Hebrew Lexicon Documentation

by David Troidl

The Hebrew Lexicon is comprised of three XML files, for recording Hebrew lexical information.

1. BrownDriverBriggs.xml contains the current BDB content. It remains a work in progress. Entries will be filled out over time, especially in the area of completing the scripture references, and Hebrew words.
2. HebrewStrong.xml contains the content for Strong's Hebrew Dictionary. Numerous corrections have been made, since the initial offering, in a significantly different format. The layout now follows the Hebrew Mesh, with duplicated entries recombined.
3. LexicalIndex.xml is now the meeting place of BDB, Strong and the TWOT[1] numbers, again significantly improved from the original Strong's Dictionary, and incorporating corrections that have arisen in the process of constructing the new lexicon.

These are supplemented with an XML schema for each, and AugIndex.xml, which maps the augmented Strong numbers, used in the Open Scriptures Hebrew Bible, to Lexical Index ID values. The lexicon in this form is intended to be easy to use, easy to understand and easy to transform.

The lexical index is meant to bridge the gap between the accessibility and ubiquity of Strong's Dictionary, and the comprehensiveness and accuracy of BDB. For those versed in Hebrew, and wanting the depth of BDB, it can now be accessed on its own terms. For those more familiar with Strong's numbers, or having current applications based on them, the lexical index provides access directly to entries via the Strong's numbers.

Brown, Driver, Briggs

The document is contained in a `<lexicon>` element.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon xsi:schemaLocation="http://openscriptures.github.com/morphhb/namespace BdbSchema.xsd"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns="http://openscriptures.github.com/morphhb/namespace">
```

Part

Within the `<lexicon>` element is a sequence of `<part>` elements, one for each letter of the alphabet for Hebrew, and then again for Aramaic. Each part has an `id`, a `title`, indicating the Hebrew letter, and an `xml:lang` attribute, either `heb` or `arc`.

1 The *Theological Wordbook of the Old Testament*, by Gleason L. Archer and R. Laird Harris

Section

Within each <part> element is a sequence of <section> elements, to delineate the main elements in BDB, set in larger type, and often indicating root words that succeeding entries are derived from. Each section has an id. Within each <section> element is a choice of <entry> and <page> elements.

Entry

The <entry> elements contain the actual BDB entry data. For example:

```
<entry id="b.cw.aa" type="root" cite="full" mod="I"><w>בָּרָא</w> 53 <pos>vb</pos>.
      <def>shape</def>, <def>create</def>
   <sense><stem>Qal</stem> <asp>Pf</asp>.—<def>shape</def>, <def>fashion</def>,
      <def>create</def></sense>
   <sense><stem>Niph</stem>.—<em>Pass</em>.
      <sense n="1"><def>be created</def></sense>
      <sense n="2">with reference to birth</sense>
      <sense n="3">of something new, astonishing</sense>
   </sense>
   <sense><stem>Pi</stem>.
      <sense n="1"><def>cut down</def></sense>
      <sense n="2"><def>cut out</def></sense>
   </sense>
   <status p="135">base</status>
</entry>
```

The <entry> element has an id. There are also four optional attributes, representing features of the BDB text:

- type – distinguishes an entry (the default) from a root.
- cite – distinguishes whether BDB has a partial (the default) or full list of Scripture references. A full list is indicated in the text by a dagger (†).
- form – indicates whether the lexical form appears in the Hebrew Bible. Default is true.
- mod – contains a Roman numeral, where BDB distinguishes entries that way.

The <entry> element will contain mixed content, with <w>, <pos>, <stem>, <asp>, <def>, , <ref>, <foreign> and <sense> elements in various combinations. The final element within the entry will be a <status>.

Page

The <page> element marks the approximate position of a page break in the BDB text. Its p attribute lists the number of the new page. Once the lexicon is complete, the page breaks should be moved inside the entries, to more accurately reflect the text.

Word

The <w> element is a container for the Hebrew words. Generally, it will appear without attributes. Though in cross references to other words in the lexicon, it may contain a src attribute, which is an IDREF to the ID of the referenced word. It may also contain a mod attribute, when the reference is distinguished that way.

Part of Speech

The <pos> element contains the BDB abbreviation of the part of speech being referenced. A list of abbreviations will be included with the release of the prefatory material of the lexicon.

Stem

The <stem> element contains the BDB abbreviation of the stem of a Hebrew verb distinguished in a part of the entry. BDB often breaks down the references for a verb according to its stems.

Aspect

The <asp> element performs a similar function for the aspect of a Hebrew verb, within the various stems.

Definition

The <def> element contains a definition, as distinguished in BDB either by bold face or by italics. Normally, each in a series of definitions will have a separate <def> element. The intent is to make individual definitions easily identifiable.

Emphasis

For the occasional use of italics for emphasis, we have the element.

Sense

The <sense> element represents a distinguishable sense in the definition of the word. <sense> elements appear as children of the <entry> or other <sense> elements, to any degree of nesting required. <sense> elements have a single, optional attribute: n, to distinguish the various <sense> elements. Senses, as listed in BDB, can take numbers, letters, Roman numerals, or even Greek letters in parentheses. Often on verb entries, senses are distinguished, instead, by <stem> elements that begin the sense content.

The <sense> element can contain choices of <pos>, <stem>, <asp>, <def>, , <w>, <ref>,

<foreign> and other <sense> elements, in any order. In the document, <sense> elements are generally indented to indicate their level of nesting.

Reference

The <ref> element contains a Scripture reference. The content of the element is the BDB form of the reference, except that superscript verse references and listed with the colon notation. The r attribute contains the SBL[2] form the reference.

Foreign

The <foreign> element contains foreign words and phrases, mainly in Greek, Arabic, Aramaic, Syriac and Ethiopic, in their own script, but with an array of other languages, generally transliterated. Mostly, at least initially, it will serve as a placeholder for missing content, until someone proficient in those languages can fill in the content. The xml:lang attribute contains the standard ISO 639 codes. See http://www.sil.org/iso639-3/codes.asp.

The following table contains the languages included in the Hebrew Lexicon so far:

BDB	Language	ISO
Ar.	Arabic	ara
Aram.	Aramaic	arc
As.	Assyrian	akk
Eth.	Ethiopic	gez
𝔊	Greek	grc
Heb.	Hebrew	heb
𝔙	Latin	lat
Pers.	Persian	fas
Skr.	Sanskrit	san
Syr.	Syriac	syr

Status

The <status> element has a single attribute: p, indicating the page number of the BDB entry in my source text. Individual pages can be accessed as (for page 1) http://en.wikisource.org/wiki/Page:A_Hebrew_and_English_Lexicon_%28Brown-Driver-Briggs%29.djvu/25, where the last number should be the page number plus 24. The content of the element

2 Society of Biblical Literature.

is the current status of that entry in the development of the lexicon. Status values currently consist of:

- `base`: entries that have basic definitions from BDB.
- `ref`: entries that have scripture references added, but remain incomplete.
- `done`: entries that fully represent the BDB data.
- `new`: root entries from BDB, that have no corresponding Strong entry.
- `added`: root entries that had to be added to accommodate internal BDB information.
- `made`: cross-reference entries, noting the word in alphabetical order, but referencing the actual definition.

Strong's Hebrew Dictionary

The Hebrew Strong document is also contained in a `<lexicon>` element:

```
<?xml version="1.0" encoding="utf-8"?>
<lexicon xsi:schemaLocation="http://openscriptures.github.com/morphhb/namespace StrongSchema.xsd"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xmlns="http://openscriptures.github.com/morphhb/namespace">
```

Entry

Within the `<lexicon>` element is a sequence of `<entry>` elements. Each has an `id`, made up of the Strong's number preceded by an 'H', to conform with XML id requirements. An `<entry>` element will generally contain `<w>`, `<source>`, `<meaning>` and `<usage>` elements, in that order, though in some cases not all of these will be necessary. A `<note>` will occasionally be added after the `<w>`, to indicate corrections to errors in the original. For example:

```
<entry id="H1254">
    <w pos="v" pron="baw-raw'" xlit="bârâ'" xml:lang="heb">בָּרָא</w>
    <source>a primitive root;</source>
    <meaning>(absolutely) to <def>create</def>; (qualified) to <def>cut down</def> (a wood), <def>select</def>,
        <def>feed</def> (as formative processes)</meaning>
    <usage>choose, create (creator), cut down, dispatch, do, make (fat).</usage>
</entry>
```

Word

The initial `<w>` element contains the main word of the entry. The following attributes are allowed:

- `pos`: to contain the part of speech.
- `pron`: to contain the pronunciation.
- `xlit`: to contain the transliteration.

- `xml:lang`: to indicate the language, `heb` or `arc`.

In Strong's Hebrew dictionary, Hebrew and Aramaic are intermingled, in rough alphabetical order.

Note

A `<note>` element contains a simple textual description of an error in the original, that has been corrected in our lexicon.

Source

The `<source>` element strives to isolate information about the derivation of the entry word. Recasting the entries in a mixed content format, similar to the BDB lexicon, may be truer to the original, but we are working with existing digital sources that came to us in this format. The `<source>` element may contain mixed content, with <w> elements, to refer to other forms or other entries in the lexicon, `<note>` elements, for corrections, and <def> elements, to indicate definitions.

Meaning

Properly, definitions fall under the `<meaning>` element, using the `<def>` element. The `<meaning>` may also contain mixed content, with `<w>` and `<note>` elements, as above.

Usage

The `<usage>` element lists the various forms the word takes in the Authorized Version. Due to the format for the listing in Strong's dictionary, there may occasionally be `<w>`, `<note>` or `<def>` elements.

Word References

A `<w>` element within one of the other elements of the entry may have a `src` attribute, when it refers to another entry in the Strong's dictionary. The `src` is an IDREF. Otherwise, it may contain alternate forms of the main entry word, in which case it can have the same `pron` and `xlit` attributes as the main `<w>`.

Lexical Index

The main function of the lexical index is as a lookup table for the other data. My goal, however, was to make it a little more informative. It is contained in an `<index>` element.

```
<?xml version="1.0" encoding="UTF-8"?>
<index xsi:schemaLocation="http://openscriptures.github.com/morphhb/namespace LiSchema.xsd"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns="http://openscriptures.github.com/morphhb/namespace">
```

Part

The <index> is divided into two <part> elements, separating the Hebrew and Aramaic entries, as in BDB. Each <part> has an xml:lang attribute, either heb or arc.

Entry

Each <part> consists of a sequence of <entry> elements. The entry will have an id, intended for internal reference. For example:

```
<entry id="bxy">
   <w xlit="bārāʾ">בָּרָא</w> <pos>V</pos> <def>shape</def>
   <xref bdb="b.cw.aa" strong="1254" aug="a" twot="278"/>
   <etym root="ברא" type="main">byx</etym>
</entry>
```

An <entry> may contain <w>, <pos>, <def>, <xref> and <etym> elements. I normally list the first three on the same line, separated by spaces, with the <xref> and <eytm> on separate lines.

Word

The <w> element is the main entry word. It will have an xlit attribute, for the SBL[3] transliteration.

Part of Speech

The <pos> element indicates the part of speech, using the Hebrew Morphology Codes.[4] This is meant to give the most common part of speech, in the case where more than one applies.

Definition

The <def> element contains a simple definition. It generally will be the first BDB definition, except in the case of proper names, where the first Strong usage is taken. Problematic cases are decided on their own merit.

Cross References

The <xref> element is really the heart of the lexical index. It will contain the following attributes, when they apply:

- bdb: the id of the corresponding BDB entry.
- strong: the id of the corresponding Strong entry.

3 Society of Biblical Literature.

4 http://openscriptures.github.com/morphhb/parsing/HebrewMorphologyCodes.html

- `aug`: the augment, for cases where a Strong entry spans more than one BDB entry.
- `twot`: the corresponding TWOT[5] number.

Etymology

The `<etym>` element in meant to tie together related words. It has two different formats, based on the `type`.

With `type="main"`, the entry represents a main word, that others are derived from. In this case, the content will be a list of `id` values for its derivatives. It may also have a `root` attribute, listing the consonantal root, mainly for aid in parsing.

With `type="sub"`, the entry represents a derived word, and the content will be the `id` of its main word.

There are also entries with `type="single"`, and others with no `<etym>` element at all. In either case, more work will have to be done to establish etymological relationships. BDB actually contains a wealth of information about cognate languages, but this remains beyond the scope of the lexical index.

In my implementation, for my local software, I take the augmented Strong number, from a selected word in the OSHB. I then use AugIndex.xml, to look up the Lexical Index id for the word. From the entry in the Lexical Index, I locate the main word, and get all its derivatives. I use these to display a word group, along with the BDB entry and Strong entry of the original word.

5 The *Theological Wordbook of the Old Testament*, by Gleason L. Archer and R. Laird Harris