# Issues in the Representation of Pointed Hebrew in Unicode

**L2/03-299**

Third draft, Peter Kirk, August 2003

## *1. Introduction*

The Hebrew block of the Unicode Standard (http://www.unicode.org/charts/PDF/U0590.pdf) is intended to include all of the characters needed for proper representation of Hebrew texts from all periods of the Hebrew language, including fully pointed and cantillated ancient texts such as that of the Hebrew Bible. It is also intended to cover other languages written in Hebrew script, including Aramaic as used in biblical and other religious texts[1] as well as Yiddish and a few other modern languages.

In practice there are a number of issues and minor deficiencies in the Hebrew block as currently defined, in version 4.0 of the Unicode Standard (http://www.unicode.org/versions/Unicode4.0.0/), which affect its usefulness for representation of pointed Hebrew texts and of Hebrew script texts in some other languages. Some of these simply require clarification and agreed guidelines for implementers. Others require further discussion and decision, and possibly additions to the Unicode standard or other action by the Unicode Technical Committee. The conclusion reached in this paper is that two new Unicode characters should be proposed; other issues can be resolved by use of suitable sequences of existing characters, provided that such use is generally agreed by content providers and rendering systems.

Several of these issues relate to different typographical conventions for publishing of Hebrew texts. It seems that a particular set of conventions is used for general publications in Hebrew, especially in Israel, but various other conventions, in which more fine distinctions are made, are used mainly for quality editions of biblical and other religious texts. A major aim of this paper is to document these different conventions, and to define ways in which the finer distinctions made in the latter conventions can be supported in Unicode without increasing complexity for those who use the former set of conventions.

The text below refers in several places to Convention A# and Convention B#, where # stands for a digit. These are independent conventions for each issue. For each issue Convention A# is the one commonly used for general publication in Israel, and Convention B# is the one used in BHS[2], the most widely used scholarly edition of the Hebrew Bible. But there are many other publications which use some of the Conventions A# and some of the Conventions B#.

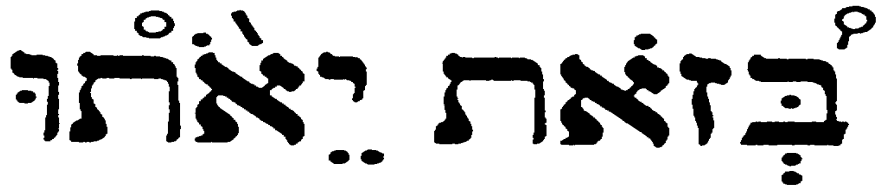
For conciseness abbreviated names are used for Unicode characters and sequences. For the full names and Unicode code points, see Appendix A.

## *2. Issues relating to all pointed Hebrew texts*

These issues may be encountered in the encoding of any Hebrew text with *vowel points*, including modern Hebrew texts in which *vowel points* are written as an aid to pronunciation. They may also apply to texts in other languages written in Hebrew script.

### 2.1. *Holam* and *alef*

The *vowel point holam* consists of a dot which is usually placed either above the top left corner of any Hebrew base character or a little further to the left, above the space between this base character and the following (to the left) one, to indicate a long O sound following that of the base character. By Convention B1 but not by Convention A1, when the base character is *alef* and is not word initial, the point *holam* may also appear above its top right corner; this also indicates a long O sound, and the *alef* is not pronounced. This *holam* is not in fact logically associated with the *alef*, but is associated with the preceding base character. It is shifted from above and to the left of the preceding (to the right) letter to above the top right of the *alef* as a typographical convention. This shift generally takes place only when the *alef* is silent, which is when there is no *vowel point* or *dagesh* combined with it and it is not followed by *vav shruqa* or *holam male*.



bəzō't yē'ōtû *"on this [they] will agree"*, from Genesis 34:22 BHS[3]

<*bet*, *sheva*, *dagesh*, *telisha gedola*, *zayin*, *holam*, *alef*, *tav*, *space*, *yod*, *tsere*, *alef*, *holam*, *qadma*, *tav*, *masora circle*, *vav*, *dagesh*>[4]
*Holam* above the right of *alef* (right hand word) and above the left of *alef* (left hand word)

In principle the options for encoding of collocations of *holam* and *alef* are the same as for collocations of *holam* and *vav* as described below. In practice the encoding described here is already in general and uncontroversial use, and so there is no good reason to change it. However, if option 5 below is chosen for *holam* and *vav*, it might be sensible to use the same new character also when *holam* appears above the right of *alef*.

### *Encoding guidelines*

When a collocation of *alef* and *holam* is intended to be pronounced as the consonant sound *alef* followed by the vowel sound *holam*, and the *holam* is intended to be positioned above and to the left of *alef*, the collocation should be encoded as <*alef*, *holam*>. When Convention B1 is in use and a collocation of *alef* and *holam* is intended to be pronounced as the vowel sound *holam* alone with the *alef* not pronounced, with the *holam* positioned over the top right of *alef*, the collocation should be encoded as <*holam*, *alef*>, with the *holam* combined with the preceding base character. Where, exceptionally, a *holam* should not be shifted on to a following *alef* regardless of the Convention, the encoding <*holam*, *ZWNJ*, *alef*> should be used. Where, exceptionally, a *holam* should be shifted on to a following *alef*, when Convention B1 is used, in a context where this shift would not normally happen, the encoding <*holam*, *ZWJ*, *alef*> should be used. The encoding <*RLM*, *holam*, *alef*>[5] may be used for special purposes to represent an isolated or word initial *alef* with *holam* above its top right, but will be rendered as such only by Convention B1.

### *Rendering guidelines*

Convention A1

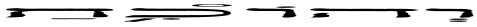
No special rendering is required for collocations of *alef* and *holam*, as *holam* is never shifted on to following *alef*. The sequence <*RLM*, *holam*, *alef*> should be rendered as an isolated *holam* followed by an *alef*.

Convention B1

When a composite character sequence including *holam* is followed by *alef*, the *holam* should be shifted in rendering from its regular position and rendered above the top right of the *alef*. But this shift should not take place when the *alef* is combined with any *vowel point* or with *dagesh*. It should also not take place when the *alef* is followed immediately by *vav shruqa*[6] or *holam male*. It should never take place when the *alef* is preceded by *ZWNJ*, but always when the *alef* is preceded by *ZWJ*. The sequence <*RLM*, *holam*, *alef*> should be rendered as an *alef* with a *holam* above its top right. When a *holam* follows an *alef* within the same composite character sequence, the *holam* should be rendered in the regular way above and to the left of the *alef*.

## 2.2. Furtive *patah*

Furtive *patah* is a *patah* or short A vowel sound pronounced in Hebrew before the consonants *ayin*, *het*, and *he* with *dagesh*, when these are word final or followed by *maqaf*. Although furtive *patah* is pronounced before the word final consonant, it is represented by a *patah* glyph positioned under this final consonant. By Convention B2 but not by Convention A2, furtive *patah* is positioned under the right side of the final consonant, thus distinguishing it from regular *patah* which is centred under the consonant. In Hebrew any *patah* which appears under the final base character of a word, or under a base character followed by *maqaf*, is a furtive *patah*, but this rule may not apply to other languages written in Hebrew script.[7]



wəhôkiax *"and [he] was complaining"*, from Genesis 21:25 BHS
<*vav*, *sheva*, *he*, *holam*, *vav*, *kaf*, *hiriq*, *merkha*, *masora circle*, *het*, *patah*>
Furtive *patah* displaced to the right

Furtive *patah* is generally encoded in Unicode and in legacy encodings as *patah* following the word final base character, although this does not correspond to the pronunciation order. Any change to a more logical encoding would further complicate the issue of multiple *vowel points* described below, and so no change is suggested.

***Encoding guidelines***

Furtive *patah* should be encoded as *patah* following the word final consonant.

***Rendering guidelines***

Convention A2

Furtive *patah* is rendered as any other *patah*.

Convention B2

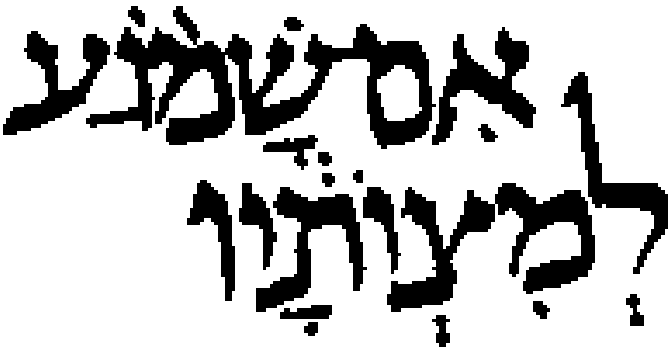
*Patah* should be shifted from its regular place to below the right side of the base character when this base character is one of *ayin*, *he* and *het*, and it is the last base character in a word or the next base character is *maqaf*. This Convention is suitable for Hebrew but may not be suitable for every other language written in Hebrew script.

## 2.3. *Holam* and *vav*

By Convention B3, the *vowel point holam* may appear in two positions relative to the base letter *vav*. The first position is above its top left or a little further to the left; the second position is above its top right or its top centre.[8] Thus a similar distinction is made to the one with *alef*. However, because *vav* is a narrow letter, the typographical distinction between these forms is often small. Sometimes, to make the distinction more clear, *holam* which would otherwise appear above the top left of *vav* is shifted further to the left, to over the space between the *vav* and the following base character. As with *holam* and *alef*, so similarly with *holam* and *vav*: *holam* above and to the left of *vav* is pronounced as a long O sound following the V or W sound of the *vav*; *holam* above the top right or centre of *vav* is pronounced as a long O sound but the *vav* is not pronounced.[9] This *holam* above the top right or centre of *vav* is not in fact logically associated with the *vav*, but is associated with the preceding base character. It is shifted from above and to the left of the preceding (to the right) letter to above the top right of the *vav* as a typographical convention, to form what is sometimes considered to be a compound character, known in Hebrew as *holam male*. This shift takes place only when the *vav* is silent, which is when there is no *vowel point* or *dagesh* combined with it and it is not followed by *vav shruqa* or *holam male*.

By Convention A3, no distinction is made between the two positions of *holam* above *vav*, so that *holam male* and consonantal *vav* with *holam* are graphically identical.

gādôl ʿăwōnî *"great is my iniquity"*, from Genesis 4:13 BHS
<*gimel*, *qamats*, *dagesh*, *dalet*, *holam*, *merkha*, *vav*, *lamed*, *space*, *ayin*, *hataf patah*, *vav*, *holam*, *nun*, *hiriq*, *tipeha*, *yod*>
*Holam* above the right of *vav*, i.e. *holam male* (right hand word), and *holam* above the left of *vav* (left hand word)[10]



The words ʾim-šāmôaʿ *"if hearing"* and ləmiṣwōtayw *"to his commandments"*,
from Exodus 15:26 in a facsimile of the Codex Leningradensis (dated 1008-9 CE)[11]
<*alef*, *hiriq*, *final mem*, *maqaf*, *shin*, *qamats*, *shin dot*, *mem*, *qadma*, *holam*, *vav*, *ayin*, *patah*>
<*lamed*, *sheva*, *mem*, *hiriq*, *tsadi*, *sheva*, *vav*, *holam*, *tav*, *qamats*, *zaqef qatan*, *yod*, *vav*>
*Holam* above the right of *vav*, i.e. *holam male* (upper word), and *holam* above the left of *vav* (lower word)
Also, furtive *patah* displaced to the right under the last letter of the upper word

baʿăwōnô *"in his iniquity"*, from Joshua 22:20 in the Aleppo Codex (dated c. 900 CE)
<*bet*, *patah*, *dagesh*, *ayin*, *hataf patah*, *vav*, *holam*, *nun*, *holam*, *meteg*, *vav*>
*Holam* above the left of *vav* (third letter from the end (left)) and *holam male* (last letter, with *holam* well to the right)

Exceptionally, when a collocation of *holam* and *vav* occurs within the divine name as written in the Bible text, i.e. in a word whose base characters are <*yod*, *he*, *vav*, *he*> (sometimes preceded by other base characters and/or followed by *maqaf* and another word), *holam* may appear above the top right of *vav* when there is another *vowel point* combined with and positioned below the same *vav*. See also section 3.1 below. Other exceptional collocations may occur in some texts.

The divine name, from Genesis 3:14 BHS
<*yod*, *sheva*, *he*, *holam*, *ZWJ*, *vav*, *qamats*, *qadma*, *he*>
*Holam* above the right of *vav* and *qamats* below it

Six options have been considered for encoding and rendering of collocations of *holam* and *vav*. The preferred option, which is described here, is based on the consensus reached in discussions on the Unicode Hebrew list. This option treats collocations of *holam* and *vav* in the same way as collocations of *holam* and *alef*. Arguably, it corresponds best to the historical development of Hebrew script and the linguistic properties of the Hebrew language. The other options are described in Appendix B, section B.1.

***Encoding guidelines***

The following guidelines should be used for texts which may be rendered according to either Convention A3 or Convention B3:

When a collocation of *vav* and *holam* is intended to be pronounced as the consonant sound *vav* followed by the vowel sound *holam*, and the *holam* is intended to be positioned above and to the left of *vav* when Convention B3 is used, the collocation should be encoded as <*vav*, *holam*>. When a collocation of *vav* and *holam* is *holam male*, and so intended to be pronounced as the vowel sound *holam* alone with the *vav* not pronounced, the collocation should be encoded as <*holam*, *vav*>, with the *holam* combined with the preceding base character. Where, exceptionally, a *holam* followed by a *vav* with no vowel should not be taken together as *holam male*, the encoding <*holam*, *ZWNJ*, *vav*> should be used. Where, exceptionally as for example in the divine name as shown above, a *holam* should appear above the top right of a following *vav* in a context where *holam male* would not normally be formed, the encoding <*holam*, *ZWJ*, *vav*> should be used. An isolated or word initial *holam male*, which might be used for example in a list of Hebrew characters or possibly in other languages written in Hebrew script, should be encoded as <*RLM*, *holam*, *vav*>.[12]

The following guidelines may be used for texts which are intended for rendering only according to Convention A3, and in practice have been used for existing modern Hebrew texts:

All collocations of *vav* and *holam* should be encoded as <*vav*, *holam*>.

A possible mechanism for requesting a distinctive rendering of consonantal *vav* with *holam* with rendering systems using Convention A3, to distinguish them from the commoner *holam male*, is to encode such collocations as <*vav*, *ZWNJ*, *holam*>.[13] This should be understood as a request for the *holam* not to be rendered above the *vav* at all but to be moved to the left and above it. This encoding is recommended for use only where there is no other way to make the required distinction.

***Rendering guidelines***

Convention A3

A general rendering system for Convention A3 should follow the same rendering guidelines as for Convention B3 below, except that the position of the shifted *holam* should be the same as for rendering of <*vav*, *holam*> sequences. If a rendering system is intended to render only texts in which all collocations of *vav* and *holam* are encoded as <*vav*, *holam*>, no special rendering is required. In any case the sequence <*vav*, *ZWNJ*, *holam*> (with accents possibly inserted) should be rendered as a *vav* with a *holam* to the left of it and above.

Convention B3

When a composite character sequence including *holam* is followed by *vav*, the *holam* should be shifted in rendering from its regular position and rendered above the top right of the *vav*. But this shift should not take place when the *vav* is combined with any *vowel point* or with *dagesh*. It should also not take place when the *vav* is followed immediately by *vav shruqa*[14] or *holam male*. It should never take place when the *vav* is preceded by *ZWNJ*, but always when the *vav* is preceded by *ZWJ*. The sequence <*RLM*, *holam*, *vav*> should be rendered as a *vav* with a *holam* above its top right. When a *holam* follows a *vav* within the same composite character sequence, the *holam* should be rendered in the regular way above and to the left of the *vav*. The sequence <*vav*, *ZWNJ*, *holam*> (with accents possibly inserted) should be rendered as a *vav* with a *holam* to the left of it and above, i.e. in many cases further to the left than the regular positioning of a *holam* above the top left of a *vav*.

## 2.4. *Holam* and *shin dot*

When a composite character sequence including *holam* is followed by a composite character sequence including *shin* and *shin dot*, by Convention A4 but not Convention B4 the *holam* dot is notionally shifted from above and to the left of the preceding (to the right) letter to the top right of the *shin* and merged with the *shin dot* which is in the same position; this may be in effect equivalent to deletion of the *holam*, although the merged dot is not necessarily rendered exactly like *shin dot*. This shift does not apply to any *holam* which should be positioned above the top right of *alef* or *vav* as

above.

***Encoding guidelines***

*Holam* should normally be encoded in such cases even though it may not be visibly rendered. If a content provider wishes to ensure that such a *holam* is never displayed with a particular text, regardless of the Convention, the *holam* may be omitted.

***Rendering guidelines***

Convention A4

*Holam* should not be rendered, i.e. simply omitted, above and to the left of any base character when followed by a composite character sequence including *shin* and *shin dot*. The rendering of *shin dot* may be adjusted e.g. replaced by the glyph for *holam*. Rendering of *holam* above the top right of any base character should not be adjusted.

Convention B4

The *holam* should be rendered in its regular position.

## 2.5. *Holam* and *sin dot*

When a composite character sequence includes *shin*, *sin dot* and *holam*, by Convention A5 but not Convention B5 the *holam* point is notionally merged with the *sin dot* which is in the same position; this is in effect equivalent to deletion of the *holam* although the merged dot is not necessarily rendered exactly like *sin dot*. (Probably Convention A5 goes with Convention A4 and Convention B5 with Convention B4).

***Encoding guidelines***

*Holam* should normally be encoded in such cases even though it may not be visibly rendered. If a content provider wishes to ensure that such a *holam* is never displayed with a particular text, regardless of the Convention, the *holam* may be omitted.

***Rendering guidelines***

Convention A5

*Holam* should not be rendered at all when part of a composite character sequence including *shin* and *sin dot*. The rendering of *sin dot* may be adjusted e.g. replaced by the glyph for *holam*. This guideline should take precedence over any merging with *shin dot*. Where *shin* with *sin dot* is followed by *holam male*, the *holam male* should be rendered fully i.e. including the *holam* dot; but where *shin* with *sin dot* is followed by *holam* and *alef*, merging of the *holam* with the *sin dot* should take precedence over the shift of the *holam* according to Convention B1.

Convention B5

The *holam* should be rendered adjacent to the *sin dot*, or shifted according to the normal rules when followed by *alef* or *vav*.

## 2.6. Punctuation issues

Certain Hebrew punctuation marks are not correctly described in Unicode 4.0.

*Sof pasuq* is used to indicate the end of a verse in the Hebrew Bible (although it is missing from the end of a few verses in some texts, and completely absent from some others) and as the equivalent of a full stop in other Hebrew writings

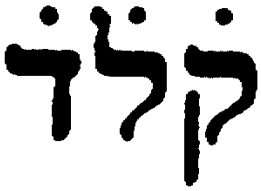
such as prayer books. It should be classed and processed as Terminal_Punctuation and as Sentence_Terminal.

*Paseq* is also used only at the ends of words, and so should also be classed as Terminal_Punctuation, but not as Sentence_Terminal. *Paseq* has two uses, one as part of the Hebrew accent system and the other as a special textual mark in the Hebrew Bible; it is normally found only in the Hebrew Bible and in quotations from it.

*Maqaf* is also generally considered to be a word divider and so should also be classed as Terminal_Punctuation. As its usage is analogous to that of *hyphen* and line breaks commonly occur after it in pointed Hebrew texts, it should also be listed in Unicode Standard Annex #14, along with *hyphen*, as a "break opportunity after".

## 2.7. Number dots

In post-biblical but pre-modern Hebrew texts, including the Masora (mediaeval editorial notes) which are printed with many Hebrew Bibles, numbers are indicated by Hebrew letters. Commonly but not always, one dot is placed above each of these letters to indicate this numerical use; sometimes two dots are used to indicate numbers greater than 1000.[15] The same dots are apparently used for other purposes in the Masora, such as indicating abbreviations. These dots do not normally co-occur with other Hebrew points.



The number 134, indicated by number dots, from the introduction (p. XV) to BHS
<*qof*, *dot above*, *lamed*, *dot above*, *dalet*, *dot above*>

There are no specific Unicode characters for these functions, but it is possible to use general purpose combining diacritical marks. (See section 3.6 below concerning the suitability for this purpose of the Unicode Hebrew mark *upper dot*.) And so it is proposed that the single number dot should be represented by *dot above* and the double number dot by *diaeresis*. Rendering systems need to be able to handle this use of general purpose diacritical marks with Hebrew base characters.

For an alternative involving use of *upper dot*, see Appendix B, section B.5.

## 2.8. Generic accent

Hebrew grammars, dictionaries and pedagogical materials commonly use a mark similar in shape to < (the *less-than* sign) positioned above a base character to indicate the stress accent in Hebrew words. This mark has a glyph and positioning similar to the accent *ole* used in the cantillated Bible text. The simplest suggestion here is to use the character *ole* also as the generic accent.

### *Encoding guidelines*

The generic accent should be encoded as *ole*.

### *Rendering guidelines*

Rendering systems intended for rendering of pointed Hebrew text should render *ole* appropriately even if they do not support the fully cantillated biblical text.

## *3. Issues relating to the Hebrew Bible text*

These issues are known to apply to the Hebrew Bible text. Some of them may also apply to other pre-modern texts in Hebrew and Aramaic. They are not normally relevant to modern works, except perhaps to quotations.

## 3.1. *Ketiv* and *Qere*

The Hebrew Bible text was originally written with consonants only, and with some vowels represented, but not fully specified, by *yod*, *vav* and *he*, perhaps also *alef*. This consonantal text came to be regarded as sacred and unchangeable. In parallel a tradition of reading aloud and chanting the text developed. After many centuries efforts were made to codify this pronunciation tradition by adding marks to the consonantal text, to indicate in detail vowels and chanting (cantillation) patterns, without actually changing the consonants. The result is the current Hebrew Bible text, consisting of consonants as base characters, and vowels and accents as combining marks (and a few as base characters e.g. *maqaf*, *paseq*).

In some cases there was a mismatch between the consonantal text and the pronunciation tradition, such that a word to be pronounced, known as *Qere*, required different consonants from the word written in the text, known as *Ketiv*. In such cases the consonants of the *Qere* were written in the margin of the text but the *vowel points* to be used to pronounce it were arranged around the unchangeable *Ketiv* consonants. The resulting form in the text was a hybrid, never intended to be pronounced as written. In some extreme cases it would in fact be unpronounceable, as for example in a few cases where there is a *Qere* word with no corresponding *Ketiv*, represented in the biblical text as blank space (or in some editions an asterisk or a similar glyph) surrounded by *vowel points* and accents.[16]

The *vowel points* of the *Qere* ʾēlay *"to me"*, from Ruth 3:17 BHS
<*RLM*, *NBSP*, *masora circle*, *NBSP*, *tsere*, *NBSP*, *patah*, *zaqef qatan*>
*Qere* without *Ketiv*

When a particular *Qere* form is used regularly with the same *Ketiv*, the *Qere* is not written in the margin but is assumed to be known, as a *perpetual Qere*. All that is visible is the sometimes anomalous combination of the *Ketiv* consonants and the *Qere* vowels. One well known example is the divine name, written with the *Ketiv* consonants *yod*, *he*, *vav*, *he* but with the *Qere* vowels *sheva*, *holam*, *qamats* or *sheva*, *holam*, *hiriq* taken from one of two alternative words used in pronunciation – although often the *holam* is not actually written. (The example of the divine name given above, from Genesis 3:14, is one of the exceptional cases in BHS where the *holam* is written.) For another important example of *perpetual Qere*, see section 3.3 below.

### *Encoding guidelines*

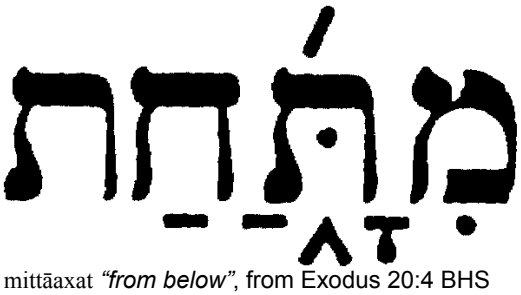
Various approaches may be taken to encoding each case of *Qere* and *Ketiv*. It would be possible to encode only the *Qere*, or only the *Ketiv*, or only the mixed form actually seen in the text. If the *Ketiv* is encoded, it may be left unpointed, or points may be added according to a reconstruction of its intended pronunciation. In cases of *perpetual Qere* the mixed forms are commonly encoded. Alternatively, any two or three of these options may be combined, with a higher level mark-up to indicate which form is which. The mixed forms in the text may be highly anomalous, and so the correct encoding for them may be unclear. One approach in cases like *Qere* without *Ketiv* is to encode *NBSP* in place of each missing base character followed by the associated diacritical marks; in this case the word should be preceded by *RLM* to ensure correct directionality as *NBSP* is directionally neutral. See also section 3.3 below.

### *Rendering guidelines*

Rendering systems should be aware that anomalous combinations are likely to occur, especially in mixed *Ketiv* and *Qere* forms. So they should not be unduly restrictive in failing to render properly sequences which are not normally accepted in Hebrew.

## 3.2. Double pointing and cantillation in the Ten Commandments

For historical reasons, in the two passages in the Hebrew Bible which list the Ten Commandments (Exodus 20:2-17 and Deuteronomy 5:6-21) two different pronunciation traditions are recorded in the pointing. The main consequence of this is that two different accents (cantillation marks) are attached to many of the words in these passages, and as the accent is generally positioned above or below the stressed syllable in the word, in most cases the two different accents are combined with the same base character. When two accents on the same base character would otherwise be in the same position, they are displaced sideways so that one is to the left of the other. Also both *rafe* and *dagesh* are combined with some base characters in these passages. For one unique word in Exodus 20:4 with two *vowel points*, see section 3.3 below.



mittāaxat *"from below"*, from Exodus 20:4 BHS
<*mem*, *hiriq*, *tav*, *dagesh*, *qamats*, *etnahta*, *CGJ*, *patah*, *geresh*, *het*, *patah*, *tav*>
Two *vowel points* and two accents, also *dagesh*, combined with one base character

***Encoding guidelines***

These passages should be encoded with two accents combined with a single base character where necessary. When the two accents would normally be in the same position, they have the same canonical combining class and so their relative ordering is significant; these accents should be encoded such that the one to the right precedes the one to the left. Co-occurrence of *rafe* and *dagesh* causes no problem as they have different canonical combining classes and do not interfere typographically.

***Rendering guidelines***

Rendering systems should expect to have to render two accents combined with a single base character, and if necessary to adjust their positioning relative to one another, and relative to *vowel points*, as required. A rather small set of combinations is actually attested in the biblical text, and so positioning algorithms can be tested exhaustively. Rendering systems should also expect to render *rafe* and *dagesh*, in either canonically equivalent order, combined with the same base character.

## 3.3. Multiple *vowel points*

The general rule in the Hebrew writing system is that no more than one *vowel point* is combined with any one base character (and indeed that, except at the end of a word or before *vav shruqa* or *holam male*, exactly one *vowel point* should be combined with each base character which is not silent).

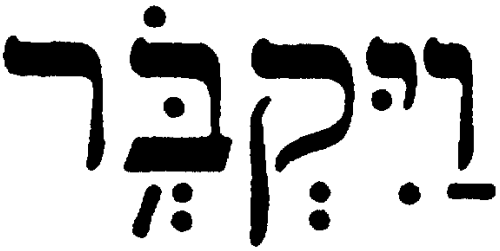
The few exceptions to this rule are in fact almost all cases where the consonants of a *Ketiv* word have been combined with the *vowel points* of a *Qere* word. But this is not immediately apparent in the one repeated example of *perpetual Qere* in which two *vowel points* appear with one consonant, because the mixed form with two *vowel points* is the only one which appears in standard electronic texts. This repeated example is the biblical form of the name of the city Jerusalem, written with the base characters <*yod*, *resh*, *vav*, *shin*, *lamed*, *final mem*>, with both *patah* (or sometimes *qamats*) and *hiriq* combined with the *lamed*, but pronounced, as written in modern Hebrew, as if ending <..., *lamed*, *patah* (or *qamats*), *yod*, *hiriq*, *final mem*>. This form is encoded in existing standard texts as it is printed, as <..., *lamed*, *patah* (or *qamats*), *hiriq*, *final mem*>, so with two *vowel points* combined with the *lamed*. There are also a few variant forms of this word with a directional *he* suffix and *sheva* replacing *hiriq*, so encoded <..., *lamed*, *patah* (or *qamats*), *sheva*, *mem*, *qamats*, *he*>.[17]

yərûšālaim, the name of Jerusalem, from Joshua 10:1 BHS
<*yod*, *sheva*, *resh*, *vav*, *dagesh*, *shin*, *qamats*, *shin dot*, *lamed*, *patah*, *revia*, *CGJ*, *hiriq*, *final mem*>
Sequence of *patah* and *hiriq*

yərûšālaəmāh, the name of Jerusalem with suffixed *he*, from 1 Kings 10:2 BHS
<*yod*, *sheva*, *resh*, *vav*, *dagesh*, *shin*, *qamats*, *shin dot*, *masora circle*, *lamed*, *patah*, *revia*, *CGJ*, *sheva*, *mem*, *qamats*, *he*>
Sequence of *patah* and *sheva*

Other mixed forms of *Ketiv* consonants and *Qere* *vowel points* have multiple *vowel points* with a single base character, as well as other more complex anomalies as described above in section 3.1 above.

A very few other anomalous cases are found in Hebrew Bible texts. For example, in the BHS Hebrew Bible and in the Leningrad Codex on which it is based, the fourth base character of the first word of 2 Kings 21:26 carries both *sheva* and *holam*. This was probably originally a scribal error as many manuscripts lack the *sheva*, but it has been perpetuated in a standard scholarly edition and in electronic texts based on it.



wayyiqəbəōr *"and he buried"*, from 2 Kings 21:26 BHS
<*vav*, *patah*, *yod*, *hiriq*, *dagesh*, *qof*, *sheva*, *bet*, *sheva*, *dagesh*, *merkha*, *CGJ*, *holam*, *resh*>
*Sheva* and *holam* with the same base character

In just one unique word in the Ten Commandments, the twelfth word (counting *maqaf* as a word divider) in Exodus 20:4, two different *vowel points*, *qamats* and *patah*, as well as *dagesh* and two accents are combined with a single base character *tav*. In this word only, the two *vowel points* represent alternative pronunciations and are not to be pronounced sequentially.[18]

Encoding of these cases presents a problem because these pairs of *vowel points* combined with a single base character are combining characters with different canonical combining classes, and so each pair is canonically equivalent to the same pair reversed. In fact, by chance, most of the attested pairs are in descending order of combining class and so are reversed by Unicode normalisation processes. It seems the possibility of multiple *vowel points* on a single base character was ignored by those who defined the combining classes for the *vowel points*. Unfortunately it is not possible to adjust the combining classes so that they meet the requirements of encoding such cases; these adjustments are prohibited by the Unicode Stability Policy because they would destabilise the normalisation of existing texts.

Several proposals have been put forward for working around this problem. Of these various proposals, the most promising is based on the character *CGJ*, which is a combining character with no visual representation. As its canonical combining class is zero it inhibits canonical reordering of combining characters between which it is encoded. The

alternative proposals are outlined in Appendix B, section B.2.

***Encoding guidelines***

The recommended encoding for any pair of *vowel points* combined with a single base character is for *CGJ* to be inserted between them, at least if the ordering of the pair is significant, either semantically or visually, and is different from the canonical ordering. Thus, for example, the city name Jerusalem should be encoded <..., *lamed*, *patah* (or *qamats*), *CGJ*, *hiriq*, *final mem*>; and the special case word in Exodus 20:4 should be encoded with a composite character sequence <*tav*, *dagesh*, *qamats*, *etnahta*, *CGJ*, *patah*, *geresh*> – hopefully the longest composite character sequence in the Hebrew Bible.

*Note:* In any Hebrew composite character sequence including *CGJ*, any of *dagesh*, *rafe*, *sin dot*, and *shin dot* should be encoded before the *CGJ*. This is for the purpose of collation, to ensure that after normalisation these characters are suitably positioned relative to the base character for collation contractions to be applied.

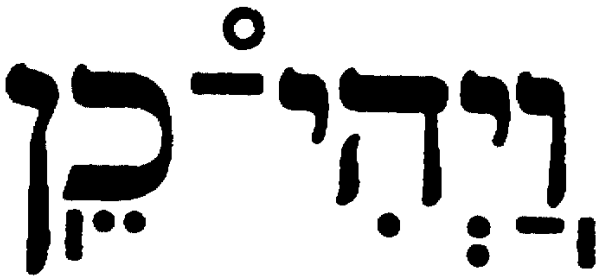
***Rendering guidelines***

Rendering systems should simply ignore *CGJ* when it appears between two *vowel points*, and render any pair of *vowel points*, if both are positioned below the base character, in right to left order below the base character. In the case of the city name Jerusalem, one practice, as in BHS, is to centre the *patah* or *qamats* under the *lamed* and shift the *hiriq* to below the right side of the *final mem*. In the special case word in Exodus 20:4, the three combining marks *qamats*, *etnahta*, *patah* should all appear in right to left order under one base character.

## 3.4. Right *meteg*

The point *meteg*, which is really part of the Hebrew accent system but is sometimes used e.g. to indicate secondary stress in texts which are not fully cantillated, is a short vertical line which is positioned below a base character and generally to the left of any *vowel point* below the base character. Note that the *vowel point* is not generally shifted from its centred position below the base character. See section 3.5 below for special rules with the *hataf vowels*.

In a minority of cases in some manuscripts and printed texts of the Hebrew Bible, *meteg* occurs not to the left but to the right of a low positioned *vowel point*. This is especially common at the beginning of a word. The distinction between these two positions seems to be not semantic but only graphical, and is not made in Convention A6. Nevertheless it is important to retain the distinction in scholarly texts, using Convention B6.[19]



wayəhî-kēn *"and it was so"*, from Genesis 1:7 BHS
<*vav*, *meteg*, *CGJ*, *patah*, *yod*, *sheva*, *he*, *hiriq*, *yod*, *maqaf*, *masora circle*, *kaf*, *tsere*, *meteg*, *final nun*>
Right *meteg* (on the right) and regular *meteg* (on the left)

wayəhî *"and it was"*, from Genesis 8:6 in a facsimile of the Codex Leningradensis[20]

<vav, meteg, CGJ, patah, yod, sheva, he,[21] hiriq, zaqef gadol, yod>
Right meteg

Encoding of meteg to the right of a vowel point presents a problem because the meteg and the vowel point are combining characters with different canonical combining classes, and so each pair is canonically equivalent to the same pair reversed. In fact the canonical combining class of meteg is higher than that of any vowel point, and so the canonical ordering corresponds to the more common graphical ordering with meteg on the left. Again, it is not possible to adjust the canonical combining classes.

There are three options for encoding meteg to the right of a vowel point. The preferred option is to use the CGJ character to specify the order of meteg and a vowel point. It is in accordance with the similar preference to use CGJ for encoding of multiple vowel points. The alternative proposals are outlined in Appendix B, section B.3.

***Encoding guidelines***

When a content provider wishes to make a positioning distinction, meteg to the right of a vowel point should be encoded <meteg, CGJ, vowel point>; and meteg to the left of a vowel point (other than a hataf vowel) should be encoded <vowel point, meteg>. When no distinction is intended, all collocations of meteg and vowel points should be encoded <vowel point, meteg>.

***Rendering guidelines***

Convention A6

When meteg and a vowel point positioned under the base character (other than a hataf vowel) are part of the same composite character sequence, the vowel point should be centred under the base character and the meteg glyph positioned to its left, and any CGJ should be ignored. When meteg occurs in a composite character sequence containing no vowel point or only holam, it should be centred below the base character.

Convention B6

When meteg and a vowel point positioned under the base character (other than a hataf vowel) are part of the same composite character sequence, the vowel point should be centred under the base character and the meteg glyph positioned to its left, or to its right if the meteg is encoded before the vowel point and is separated from it by CGJ. When meteg occurs in a composite character sequence containing no vowel point or only holam, it should be centred below the base character.

## 3.5. Meteg with hataf vowels

The three hataf vowels are made up graphically of sheva to the right of another vowel point, but this graphical equivalence should not be taken to have a semantic significance. In texts rendered according to Convention B7, on the rather rare occasions when meteg occurs with the same base character as a hataf vowel, the meteg glyph is most frequently positioned medially, i.e. between the two graphical elements of the hataf vowel, but it is on occasion positioned to the left or to the right of the hataf vowel.[22] When Convention A7 is used the relative positioning is always the same, with the meteg either always to the left or always medial. (Probably Convention A7 goes with Convention A6 and Convention B7 with Convention B6).

ʾăšer *"which"*, from Leviticus 21:10 BHS
<alef, hataf patah, meteg, shin, segol, shin dot, resh>

Medial *meteg* within *hataf patah*

There are two options for encoding collocations of *meteg* with *hataf vowels*. Again, the preferred option is based on the *CGJ* character.[23] The alternative is outlined in Appendix B, section B.4.

***Encoding guidelines***

When a content provider wishes to make a positioning distinction, *meteg* positioned medially within a *hataf vowel* should be encoded <*hataf vowel*, *meteg*>; *meteg* to the left of a *hataf vowel* should be encoded <*hataf vowel*, *CGJ*, *meteg*>; and *meteg* to the right of a *hataf vowel* should be encoded <*meteg*, *CGJ*, *hataf vowel*>. When no distinction is intended, all collocations of *meteg* and *hataf vowels* should be encoded <*hataf vowel*, *meteg*>.

***Rendering guidelines***

Convention A7

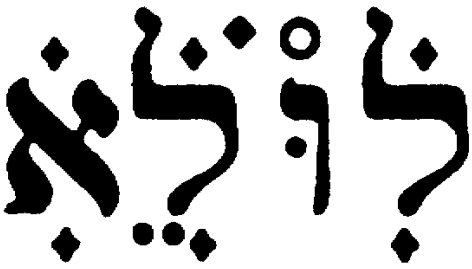
When *meteg* and a *hataf vowel* positioned under the base character are part of the same composite character sequence, the *hataf vowel* should be centred under the base character and the *meteg* glyph positioned either always to the right of it or always in the centre of it, as a typographical choice, and any *CGJ* should be ignored.

Convention B7

When *meteg* and a *hataf vowel* positioned under the base character are part of the same composite character sequence, the *hataf vowel* should be centred under the base character and the *meteg* glyph positioned in the centre of it, or to its right if the *meteg* is encoded before the *vowel point* and is separated from it by *CGJ*, or to its left if the *meteg* is encoded after the *vowel point* and is separated from it by *CGJ*.

## 3.6. Upper and lower extraordinary points

A small number of words in the Hebrew Bible are marked with a special or "extraordinary" point (Latin: *punctum extraordinarium*) above the word, or often but not always such a point above every letter of the word. Just one word, in Psalm 27:13, is marked by such a point both above each letter and below each letter. In BHS these points are diamond-shaped and larger than the round dots which make up many *vowel points*, but smaller than the accent *revia*; but typographical conventions for size and positioning vary. These points are an ancient part of the Hebrew Bible text, probably predating all other points, and are commonly printed in otherwise unpointed Bible texts.[24]



lûlē' *"unless"*, from Psalm 27:13 BHS
<*lamed*, *lower dot*, *upper dot*, *vav*, *dagesh*, *masora circle*, *lamed*, *tsere*, *lower dot*, *revia*, *upper dot*, *alef*, *lower dot*, *upper dot*>
Upper and lower extraordinary points (the largest dot at the top is *revia*)

The Unicode Hebrew mark *upper dot* is apparently intended for use as the upper extraordinary point. It may also have been intended originally for use as the number dot, but the differences in shape and positioning make it inappropriate to use the one character for both purposes, and general purpose combining diacritical marks are more suitable for the number dots. So *upper dot* should be used for the upper extraordinary point. As there is no corresponding character already defined for the lower extraordinary point, it will be necessary to define in Unicode one new combining character, HEBREW MARK LOWER DOT. As this should be centred below the base character, it should have a canonical combining class of 220.

Note the typographical convention in the example illustrated, that the *upper dot* displaces *revia* to the right (as they have the same canonical combining class, this implies that the encoding should be <*revia*, *upper dot*>), but the *lower dot* displaces *tsere* to the left; there are other conventions in other editions.

For an alternative involving use of the general purpose combining diacritical marks, see Appendix B, section B.5.

### 3.7. *Setuma* and *petuha*

Two symbols are used frequently in the Hebrew Bible to indicate paragraph breaks (independently of chapter numbering which was introduced in about the 16th century CE). The symbol *setuma* is identical to the Hebrew letter *samekh*, and the symbol *petuha* is identical to the letter *pe* (the non-final form), except that in some texts they are printed in a smaller size. Semantically, these are punctuation marks. Even so, it is probably adequate to represent them in Unicode as *samekh* and *pe* respectively, with higher level formatting if required.

### 3.8. *Inverted nun*

In a few places in the Hebrew Bible a symbol known as *inverted nun* is found. This is a punctuation mark, not a variant form of the letter *nun*. As its resemblance to *nun* is only approximate and has no semantic significance, it is best considered a completely separate character. It will therefore be necessary to propose a new Unicode base character HEBREW PUNCTUATION INVERTED NUN.[25]

*Inverted nun*, from Numbers 10:34 BHS[26]

### 3.9. Unusual letter forms

In a number of places in the Hebrew Bible unusual letter forms are traditionally used, though not all are found in all printed editions. These include enlarged letters, reduced letters and raised letters.[27] These are semantically variant forms of the letters, and so they do not require specific representation in Unicode. These variations may be indicated by higher level formatting. There are also a few "broken" letters with variant glyphs, for which it may be necessary to add new characters to Unicode, or perhaps to define a use of variation selectors to indicate that a variant glyph is required.[28]

mənaššeh *"Manasseh"*, from Judges 18:30 BHS
<*mem*, *sheva*, <raised> *nun*, *patah* <\raised>, *shin*, *segol*, *dagesh*, *shin dot*, *geresh*, *masora circle*, *he*>
Raised *nun*, and a six character composite character sequence in normal text

## *4. Issues relating to non-Tiberian pointing*

Several different pointing systems have been used for Hebrew and Aramaic texts. Unicode currently supports only one of these systems, the Tiberian system, which is in current use for printed Bibles and most liturgical texts, and is used for modern Hebrew when it is written with *vowel points*. Three other pointing systems, known as Babylonian, Palestinian and Samaritan, are important for scholarly purposes, and at least the Babylonian system is still in liturgical use in some Jewish communities. There is no simple one-to-one correspondence between the systems. Further study is needed of how these other systems might be represented in Unicode. Probably several new Unicode characters will need to be

proposed.

## 5. Issues relating to languages less commonly written in Hebrew script

Unicode 4.0 lacks support for a few combining marks found in some variants of the Hebrew script used for less well known languages, or by specific communities for languages more usually written with other scripts. A few such marks have already been identified, and more may be identified as study continues.

Hebrew script may also be used, in dictionaries and pedagogical materials, to represent for modern Hebrew speakers the pronunciation of languages, e.g. English, not otherwise written in Hebrew script. For such purposes anomalous character sequences may be used. Further study may lead to the identification of marks used for this purpose which are not normally used in Hebrew.

Many of the marks which may be needed for these purposes, if not already defined as Hebrew combining characters, are defined as combining characters in other Unicode blocks. If additional marks are required which cannot be represented in this way, it will be necessary to propose these marks to Unicode as new characters.

### *Rendering guidelines*

Rendering systems should be aware that anomalous combinations are likely to occur when Hebrew script is used for languages other than Hebrew. These may include combinations of Hebrew base characters with combining marks from other Unicode blocks. So rendering systems should not be unduly restrictive in failing to render properly sequences which are not normally accepted in Hebrew.

## 6. Summary of recommendations

In order to complete support for Hebrew with Tiberian pointing in Unicode, it would appear that the following two new characters should be added to Unicode:

- HEBREW MARK LOWER DOT
- HEBREW PUNCTUATION INVERTED NUN

Other ambiguities in the encoding and rendering of Hebrew in Unicode can be resolved by following the guidelines set out in this paper. A few minor features of the Hebrew Bible text can best be resolved by higher level formatting.

Additional new characters will be required for support of non-Tiberian pointing systems and perhaps for other languages written in Hebrew script.

## *APPENDIX A: Abbreviated names of Unicode characters and sequences*

### Existing Unicode characters:

| *Abbreviation* | *Unicode code point* | *Unicode character name or explanation* |
|---|---|---|
| *alef* | U+05D0 | HEBREW LETTER ALEF |
| *ayin* | U+05E2 | HEBREW LETTER AYIN |
| *bet* | U+05D1 | HEBREW LETTER BET |
| *CGJ* | U+034F | COMBINING GRAPHEME JOINER |
| *dagesh*[29] | U+05BC | HEBREW POINT DAGESH OR MAPIQ |
| *dalet* | U+05D3 | HEBREW LETTER DALET |
| *diaeresis* | U+0308 | COMBINING DIAERESIS |
| *dot above* | U+0307 | COMBINING DOT ABOVE |
| *dot below* | U+0323 | COMBINING DOT BELOW |
| *etnahta* | U+0591 | HEBREW ACCENT ETNAHTA |
| *final mem* | U+05DD | HEBREW LETTER FINAL MEM |
| *final nun* | U+05DF | HEBREW LETTER FINAL NUN |
| *geresh* | U+059C | HEBREW ACCENT GERESH |
| *gimel* | U+05D2 | HEBREW LETTER GIMEL |
| *hataf patah* | U+05B2 | HEBREW POINT HATAF PATAH |
| *hataf qamats* | U+05B3 | HEBREW POINT HATAF QAMATS |
| *hataf segol* | U+05B1 | HEBREW POINT HATAF SEGOL |
| *he* | U+05D4 | HEBREW LETTER HE |
| *het* | U+05D7 | HEBREW LETTER HET |
| *hiriq* | U+05B4 | HEBREW POINT HIRIQ |
| *holam* | U+05B9 | HEBREW POINT HOLAM |

| *Abbreviation* | *Unicode code point* | *Unicode character name or explanation* |
|---|---|---|
| *hyphen* | U+2010 | HYPHEN |
| *kaf* | U+05DB | HEBREW LETTER KAF |
| *lamed* | U+05DC | HEBREW LETTER LAMED |
| *less-than* | U+003C | LESS-THAN SIGN |
| *maqaf* | U+05BE | HEBREW PUNCTUATION MAQAF |
| *masora circle* | U+05AF | HEBREW MARK MASORA CIRCLE |
| *mem* | U+05DE | HEBREW LETTER MEM |
| *merkha* | U+05A5 | HEBREW ACCENT MERKHA |
| *meteg* | U+05BD | HEBREW POINT METEG |
| *NBSP* | U+00A0 | NO-BREAK SPACE |
| *nun* | U+05E0 | HEBREW LETTER NUN |
| *ole* | U+05AB | HEBREW ACCENT OLE |
| *paseq* | U+05C0 | HEBREW PUNCTUATION PASEQ |
| *patah* | U+05B7 | HEBREW POINT PATAH |
| *pe* | U+05E4 | HEBREW LETTER PE |
| *qadma* | U+05A8 | HEBREW ACCENT QADMA |
| *qamats* | U+05B8 | HEBREW POINT QAMATS |
| *qof* | U+05E7 | HEBREW LETTER QOF |
| *qubuts* | U+05BB | HEBREW POINT QUBUTS |
| *rafe* | U+05BF | HEBREW POINT RAFE |
| *resh* | U+05E8 | HEBREW LETTER RESH |
| *revia* | U+0597 | HEBREW ACCENT REVIA |

| *Abbreviation* | *Unicode code point* | *Unicode character name or explanation* |
|---|---|---|
| *RLM* | U+200F | RIGHT-TO-LEFT MARK |
| *samekh* | U+05E1 | HEBREW LETTER SAMEKH |
| *segol* | U+05B6 | HEBREW POINT SEGOL |
| *sheva* | U+05B0 | HEBREW POINT SHEVA |
| *shin* | U+05E9 | HEBREW LETTER SHIN |
| *shin dot* | U+05C1 | HEBREW POINT SHIN DOT |
| *sin dot* | U+05C2 | HEBREW POINT SIN DOT |
| *sof pasuq* | U+05C3 | HEBREW PUNCTUATION SOF PASUQ |
| *space* | U+0020 | SPACE |
| *tav* | U+05EA | HEBREW LETTER TAV |
| *telisha gedola* | U+05A0 | HEBREW ACCENT TELISHA GEDOLA |
| *tipeha* | U+0596 | HEBREW ACCENT TIPEHA |
| *tsadi* | U+05E6 | HEBREW LETTER TSADI |
| *tsere* | U+05B5 | HEBREW POINT TSERE |
| *upper dot* | U+05C4 | HEBREW MARK UPPER DOT |
| *vav* | U+05D5 | HEBREW LETTER VAV |
| *vav with holam* | U+FB4B | HEBREW LETTER VAV WITH HOLAM |
| *yod* | U+05D9 | HEBREW LETTER YOD |
| *zaqef gadol* | U+0595 | HEBREW ACCENT ZAQEF GADOL |
| *zaqef qatan* | U+0594 | HEBREW ACCENT ZAQEF QATAN |
| *zayin* | U+05D6 | HEBREW LETTER ZAYIN |
| *ZWJ* | U+200D | ZERO WIDTH JOINER |

| *Abbreviation* | *Unicode code point* | *Unicode character name or explanation* |
|---|---|---|
| *ZWNJ* | U+200C | ZERO WIDTH NON-JOINER |

### Sequences and groups of Unicode characters:

| *Abbreviation* | *Unicode code point* | *Unicode character name or explanation* |
|---|---|---|
| *hataf vowel* | Any of U+05B1, U+05B2 and U+05B3 | Any of *hataf segol*, *hataf patah*, *hataf qamats* |
| *holam male* | (U+05B9 U+05D5) | A collocation of *holam* and *vav* when pronounced as a vowel |
| *vav shruqa*[30] | U+05D5 U+05BC | A composite character sequence including *vav* and *dagesh* with no *vowel points* (but possibly including other combining characters), when pronounced as a vowel |
| *vowel point* | Any of U+05B0 to U+05B9 and U+05BB | Any of *sheva*, *hataf segol*, *hataf patah*, *hataf qamats*, *hiriq*, *tsere*, *segol*, *patah*, *qamats*, *holam*, *qubuts* |

### Proposed new characters:

| *Abbreviation* | *Unicode character name or explanation* |
|---|---|
| *inverted nun* | Proposed HEBREW PUNCTUATION INVERTED NUN |
| *lower dot* | Proposed HEBREW MARK LOWER DOT |

### Possible proposed new characters according to alternative proposals:

| *Abbreviation* | *Unicode character name or explanation* |
|---|---|
| *double upper dot* | (Alternative proposed HEBREW MARK DOUBLE UPPER DOT) |
| *hataf patah meteg* | (Alternative proposed HEBREW POINT HATAF PATAH WITH METEG) |
| *hataf qamats meteg* | (Alternative proposed HEBREW POINT HATAF QAMATS WITH METEG) |
| *hataf segol meteg* | (Alternative proposed HEBREW POINT HATAF SEGOL WITH METEG) |
| *holam male* | (Alternative proposed HEBREW LETTER HOLAM MALE) |
| *right holam* | (Alternative proposed HEBREW POINT RIGHT HOLAM) |
| *right meteg* | (Alternative proposed HEBREW POINT RIGHT METEG) |

## *APPENDIX B: Alternative proposals and guidelines*

The following alternative proposals and corresponding guidelines have been considered but, in the light of discussions on the Unicode lists, provisionally rejected.

## B.1. Alternatives for *holam* and *vav*

### *B.1.1. Alternative 1*

This alternative is based on current practice, mostly for modern Hebrew, in which the two positions of *holam* on *vav* are generally not distinguished for simplicity of rendering. The collocation of *vav* and *holam* which is intended as *holam male* is encoded not in its logical order in a similar way to furtive *patah*.

#### *Encoding guidelines*

All collocations of *vav* and *holam* should be encoded as <*vav*, *holam*>. For exceptional cases (in this case not including the divine name example given), *ZWJ* may be inserted before the *vav* to indicate that the collocation should be understood as *holam male*, i.e. as a vowel associated with what precedes it, and *ZWNJ* may be inserted to indicate that it should be understood as the consonant *vav* with *holam*, i.e. as a separate syllable.

#### *Rendering guidelines*

Convention A3

All composite character sequences of *vav* and *holam* should be rendered in the same way.

Convention B3

In a composite character sequence including *vav* and *holam*, *holam* should always be rendered above and to the left of *vav* when the sequence also includes *dagesh*. It should also be rendered above and to the left of *vav* when it is word initial, and when it is preceded by *maqaf* or by any composite character sequence including one or more *vowel point*. Otherwise the *holam* should be rendered above the top right of the *vav*. However, when the sequence is preceded by *ZWJ* *holam* should always be rendered above the top right of *vav*, and when it is preceded by *ZWNJ* *holam* should always be rendered above and to the left of *vav*.

*Note: This procedure depends on a simplified assumption that a composite character sequence including vav and dagesh preceding a combination of vav and holam is always to be understood as the consonant vav with dagesh, and never as the long vowel vav shruqa which is encoded identically. Consonantal vav with dagesh followed by a collocation of vav and holam is found in Hebrew texts. Vav shruqa followed by such a collocation does not appear to be found in the Hebrew Bible. As holam should appear above and to the left of vav when following vav shruqa, an exhaustive solution here requires detection of the difference between vav shruqa and consonantal vav with dagesh, which is a similar procedure to the above which must in principle be applied recursively back to the start of a sequence of composite character sequences including the characters vav and dagesh.*

*As an alternative to this recursive algorithm, a rendering engine may keep a record from the beginning of the word of whether the last character was a consonant or a vowel, and use this to disambiguate collocations of holam and vav. Note that in Hebrew a word initial vav and dagesh combination is in fact generally, possibly always, the vowel vav shruqa, the only permitted word initial vowel. This algorithm is not necessarily appropriate for other languages written in Hebrew script.*[31]

### *B.1.2. Alternative 2*

This alternative is a combination of the preferred guidelines and alternative 1, intended to provide compatibility with existing texts which may be encoded according to either of these options.

***Encoding guidelines***

Collocations of *vav* and *holam* may be encoded either according to the preferred guidelines or according to alternative 1. One or other of these options could be specified as the preferred encoding. A folding rule should be introduced such that both encodings are considered equivalent for collation purposes.

***Rendering guidelines***

The preferred rendering guidelines should first be applied. Any remaining composite character sequence including *vav* and *holam* in which the *holam* has not been shifted should then be processed according to the rendering guidelines of alternative 1.

### *B.1.3. Alternative 3*

This alternative requires one new base character to be added to Unicode, HEBREW LETTER HOLAM MALE, with a glyph consisting of a *vav* with a *holam* point above its top right. This character should not be considered a pre-composed form, and it should not have any canonical decomposition. In these respects it is distinct from the alphabetic presentation form *vav with holam*, of which general use is discouraged. It should have a compatibility decomposition to <*holam*, *vav*> according to its pronunciation.

***Encoding guidelines***

When a collocation of *vav* and *holam* is intended to be pronounced as the consonant sound *vav* followed by the vowel sound *holam*, and the *holam* is intended to be positioned above and to the left of *vav*, the collocation should be encoded as <*vav*, *holam*>. When a collocation of *vav* and *holam* is *holam male*, and so intended to be pronounced as the vowel sound *holam* alone with the *vav* not pronounced, the collocation should be encoded with the new character *holam male*.

***Rendering guidelines***

Convention A3

Composite character sequences including *vav* and *holam* and new character *holam male* should be rendered identically, as a *vav* glyph with a *holam* dot above it.

Convention B3

In a composite character sequence including *vav* and *holam*, *holam* should always be rendered above and to the left of *vav*. The rendering of the new character *holam male* is straightforward and should be distinct, with the *holam* dot glyph above the top right of the *vav* glyph.

### *B.1.4. Alternative 4*

This alternative requires one new combining character to be added to Unicode, HEBREW POINT RIGHT HOLAM, with a glyph identical to that of *holam* but positioned above the top right of the base character. It should have a compatibility decomposition to 05B9 *holam*. Note that if this option is chosen it might be sensible to use this character also for *holam* above the right of *alef*.

***Encoding guidelines***

When a collocation of *vav* and *holam* is intended to be pronounced as the consonant sound *vav* followed by the vowel sound *holam*, and the *holam* is intended to be positioned above and to the left of *vav*, the collocation should be encoded as <*vav*, *holam*>. When a collocation of *vav* and *holam* is *holam male*, and so intended to be pronounced as the vowel sound *holam* alone with the *vav* not pronounced, the collocation should be encoded as <*vav*, *right holam*>.

***Rendering guidelines***

Convention A3

Composite character sequences including *vav* and *holam* and those including *vav* and *right holam* should be rendered identically, as a *vav* glyph with a *holam* dot above it.

Convention B3

In a composite character sequence including *vav* and *holam*, *holam* should always be rendered above and to the left of *vav*. In a composite character sequence including *vav* and *right holam*, *right holam* should always be rendered above the top right of *vav*.

#### *B.1.5. Alternative 5*

This alternative treats collocations of *vav* and *holam* as generally representing *holam male*, and the relatively rare case of consonantal *vav* with *holam* as a special case.

***Encoding guidelines***

When a collocation of *vav* and *holam* is intended to be pronounced as the consonant sound *vav* followed by the vowel sound *holam*, the collocation should be encoded as <*vav*, *LCC*, *holam*>, where *LCC* is a layout control character. One current text uses *ZWJ* as this layout control character, but it might be theoretically better to use *ZWNJ* or *CGJ*: *ZWNJ* because the function is more of separation than of joining; and *CGJ* because this avoids use of a defective combining sequence. When a collocation of *vav* and *holam* is *holam male*, and so intended to be pronounced as the vowel sound *holam* alone with the *vav* not pronounced, the collocation should be encoded simply as <*vav*, *holam*>.

***Rendering guidelines***

Convention A3

Composite character sequences including *vav* and *holam* and sequences with an additional layout control character should be rendered identically, as a *vav* glyph with a *holam* dot above it.

Convention B3

In a composite character sequence including *vav* and *holam*, *holam* should be rendered above the top right of *vav*. In a composite character sequence including the chosen layout control character and *holam*, *holam* should always be rendered above and to the left of the preceding base character.

## B.2. Alternatives for multiple *vowel points*

The only technically acceptable alternative solution to the problem of multiple *vowel points* which has been put forward involves adding to the Hebrew block of Unicode a complete second set of *vowel points* with the correct canonical combining classes. Three variants of this have been suggested: the new set of *vowel points* should be used only for biblical Hebrew; they should be used in place of the existing *vowel points* for all Hebrew text, with the currently defined *vowel points* being deprecated; or they should be used only for the second (and in principle any subsequent) *vowel point* in any composite character sequence. The proposal to use different *vowel points* for biblical and modern Hebrew is fundamentally unacceptable to many Hebrew users. The proposal to change the *vowel points* in all existing texts has the serious drawback that it invalidates all existing pointed Hebrew texts. While there are no serious problems with using a second set of *vowel points* in only these rather few anomalous cases, it does seem to be an inefficient solution to define so many new characters which will be used so rarely.

## B.3. Alternatives for right *meteg*

#### *B.3.1. Alternative 1*

The first alternative for right *meteg* is to add a new combining character to Unicode, HEBREW POINT RIGHT METEG, with a canonical combining class different from that of any *vowel point*.[32] This new *right meteg* character should have a compatibility equivalence to *meteg* and should be treated as equivalent in searching and collation.

***Encoding guidelines***

*Meteg* to the right of a *vowel point* should be encoded as <*right meteg*, *vowel point*>. *Meteg* to the left of a *vowel point* should be encoded as <*vowel point*, *meteg*>.

***Rendering guidelines***

Convention A6

When either *meteg* or *right meteg* and a *vowel point* positioned under the base character are part of the same composite character sequence (in either of the canonically equivalent orders), the *vowel point* should be centred under the base character and the *meteg* glyph positioned to its left. When either *meteg* or *right meteg* occur in a composite character sequence containing no *vowel point* or only *holam*, the *meteg* glyph should be centred below the base character.

Convention B6

When *right meteg* and a *vowel point* positioned under the base character are part of the same composite character sequence (in either of the canonically equivalent orders), the *vowel point* should be centred under the base character and the *meteg* glyph positioned to its right. When *meteg* and a *vowel point* positioned under the base character are part of the same composite character sequence (in either of the canonically equivalent orders), the *vowel point* should be centred under the base character and the *meteg* glyph positioned to its left. When either *meteg* or *right meteg* occur in a composite character sequence containing no *vowel point* or only *holam*, the *meteg* glyph should be centred below the base character.

### *B.3.2. Alternative 2*

This second alternative depends on acceptance of one of the proposals to fix the multiple *vowel points* problem by defining a complete additional set of *vowel points*. If one of these is agreed despite its serious drawbacks, it is sensible also to define an additional *meteg* character with an appropriate canonical combining class for use with these *vowel points*.

## B.4. Alternative for *meteg* with *hataf vowels*

The alternative for *meteg* with *hataf vowels* is to add three new combining characters to Unicode, HEBREW POINT HATAF SEGOL WITH METEG, HEBREW POINT HATAF PATAH WITH METEG and HEBREW POINT HATAF QAMATS WITH METEG[33], with canonical combining classes probably the same as those of the corresponding *hataf vowels* without *meteg*. The glyphs should be those of the *hataf vowels* with the *meteg* glyph positioned between the two halves. Each of these new characters should have a compatibility decomposition to the corresponding *hataf vowel* followed by *meteg* and should be treated as equivalent to such a sequence in searching and collation.

***Encoding guidelines***

Combinations of *hataf vowels* with medial *meteg* should be encoded with these newly defined characters. Other collocations of *meteg* with *hataf vowels* should be encoded as for collocations of *meteg* with other *vowel points*.

***Rendering guidelines***

Convention A7

These newly defined characters should be rendered with their combined glyphs centred below the base character. Other collocations of *meteg* with *hataf vowels* should also be rendered with these combined glyphs.

Convention B7

These newly defined characters should be rendered with their combined glyphs centred below the base character. Other collocations of *meteg* with *hataf vowels* should be rendered as for collocations of *meteg* with other *vowel points*.

## B.5. Alternative for number dots and extraordinary points

The alternative for number dots and upper and lower extraordinary points is to use *upper dot* and a new character for the former and general purpose combining diacritical marks for the latter. This is not preferred because the extraordinary points seem to have distinctive shapes and positioning, but the number dots do not.

With this alternative, *upper dot* is used for the single number dot. There is no corresponding double dot in Unicode for the thousands marker. It has been suggested that a sequence of two *upper dots* should be used for this. A better long term solution would be to define a new Unicode combining character HEBREW MARK DOUBLE UPPER DOT. The canonical combining class of this *double upper dot* should be 230, the same as that of *upper dot*.

With this alternative, the general purpose combining diacritical marks *dot above* and *dot below* are used for the upper and lower extraordinary points respectively. Rendering systems need to be able to handle this use of general purpose diacritical marks with Hebrew base characters, in this case in combination with Hebrew *vowel points* and accents, and to use suitable distinctive glyphs.

1 The Unicode Hebrew block is also suitable for representation of Aramaic as used in inscriptions and papyri, especially those from c. 500 BCE. The Hebrew "square" script, on which the Unicode Hebrew block is based, was copied from the Aramaic script in use at that time, and the basic glyphs have changed very little since then. See http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2311.pdf for more details. However, a proposal has been made to the Unicode Technical Committee for a separate "Early Aramaic" block (http://std.dkuug.dk/JTC1/SC2/WG2/docs/n2042.pdf). Although most of these letter shapes are only slight variants of the corresponding Hebrew letters, this proposed new block has been included in the Unicode Roadmap (http://www.unicode.org/roadmaps/bmp/).

2 *Biblia Hebraica Stuttgartensia*, 4th edition, German Bible Society 1990.

3 These examples are scanned from BHS. For scanned examples from other printed Bibles, see for example Haralambous, *Typesetting the Holy Bible in Hebrew, with TEX*, 1994, http://genepi.louis-jean.com/omega/biblical-hebrew94.pdf or http://omega.enstb.org/yannis/pdf/biblical-hebrew94.pdf, also http://www.moses.uklinux.net/sample/sample.pdf and http://www.mfa.gov.il/mfa/go.asp?MFAH0kxu0.

4 This is how the words displayed would be encoded according to the guidelines in this paper, with Unicode canonical ordering. This ordering has been chosen as an illustration of the problems with it.

5 In this sequence the *holam* technically forms a defective combining sequence. *RLM* has two functions: to prevent the *holam* combining with a previous character, e.g. a *space*; and to ensure that the *holam* (which is directionally neutral as a non-spacing mark) is rendered with right-to-left directionality.

6 To avoid excessive complexity in the algorithm, the shift should not take place in any sequence <*holam*, *alef*, *vav*, *dagesh*>. If exceptionally the *vav* here is consonantal and so the *alef* is silent, *ZWJ* should be inserted before the *alef* to ensure the correct rendering.

7 Furtive *patah* is found about 2683 times in the Hebrew Bible (in the Michigan-Claremont-Westminster electronic edition of the BHS text).

8 Haralambous, p.8, notes that the *holam* in the second position may also be positioned lower than a regular *holam*. This lower position is clearly visible in the extract from a Viennese Bible of 1889 reproduced by Haralambous, p.18.

9 *Holam* above and to the left of *vav* is found about 422 times in the Hebrew Bible. *Holam* above the top right of *vav* (*holam male*) is very much more frequent.

10 The same distinction in *holam* positioning is made in the Israeli Bible editions BHL = Aron Dotan (ed.), *Biblia Hebraica Leningradensia* (Hendrickson, 2001) and *Keter Yerushalayim* (Ben Zvi, 2001). In BHL *holam* tends to be positioned above the

gaps between base letters.

11 Taken from http://www.usc.edu/dept/LAS/wsrp/educational_site/biblical_manuscripts/LeningradCodex2_e.jpg. These two words appear in the manuscript one above the other as shown. The image has been tidied by deletion of parts of other words which surround these two.

12 The note concerning the sequence <*RLM*, *holam*, *alef*> applies also to this sequence.

13 Again, the *holam* here forms a defective combining sequence, and for this reason this encoding cannot be recommended except in special circumstances.

14 To avoid excessive complexity in the algorithm, the shift should not take place in any sequence of <*holam*, *vav*, *vav*, *dagesh*>. If exceptionally the second *vav* here is consonantal and so the *holam* and the first *vav* make up *holam male*, *ZWJ* should be inserted before the first *vav* to ensure the correct rendering.

15 A slightly different system for distinguishing such numbers from words, already well supported by Unicode, is described at http://www.qsm.co.il/Hebrew/Gimatria.htm.

16 *Qere* without *Ketiv* is found in ten places in the Hebrew Bible: Judges 20:13, 2 Samuel 8:3, 16:23, 18:20, 2 Kings 19:31,37, Jeremiah 31:38, 50:21, Ruth 3:5,17. Other examples of anomalous mixed forms of *Ketiv* and *Qere* include the examples of "Letters with more than one vowel" and "isolated *dagesh*" and the 30 cases of "Missing letters" listed by Haralambous, pp.10-11.

17 The city name is found 636 times in the Hebrew Bible, including four occurrences with the suffixed *he*.

18 This author is not aware of any other genuine cases of multiple vowels which are not related to *Qere* and *Ketiv* issues. In 2 Chronicles 13:14 in the electronic BHS text there is an anomalous <*sheva*, *hiriq*> sequence in a reconstructed *Ketiv* form, but the editors have agreed that this is an error. In some electronic texts a <*hiriq*, *holam*> sequence is found in Ezekiel 12:23, but this is based on a misreading of the text.

19 Right *meteg* is found about 906 times in the electronic BHS, whereas left *meteg* occurs more than 30,000 times. The usage and positioning of *meteg* varies considerably between editions of the Hebrew Bible.

20 Taken from http://www.moses.uklinux.net/sample/b19a-preview.pdf.

21 Although this looks like a *het*, the context makes it clear that a *he* is intended.

22 In the electronic BHS, *meteg* occurs 78 times medially in a *hataf vowel*, six times to the left and twice to the right. *Meteg* never occurs in any position with *hataf qamats*. Other editions vary considerably.

23 Technically this use of *CGJ* is not fully in accordance with the Unicode standard as *CGJ* is having an effect on the formation of a ligature between the *meteg* and the *hataf vowel*. But the characters *ZWJ* and *ZWNJ* which are defined for ligation control are unsuitable in this context as they break the composite character sequence. It seems most sensible and consistent to use *CGJ* for this function as well.

24 Upper extraordinary points are found in Genesis 16:5, 18:9, 19:33, 33:4, 37:12, Numbers 3:39, 9:10, 21:30, 29:15, Deuteronomy 29:28, 2 Samuel 19:20, Isaiah 44:9, Ezekiel 41:20, 46:22, and Psalm 27:13. Lower extraordinary points are found only in Psalm 27:13. See also http://www.bayit02.freeserve.co.uk/html/dots.html.

25 *Inverted nun* is found in the Hebrew Bible at the end of Numbers 10:34 and 10:36 and before each of Psalm 107:21,22,23,24,25,26,40. In some manuscripts, however, *inverted nuns* are used in place of regular *nuns*: see for example the extract at http://www.bayit02.freeserve.co.uk/html/reversed_nuns.html in which, in place of the *inverted nuns* at Numbers 10:34 and 10:36, the first regular *nun* after each of them is replaced by an *inverted nun*.

26 Haralambous, p.10, and others consider unsatisfactory this glyph used in BHS. A more correct glyph is similar to a horizontal reversal of the glyph for *nun*. See also http://www.bayit02.freeserve.co.uk/html/nun_hafucha.html for some sample glyphs.

27 Haralambous, pp.8-10, and Gesenius, Kautzsch and Cowley, *Hebrew Grammar*, Oxford UP 1910, p.31 section 5n, list 27 enlarged letters and 19 reduced letters; raised letters at Judges 18:30, Psalm 80:14 and Job 38:13,15. The irregular uses of *final mem* in Isaiah 9:6 and non-final *nun* in Job 38:1, and in some editions of non-final *mem* in Nehemiah 2:13, are not a problem for Unicode because final and non-final forms are encoded separately. See also http://www.bayit02.freeserve.co.uk/html/large_letters.html and http://www.bayit02.freeserve.co.uk/html/small_letters.html.

28 Haralambous and Gesenius, Kautzsch and Cowley list a "broken" *vav* at Numbers 25:12 and variant forms of *qof* at Exodus 32:25 and Numbers 25:12. See also http://www.bayit02.freeserve.co.uk/html/broken_vav.html.

29 The abbreviation *dagesh* is used even in cases where this point is technically known as *mapiq* or *shuruq* rather than *dagesh*.

30 *Vav shruqa* is also known as *shuruq*, but the name *shuruq* is also used for the dot, graphically identical to *dagesh*, in the glyph for *vav shruqa*. The pre-composed character U+FB35 HEBREW LETTER VAV WITH DAGESH represents *vav shruqa* as well as consonantal *vav* with *dagesh*, but use of this character is not recommended.

31 This author knows of only one case in the Hebrew Bible which is not fully disambiguated by this algorithm, a collocation of *holam* and *vav* following *hiriq* and *yod* in Jeremiah 52:19 BHS. This is disambiguated in BHS by the positioning of an accent above the *yod* which implies that the *yod* is not silent. In many Hebrew Bible manuscripts a *dagesh* is written in the *yod*, which also clearly disambiguates the *yod* and so the following *holam* and *vav*.

32 For efficient rendering, this canonical combining class should be less than that of any *vowel point*.

33 As in BHS *meteg* is never used in any position with *hataf qamats*, if this alternative is followed further study will be needed of whether this proposed *hataf qamats with meteg* would ever actually be used.