *ga^{c}ya* is generally written to the left of the *shewa* sign, but there are MSS in which it is often written to the right. The same is generally true of *ḥaṭef shewa* signs, but in some MSS, such as L and S^{1}, the *ga^{c}ya* sign may be written between the two parts of the *ḥaṭef* sign; -ֲ-. This does not occur in A and C.

Figure 3. Vowel combinations arising from changing vocalization and a constant consonantal framework (Tov 1992, p. 43).

In printed editions of the Hebrew text, some editors have used alternate positioning of meteg-silluq to reflect editorial decisions. This was true of Kittel’s editions (BHK):

> Increasingly as we have become conscious that in L we have a reliable ben Asher text, we have attempted to reproduce this text just as it stands in the MS. I had also made arrangements with Rudolf Kittel that beginning with the Psalms, we would put all Methegs which occur in the MS. to the left of the vowel, but on the other hand we would place the Methegs which we felt ought to be added (in accordance with the statements on page XXVII of Kittel's Foreword) to the right of the vowel, so that the reader can recognize immediately where the MS. has a Metheg, and where we have added one. At the same time we decided to reduce the inserted Methegs to a minimum. (BHK, p. xxxi)

In the BHS edition, the editors have chosen to preserve positioning found in the Leningrad Codex:

> The most important differences between BHS and BHK are the following:…
>
> 2. TEXT. We have thought it best to reproduce the text of the latest hand of L with close fidelity… The addition of Silluq, which is occasionally lacking, and particularly of Metheg, which is often omitted, has been discontinued, particularly as in L itself Metheg is found both to the left and the right of the vowel pointing, and Silluq may also appear to the right. (BHS, p. xii)

Accordingly, alternate positions of silluq and meteg relative to vowel points are found in the BHS, as was true of the BHK for different reasons. So, for example, compare the combination of hataf patah and medial meteg, seen in Figure 2 above, with combinations of hataf patah and meteg in left and right positions:

Figure 4. Hataf patah with right-side meteg (Psalm 85:7, BHS).

Figure 5. Hataf patah with left-side meteg (Job 39:11, BHS).

Similarly, compare the combination of segol with left-side meteg in Figure 2 with the combination of segol and right-side meteg in Figure 6; and the combination of qamats with left-side silluq in Figure 2 with the combination of qamats and right-side silluq in Figure 7:

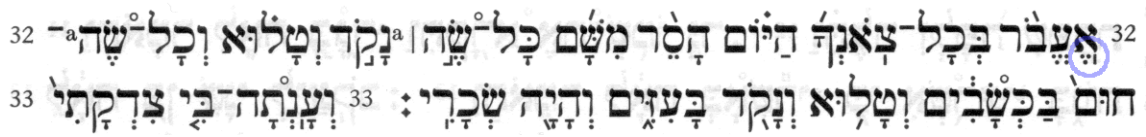

Figure 6. Segol with right-side meteg (Gen 30:32, BHS).

Figure 7. Qamats with right-side silluq (Ps 79:12, BHS).

One of the dictums of researchers involved in electronic encoding of Biblical Hebrew texts is to "encode what is *written* not what is *meant*".[2] Of course, there are limits to the kinds of visual distinctions that are appropriate for character representation. Distinctions of the sort illustrated here, however, are very definitely among the distinctions that are appropriate for representation in terms of character encoding, and that researchers are wanting to represent in terms of character encoding. In fact, encoding these very distinctions is well established practice within Biblical Hebrew encoding projects:

> Accent 75 serves both for silluq and for meteg when meteg occurs (as it does most often) to the left of its vowel. Accent 95 is reserved for meteg when it occurs to the right of its vowel, and 35 codes a meteg which falls between the components of a hatep vowel as at Judges 9:27. (Parunak 1982, §3.5.1)

It is not only scholars that need to represent such ordering distinctions in texts: publishers also need to be able to represent such distinctions, as evidenced by publications such as BHK and BHS.

Accordingly, it is considered necessary that a means be provided of representing such distinctions in the UCS.

[2] See §3.3.2 of Parunak (1982).

F.3 *Alternative solutions?* No!

In principle, what is needed in order to provide a means for representing the ordering distinctions described in § F.1 and §F.2 is that all vowels other than holam and also meteg-silluq be in the same the canonical combining class. There are two possible ways in which this can achieved: revise the canonical combining classes of the existing characters, or encode new characters with the necessary canonical combining classes. In an ideal world, we would consider the former solution to be preferable. This solution would have the effect of changing normalization forms, however, in violation of point 3.e of the Unicode Standard Stability Policy.[3] It is, therefore, not a viable alternative. Thus, the only solution that can meet the needs for encoding of Biblical Hebrew is to encode new, duplicate characters, as proposed here.

F.4 Sin/shin dots, holam

The preceding discussion has provided the rationale for all of the proposed characters other than sin dot, shin dot and holam. The rationale for proposing these three characters is presented here.

The canonical combining classes for the existing sin/shin dot and holam characters are sufficient for the basic requirement of representing necessary distinctions in the text. They are considered problematic, however, for other purposes related to implementation and usability.

In Biblical Hebrew text, it is common to have multiple combining marks co-occuring with a single base character. I do not know of the actual upper limit in existing corpora, but sequences involving three combining marks are quite frequent, sequences with four are not rare, and sequences with five or even six are certainly plausible. In these multi-mark sequences, it will typically be the case that each combining mark is in a distinct canonical combining class (the preceding discussion on the need for vowels and meteg to be in the same class notwithstanding). The effect of this is that a given document of Biblical Hebrew text can have a *vast* number of canonically equivalent representations, each different from the other only in the ordering of combining sequences.

It can be a significant burden on processes to deal with all of these alternate orderings. This is a particular concern in relation to rendering. For instance, it would be a difficult challenge for a font developer creating an OpenType Hebrew font to accommodate all of the possible orderings of combining marks, and if they managed to do this in their font, the number of rules to be processed could result in unacceptably slow rendering. The need for font developers to do this could be avoided if layout engines such as Uniscribe were to re-order the combining marks into canonical order, but even that amount of processing on a page of text could result in unacceptably slow rendering if the text contained a significant number of sequences not in canonical order.

There is a well-established practice among Biblical Hebrew users regarding the ordering of Hebrew combining marks. This order is *decidedly not* the same as the canonical order defined by existing combining classes. It should be noted that the proposed alternative characters for vowels and meteg discussed above are only a small factor in this: they are only a part in the overall sequence. Thus, the ordering in established usage is rougly as follows:

consonant < shin / sin dot < dagesh / rafe < vowel < meteg / accent

The canonical ordering of currently-defined characters, however, is substantially different:

consonant < vowel < dagesh < meteg < rafe < shin / sin dot < accent

In terms of usability, the canonical order would not be acceptable to users: they simply would not enter and edit text in this order, and would likely reject software that required them to do so. Yet developers of fonts and layout engines cannot assume the established order preferred by users since canonical order is very likely to be encountered, particularly in light of W3C recommendations.

Therefore, there is a very strong desire on the part of implementers (particularly of fonts and rendering systems) that there be one preferred order of Biblical Hebrew combining sequences and that this be the canonical order, and

[3] See http://www.unicode.org/standard/stability_policy.html.

a very strong desire on the part of Biblical Hebrew users that this preferred order be the order that is already well established.

The proposed alternative characters for vowels and meteg partially resolves this ordering issue: by putting the vowels (other than holam) and meteg into class 220, the resultant ordering (still using existing characters for dagesh, rafe and shin / sin dot) would be as follows:

consonant < holam < dagesh < rafe < shin / sin dot < vowel / meteg < accent [4]

The relative ordering of holam, dagesh / rafe and shin / sin dot would still face challenges of unacceptability from users, however.

The problems related to ordering can be entirely resolved by encoding alternative characters for holam, shin dot and sin dot for use specifically with Biblical Hebrew. The three additional characters would allow for an ordering that basically matches established practice:[5]

consonant < shin / sin dot < dagesh / rafe < vowel / meteg-silluq < accent

Given the unavoidable necessity of encoding the eleven alternative characters for Biblical Hebrew vowels and meteg as discussed above, it does not seem that there would be any significant additional detriment by encoding these three other alternative characters. We suggest that the increase in detriment would be marginal, whereas the increased benefit to Biblical Hebrew implementers and users would be significant.

G. References

Elliger, K.; W. Rudolph; and A. Schenker; eds. 1997. *Biblia Hebraica Stuttgartensia, 5th revised edition.* Stuttgart: Deutsche Bibelgesellschaft. ("BHS")

Kittel, Rudolph, ed.. 1952. *Biblia Hebraica 3, 7th edn.*. Stuttgart: Privileg. ("BHK")

Parunak , H. Van Dyke. 1982. "Code Manual for the Michigan Old Testament." Published online at http://www.wts.edu/hebrew/whmcodemanual.html.

Tov, Emanuel. 1992. Textual criticism of the Hebrew Bible, 2nd edn. Minneapolis: Fortress Press.

Yeivin, Israel. 1980. *Introduction to the Tiberian masorah.* (*The Society of Biblical Literature masoretic studies*, 5.) Translated by E.J. Revell. Scholars Press.

[4] The non-holam vowels and meteg would be in the same class as most of the below accents, and so alternate orders of these vowels or meteg with below accents would be distinct under normalization. This is an acceptable, even desirable, result.

[5] There is a minor difference from the established ordering in that meteg would be ordered together with vowels whereas existing legacy practice has meteg always after vowels. In the legacy encoding systems, however, distinctions in visual order are represented in terms of distinct code points, whereas in Unicode they would be represented in terms of alternate orderings. Thus, this difference between the established ordering and the ordering that would be achieved by this proposal is insignificant.