The Old New Thing

Crowdsourcing the Digitization of Ben-Yehuda’s Dictionary
Don’t Panic

- Hebrew is written right-to-left
- ~100 years ago, Eliezer Ben-Yehuda researched and wrote the largest and (still!) most comprehensive dictionary of the Hebrew language
- It is now public domain
So...

- Scan → Commons
- Optical Character Recognition (OCR) → Wikisource
- Proofread on Wikisource
- enrich Wiktionary
- ...right?
Well...

As they say on Facebook relationship statuses:

It’s complicated...
Hmm...

- Appealing?
- Digestible?
- Manageable?
- Doable?
(My) Solution: go granular

- Major design principle: each headword (definition) is a separate unit of work, with its own status and lifecycle
- To do that: manually (but easily) partition scans into individual definition parts.
- Definitions live their lives: get assigned, transcribed, proofread (3 times!), published, and linked to.
System automagically stitches together entries beginning in page X, continuing through page X+1, and ending in page X+2.
Example of a single (short) entry

- digestible, manageable, clearly delineated.
- from the moment entries are separated from each other, everything else happens at the level of the individual entry, not “page”.

Senecio vulg...
The Bulk of the Work

- Web-based transcription system to type, fix-up (type Greek or Arabic), and triple-proofread each individual entry
- Entries move through these stages, and are published as soon as done.
- The dictionary is thus already partially available, with gradually-filled “holes”.
A liiittle bit of markup... Looks familiar? ;)}
1..n entries still to be completed
Magic!
Biblical citations are **automagically** linked to the specific verse on Hebrew Wikisource! :)
Numbers

- ~12000 pages in the dictionary
- ~300 entries completed during limited beta, by about 8 active volunteers
- (top contributor typed ~450 entries and proofread ~1150)
- After public launch, expect to complete dictionary within 2 years.
What shall we do tomorrow?

- More links from citations to extant sources online (Wikisource, Project Ben-Yehuda, HebrewBooks.org)
- Export entries as [draft] Hebrew Wiktionary articles
- Digitize other right-to-left dictionaries (Hebrew, Arabic, Farsi, ... )
Technical details

- Tools used: Ruby, Rails, jQuery, HTML5, Canvas, ImageMagick, TinyMCE
- Running on Debian Linux :)
- Code available on GitHub: https://github.com/abartov/ebydict
- If you can read Hebrew (and Rashi script), join us! http://ebydict.benyehuda.org
Project Ben-Yehuda?

- A volunteer-created digital library of public domain Hebrew textual works. cf. [[Project Gutenberg]], [[Project Runeberg]]
- Founded 1999. (Yes, before Wikipedia…)
- over 9000 works digitized, ~200 active volunteers
- Special retro-1980s Web design! :-p

Hi, Lars! :)

---
Thank you!

Asaf Bartov -- Project Ben-Yehuda
editor@benyehuda.org