Exploring Digitization and Encoding Options for Ben Yehuda's Hebrew Dictionary

Sinai Rusinek: report on a week-long STRV visit to BCDH and ELEXIS-RS.

The seminal "Dictionary of Old and New Hebrew" was established as a significant step in the revival of Hebrew as a literary language, which started in Eastern Europe in the 18th century, and culminated in Eliezer Ben Yehuda's life's project, to make Hebrew into a spoken, living language. The publication of the dictionary lasted more than five decades and outlived Ben Yehuda, who in his lifetime saw the publication of only 5 of the 17 volumes.

The dictionary contains almost 8000 words and the entries include translations to German, French and English, Hebrew definitions, quotations from historical and contemporaneous sources and often etymological and other notes, connecting old words as well as neologisms to both Asian and European traditions, as well as to ideas about science, technology and progress. This plethora make the dictionary into an invaluable resource for the study of historical semantics, ethnography and cultural history of the period. Its main potential, however, can be fully realized when it has undergone a refined digitization process.

The digitization project of Ben Yehuda's dictionary is a crowdsourcing volunteer based project which is taking place in the framework of the larger Ben Yehuda Project, the "Israeli Gutenberg project" https://bybe.benyehuda.org/page/english, founded and spearheaded by Asaf Bartov of the Wikimedia foundation, who has also built the transliteration environment for the dictionary project. 2400 entries have already been transcribed. The result of the transcription, which is done manually, is a partly-structured HTML. The web interface enables basic search, and biblical quotations were already linked to the online Wikitext Hebrew Bible.

Funded by Elexis, a short term research visit in Elexis-RS at the Belgrade Center for Digital Humanities, hosted by Dr. Toma Tasovac, was dedicated to explore the potential of making the Ben Yehuda dictionary a structured research resource, according to the TEI Lex-0 modelling convention.

We discussed and experimented with platforms for publishing XML such as CETEICEAN and eXist-DB, which is the platform serving Raskovnik - the BCDH-designed platform for Serbian dictionaries. I was introduced to the oXygen and Github-based workflows for annotation and

editing. We spent considerable time analyzing the microstructure of the dictionary, and, finally, we created a proof-of-concept annotation of sample entries from the Ben Yehuda dictionary. The modelling process shed light on the potential as well as the challenges the lie ahead for a future annotation project.

The Ben Yehuda dictionary is highly structured with syntactic information and rich in examples, citations and notes. The following example is a snippet from the encoding of a sample noun entry exhibiting these elements:

```
<entry xml:id="BY_701" xml:lang="he" type="mainEntry">
   <form type="lemma"><orth>אֲבַטֵּחַ</orth></form>
   <!--<anchor corresp="#BY_701-NOTE1">1</anchor>-->
   <pc>,</pc>
   <form type="variant"><orth>אֲבַטִּיחַ</orth></form>
   <pc>,</pc>
   <gramGrp><gram type="pos" value="nm" norm="NOUN">ז"ש</gram></gramGrp>
   <pc>,</pc>
   <form type="inflected">
      <gramGrp><gram type="number" value="pl">ר"מ</gram></gramGrp>
      <form type="inflected"><orth>אֲבַטְּחִים</orth></form>
      <pc>,</pc>
      <form type="inflected"><orth>אֲבַטִּיחִים</orth></form>
   </form>
   <pc>,</pc>
   <pc>–</pc>
   <sense xml:id="BY_701-S1">
      <def> פרי אדמה מלא משקה, ממין הקשואים והדלועים </def>
      <pc>,</pc>
      <cit type="translationEquivalent" xml:lang="de">
         <form type="lemma"><orth>wassermelone</orth></form>
      </cit>
      <pc>;</pc>
      <cit type="translationEquivalent" xml:lang="fr">
         <form type="lemma">
            <orth>melon</orth>
         </form>
      </cit>
```

The citations, taken from biblical literature as well as from Mishnaic and Talmudic later sources, and to a lesser extent from modern Hebrew, is especially promising with regard to citation network analysis. A typology of the citation sources is required in addition to linking quotations with open resources where they may be read in context.

```
<cit type="example">
   <quote> זכרנו את הדגה אשר נאכל במצרים חנם את הקשאים ואת הָאֲבַטִּחים ואת החציר
      ואת הבצלים ואת השומים</quote>
   <ref
      target="https://www.sefaria.org.il/Bemidbar.11.5 https://he.wikisource.org/wiki/קטגוריה:במדבר_יא_ה"
      type="bibliography">
      <bibl type="biblical">(<title>במד'</title><citedRange> יא
         ה</citedRange>)</bibl></ref>
</cit>
<pc>.</pc>
<pc>–</pc>
<lbl>מ"ובתו:</lbl>
<cit type="example">
   <quote> איזהו גרנן למעשרות וכו' אבטיח משישלק </quote>
   <ref
      target="https://www.sefaria.org/Mishnah_Maasrot.1.5 https://he.wikisource.org/wiki/משנה_מעשרות_א_ה"
      type="bibliography">
      <bibl type="mishnaic">(<title>מעשר'</title>
         <citedRange>א ה</citedRange>)</bibl></ref>
</cit>
<pc>.</pc>
<cit type="example">
   <quote>האומר לחבירו הילך איסר זה וכו' באבטיח שאבור לי סופת ואוכל </quote>
   <bibl type="mishnaic">(שם <citedRange>ב ו</citedRange>)</bibl>
</cit>
<pc>.</pc>
<cit type="example">
   <quote>ובירק הקשואים והדלועים והאבטיחים</quote>
   <bibl type="mishnaic"> (שם<citedRange> א ד</citedRange>)</bibl>
</cit>
```

There are, however, some specificities of the dictionary and language that defy simple annotation: for example, inflected forms may appear in the dictionary as properties of a specific entry, or as different entries in a group. A sense is often not explicated for different inflections or form, as it is assumed to be obvious from the Hebrew inflection system. In these points the legacy dictionary conflicts with the expected regularity of the schema.

Also, an abundance of related entries is creating a complex hierarchy that calls for an elaborate system of unique IDs, covering the various categories (word family entry vs. main entry, nested (related) entry, and sense).

With an eye to automating as much of the process as possible, we formulated some preliminary rules based on a combination of regular expressions and XPath as a basis for a conversion scripts that will hopefully expedite the work in a future implementation of the project. I also familiarized myself with XProc-based sets of XSLT transformations which are used at BCDH for automatic and semi-automatic conversions of paper-based dictionaries to TEI.

I am grateful to BCDH and ELEXIS for giving me the opportunity to expand my knowledge and hone my technical skills.