Practical Linguistic Annotation

Roorda, Dirk

published in
International Journal of Humanities and Arts Computing
2017

DOI (link to publisher)
10.3366/ijhac.2017.0196

document version
Peer reviewed version

PRACTICAL LINGUISTIC ANNOTATION: THE HEBREW BIBLE[1]

DIRK ROORDA

INTRODUCTION

1. Annotation

An annotation is a piece of information attached to another piece of information.[2] Annotations generally do not have the same authorship, publishing workflow, and audience as the information sources they are attached to. Annotations serve to provide comments to sources, and these comments may involve analysis, explanation, correction, linking, evaluation, tagging, counting, and much more. In this article we focus on the logistics of information, rather than on the meaning. While it is useful to distinguish annotations for their type of content, our interest lies in the patterns of information distribution. How are annotations created, how are they published, and how do they behave in the research data cycle?

2. The Hebrew Bible

The Hebrew Bible is a family of ancient texts with a complex origin. It is recognized by several world religions, and it has pervaded large swaths of human culture. Academic research into the Bible occurs in several disciplines: linguistics, history, and theology with their specialties such as linguistic variation, historical linguistics, textual criticism, literary analysis, exegesis, and hermeneutics.

Religious communities have added their own sets of interpretations and observations. The practice of Bible translation into a great many languages of the world[3] has tuned people's antennas for interpretation. There are editions of the text of the Hebrew Bible in which the pages contain a small square of source text, surrounded by layers and layers of annotation.[4]

International Journal of Humanities and Arts Computing 11.2 (2017): 276–287
DOI: 10.3366/ijhac.2017.0196

www.euppublishing.com/ijhac

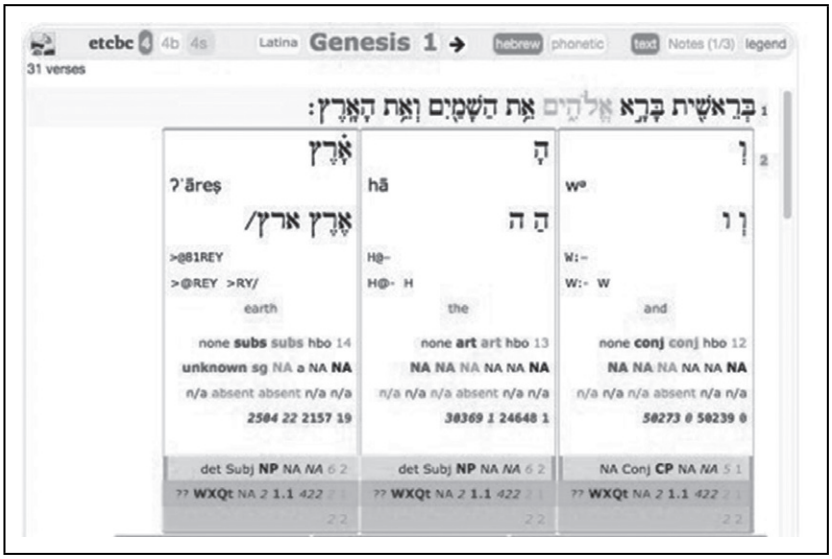

Figure 1. : Text and annotations in SHEBANQ. Clicking on a verse number hides and shows the annotations.

SHEBANQ: A SYSTEM FOR HEBREW TEXT

The ETCBC is the department of the Faculty of Theology at the Vrije Universiteit Amsterdam that has created a linguistic text database of the Hebrew Bible.[5] In 2013–2014 the SHEBANQ project has reshaped that database into a standard form: LAF[6] and has built a demonstrator to show new ways of utilizing that database in the age of internet connectedness. Indeed, the ETCBC database has been modeled as a huge set of annotations. This demonstrator is now a website in production, also called SHEBANQ.

We show how the Hebrew Bible has been captured in a system of annotations and point to a number of non-trivial, innovative uses of the concept of annotation which were not possible or practical before the digital handling of information.

1. Exhaustive linguistic annotation

Each of the more than 400,000 words carries annotations specifying its part of speech, it morphological characteristics, its various representations and more. The same holds for larger units, such as phrases and clauses. All in all, this gives tens of millions of annotated features. Before the arrival of digital information processing, this was not a feasible thing to do. But here we have it: a text with millions of annotations, online, in a working system: SHEBANQ (see Fig 1).

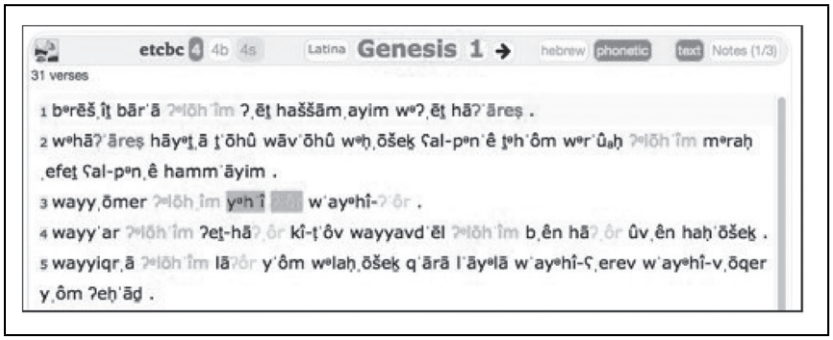

Figure 2. : Text in phonetic representation, with all markings and annotations in place.

2. Multiple textual representations as annotation

There is something else to note: the text itself exists as the content of annotations. This has to do with the peculiar fact that the older variants of biblical material were written down in a consonantal script, while the vowels were added as diacritical marks ('pointing') several centuries later, near the final consolidation of the text around 900 AD. So every word still has a consonantal representation, but also a fully 'pointed' representation. It is a clear case where the text does not have a single representation. Annotation provides a neat way to expose those representations together.

Further down that road, we also provide a phonetic representation of the text (see Fig. 2). That will help people not familiar with Hebrew to get access to the linguistic annotations and use it for their own purposes.[7] Nevertheless, the authoritative text of the Biblia Hebraica Stuttgartensia is the default representation.[8]

In SHEBANQ, the annotations are not tied to the representation of the text. So if the user switches representation, all the highlights and other annotations remain in place.

3. Queries as annotations

Now that text and linguistic annotations reside in a database, it becomes possible to query both kinds of data. An important objective of the creators of the ETCBC database has always been the ability to search for peculiar syntactic patterns. When reading the Bible, every now and then a passage is particularly problematic and requires explanation. But what kind of explanation? Has there been a text transmission error? Is there a hidden borrowing from another text? Is there a syntactic construction that belongs to another dialect or language? Is

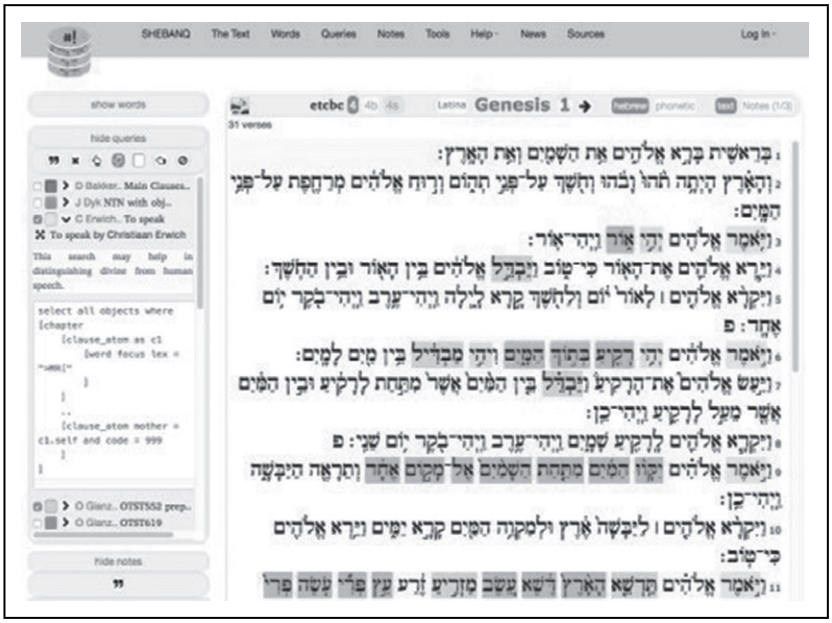

Figure 3. : Queries as notes in the margin. The reader of the passage is drawn to exegetical problems of others, and their solutions.

there deliberate use of language to achieve a literary effect? Or is there a truly special meaning lurking behind the text? Research into these problems is greatly helped by catalogues of occurrences of the same or partly the same phenomenon. By using a text database, we are able to systematically query those patterns.

It is not easy to write such queries. The data is full of unexpected patterns, it is easy to miss cases, so many checks and cross-checks are needed. A successful query is a piece of scholarly crafts(wo)manship, and should be shared and published as such.

Seen in an abstract way, a query is an annotation to all its results. One annotation targeting multiple passages is already a little bit innovative, although one might say that cross-references and indexes are examples of multi-target annotations. But here there is a bit more going on. By presenting a query as an annotation to its results, an unexpected flow of information is made possible: from result to query. When a scholar reads a difficult passage, (s)he might be interested in the exegetical queries that have results in that passage (see Fig. 3). This is exactly what SHEBANQ makes possible. Next to every chapter in the Bible a list of relevant queries is presented, and the results of those queries are highlighted in the chapter at hand.[9]

4. Semi-automatic analysis as annotation

Linguistic research into the Hebrew Bible has not ended. The meaning of Hebrew verb forms in poetry is a long-standing problem (and many occurrences in prose are far from clear for that matter), and data-driven research has the potential to produce new solutions.[10] Verb meanings are also dependent on the number and nature of constituents in the sentence (verbal valence), and it is worthwhile to devise a flow chart system to generate verb senses on the basis of signals near verb occurrences.[11] This involves a lot of trial and error. Sometimes it leads to a review of the linguistic encoding, to new syntactic and semantic distinctions. One way to organize this, is to generate the results of a flow chart as a set of annotations to be presented next to the text. The researcher can then see the decisions in full context and comment on those outcomes by manual annotations. These annotations can be harvested in turn and provide a basis for an improved algorithm. This workflow is supported on SHEBANQ, although not many people are fully utilizing it yet.

Experience, however, shows that it is cumbersome to execute this work exclusively on a website. A website such as SHEBANQ only supports that many use cases, while every research activity requires its own data preprocessing. An efficient workflow for this kind of research is to collect data, store it in spreadsheets, have the researcher work on them, and then feed the filled-in sheets back into the system. We support this workflow by means of LAF-Fabric, which is an off-line companion to SHEBANQ, based on exactly the same data. With the help of LAF-Fabric, the programming scholar can grab all data that is needed for a particular task, lay it out neatly in columns, and convert edited sheets into new sets of annotations.[12] The work of verbal valency is available on the SHEBANQ tools page (see Fig. 4). These new annotations have been bulk-imported into SHEBANQ and pubished, but they can also serve as basis for new algorithms in LAF-Fabric.[13]

5. Everything else

Although versatile, SHEBANQ cannot do everything. For example, teaching Hebrew to academic students could profit from SHEBANQ, but SHEBANQ is not optimized for it. There is a system called Bible Online Learner[14], based on the same ETCBC database, that has facilities to generate drills and exercises for students and score their answers. Rather than to try to pack all functionality into one system, it is better to have several systems around, each geared to their own task, but yet knowing of each other's existence. Every chapter page in SHEBANQ links to the corresponding chapter page in BibleOL and vice versa. Moreover, in order to compose exercises, BibleOL uses queries that are published in SHEBANQ (see Fig. 5).

Figure 4. : Verbal valence notes have been bulk-imported into SHEBANQ and are visible in notes view. Users can mute note sets and focus on the topics of their interest.

Figure 5. : Interlinking with Bible Online Learner. Clicking on the SHEBANQ logo takes you to SHEBANQ, where there is a Bible OL logo to link you back.

6. Summing up

In the digital age, annotation has become a practical paradigm to carry out scholarly work: we can use annotations in quantities unheard of, to achieve old goals in new ways, and to pursue new goals with new workflows.

The reader is invited not only to look at the screenshots, because they tend to show screens packed with information. One of the strong points of digitally displaying information is that most of the material can be hidden most of the time. SHEBANQ as an annotation tool helps the researcher to collect all data relevant to the task at hand in one or two screens, for a great variety of tasks. And where SHEBANQ falls short, the companion tool LAF-Fabric takes over, but the price is that the user must program it. This is where the digital paradigm affects (or should we say *in*fects) the daily work of the scholar: programming skills are becoming increasingly relevant.

An important characteristic mentioned in most of the cases above is the facility to share and publish annotations. The Hebrew Text database is the result of a lot of scholarly work, and that work should be published, not only for the academic

record, but also for the purposes of teaching and training.[15] Moreover, published annotations enable useful cooperation of different systems based on the same data.

REQUIREMENTS FOR SCHOLARLY ANNOTATION

In the previous section we described annotations in action. When the action is research, it is important to comply with a few essential requirements.

Archiving

We saw how annotations capture scholarly work, sometimes at a high level of abstraction and expertise. So scholars must be able to save annotations and then share and publish them. Researchers that work years from now must be able to retrieve annotations when they see the sources, and to retrieve the sources when they see the annotations. While the digital paradigm is very beneficial to transform information flexibly and distribute it globally, it is much more challenging to fix existing information rigidly and distribute it over decades to come.

The digital age calls for digital archives that recognize these challenges and do something about it. In the SHEBANQ case, the data has been archived at DANS[16], all the code sits on Github (see an overview of the sources) and repository snaphsots have been archived at Zenodo at CERN. The live website is run by DANS on a server of the Royal Netherlands Academy of Arts and Sciences.

Coupling

The particular thing about annotations is that they need the coupling to another resource in order to be 'to-the-point'. In the age of analogue resources, this coupling tended to be tight: in the margins, or as footnotes, usually within the same material container. Where the coupling was less tight, such as in endnotes, indexes, registers as separate books or volumes, it became quickly unwieldy to handle all relevant annotations.

In the digital age these problems of information logistics can be solved much more elegantly and effectively, provided certain agreements are being made by the designers of information. It is a bit like geotagging photos by means of a recorded GPS track: if the track points are coded with the same time codings as the photos, the photos can be located on the track and then on the map. For annotations we need anchors: points in sources to link to. These points should be standardized so that different scholars, as producers of annotations, use the same anchors. That will help to make their annotations interoperable.

For linguistic annotations, the LAF standard helps a lot to refer to primary data in an objective way, although these anchors are still project dependent. There are efforts to bring about a more global persistent linking system to canonical resources (see Canonical Text Services and the CITE architecture), and it is a matter of time before it will be applied to the Hebrew Bible as well.

The holy grail of this all is the Linked Open Data (http://linkeddata.org) endeavour, which is an attempt to map all entities in human discourse unto unique, persistent identifiers, and code all properties that can be expressed into triples consisting of a subject, predicate and object, according to well-defined vocabularies and ontologies. This is a huge modelling effort, and it is not always clear how computing-intensive workflows may take advantage of it. But for importing and exporting data across boundaries of project and discipline, this is definitely the way to go.

An advantage of well-coupled annotations is that they can be sorted and organized on the basis of where they point to. But we need other organizing principles as well, such as the provenance of an annotation (researcher, project, organization), time (creation, update), motivation (correction, evaluation), nature (linguistic, hermeneutical). Of these, motivation and nature can be entered in free text description fields, which in practice, sadly, quite often reveal the text 'None'.

Innovation

A lot of digital development starts with mimicking analogue concepts. After a certain period, those digital counterparts may exhibit new dynamics. This only happens if the new concepts manage to exploit typical advantages of the digital paradigm over the old ways. One of the key digital advantages is the *network effect*: for certain tasks it has become possible to mobilize many people with mostly limited contributions. Such loosely organized networks can deliver impressive results, such as Wikipedia.[17] If scholars grab the opportunity to 'socialize' parts of their workflows, they may gain results not previously possible.

SHEBANQ has socialized the art of making exegetical queries. It is being used in the classroom, and scholars can quote queries to each other and cite them in papers. Everybody may enter new queries. And everybody can comment on specific query results by means of simple manual annotations. However, we are not seeing (yet) that kind of spontaneous manual annotation.

Reflection and action

Before building SHEBANQ, we tried to design its layout and the details of how queries should be displayed to the user. Query results are structured objects, and queries may have many structured results; it was not at all clear how we could

provide the users with a good visual representation of query results, and how to show them in context.

Most of this became clear after we started construction. Only fully engaging in building this web app made us discover one unanticipated problem after another, and solve them all. For example, we decided to provide on-the-fly heat maps of query results, which give users an instant overview of how the results of a particular query are distributed in the Bible (see Fig. 6). But we refrained from presenting query results in their full complexity as structured objects. We also modified our goals. Rather than make SHEBANQ into the ultimate research tool, we developed LAF-Fabric as an off-line side tool, with more flexibility to tackle the nitty-gritty of daily research. SHEBANQ got redefined from a laboratory to a showroom of research results, where very diverse research output comes together in one context. Now SHEBANQ and LAF-Fabric together provide the facilities of a scholarly lab.

In our opinion, it makes no sense to reflect on the nature of annotations without being involved in digital construction work. The ontology of a (digital) medium is the reflection of its usage patterns. When migrating annotations from analog to digital, we are potentially upsetting those very usage patterns, and hence the ontology of annotations.

Programming skills

Just as analogue information systems presuppose the skills of reading and writing, the potential of the digital media cannot be unleashed without new skills. For researchers, this means definitely: programming. Especially where experimentation is involved, it is impractical to outsource development of new tools to 'mere' programmers. Instead, scholarly teams should insource programming skills in their own skulls. They do not need to master professional levels. Data oriented programming has become much easier by the evolution of scripting languages such as Python and additional tools such as the Jupyter notebook.[18] And not every team member needs to learn to program, if only the team as a whole is able to produce experimental or pilot solutions. Only after many experiments by scholars, it will be the right time to bring the professional coders in to turn the successful pilots into products and infrastructure.

Addendum

From the start of 2017 onwards, I have deprecated LAF-Fabric in favour of a new format and tool: Text-Fabric.[19] Thanks to the move from an XML based format into a plain text based format all data fits in a Github repository.[20]

Figure 6. : Heat map of query results. Every square represents a block of 500 words of Bible text. The color indicates how many result words the query has in that block. Every square is clickable and takes you to the corresponding passage.

END NOTES

[1] This works rests on the shoulders of the giants at the ETCBC, such as Eep Talstra and Constantijn Sikkel who conceived the database and made it work through the decades behind us. See E. Talstra and C. J. Sikkel, 'Genese und Kategorienentwicklung der WIVU-Datenbank', in C. Hardmeier et al., ed., *Ad Fontes! Quellen erfassen—lesen—deuten. Was ist Computerphilologie? Ansatzpunkte und Methodologie—Instrument und Praxis* (Amsterdam, 2000), 33–68; E. Talstra, 'Computer-assisted linguistic analysis. The Hebrew Database used

in Quest.2', in J. A. Cook, ed., *Bible and Computer. The Stellenbosch AIBI-6 Conference. 2000–07–17/21, Stellenbosch: Proceedings of the Association Internationale Bible et Informatique* (Leiden, 2000), 3–22, https://shebanq.ancient-data.org/shebanq/static/docs/methods/2000_Talstra_QuestDataTypes.pdf. The query engine of SHEBANQ is the one made by Ulrik Petersen. See U. Peterson, 'Emdros—a text database engine for analyzed or annotated text', *Proceedings of COLING 2004*, 1190–3, http://emdros.org/petersen-emdros-COLING-2004.pdf; U. Peterson, 'Principles, Implementation Strategies, and Evaluation of a Corpus Query System', *Lecture Notes in Computer Science,* 4002 (2006). 215–26, http://link.springer.com/chapter/10.1007%2F11780885_21; U. Peterson, *EMDROS. Text database engine for analyzed or annotated text*, 2002–2014, http://emdros.org. Peterson has relied on the ideas of Christ-Jan Doedens: C.-J. Doedens, *Text Databases. One Database Model and Several Retrieval Languages* (Amsterdam, 1994). Researchers, senior and junior have put data and tools to many tests: Janet Dyk, Reinoud Oosting, Oliver Glanz, Gino Kalkman, Martijn Naaijer, Christiaan Erwich, Cody Kingham plus 89 users of SHEBANQ that shared 686 queries with us.

[2] See M. Bauer and A. Zirker, 'Whipping Boys Explained: Literary Annotation and Digital Humanities', in Ray Siemens and Kenneth M. Price, eds., *Literary Studies in the Digital Age: An Evolving Anthology* (New York, 2015), https://dlsanthology.commons.mla.org/whipping-boys-explained-literary-annotation-and-digital-humanities/; and M. Bauer and A. Zirker, 'Explanatory Annotation of Literary Texts and the Reader: Seven Types of Problems', this volume.

[3] See M. Cysouw, 'Parallel Bible Corpus. 1169 unique Bible translations', n.d., http://www.paralleltext.info/data/, and C. A. Christodoulopoulos, 'A multilingual parallel corpus created from translations of the Bible', https://github.com/christos-c/bible-corpus, 22 June 2017.

[4] See also R. Siemens et al., this volume.

[5] D. Roorda, 'The Hebrew Bible as Data: Laboratory—Sharing—Experiences', in J. Odijk and A. van Hessen, eds., *CLARIN in the Low Countries*, 2015, https://arxiv.org/abs/1501.01866; D. Roorda, J. Krans, B.-J. Lietaert-Peerbolte, W. T. van Peursen, U. Sandborg-Petersen and E. Talstra, 'Scientific report of the workshop Biblical Scholarship and Humanities Computing: Data Types, Text, Language and Interpretation, held at the Lorentz Centre Leiden from 6 Feb 2012 through 10 Feb 2012', Lorentz Center, Leiden, 2012, http://www.lorentzcenter.nl/lc/web/2012/480/report.php3?wsid=480&venue=Oort, 22 June 2017.

[6] N. Ide and L. Romary, *Linguistic Annotation Framework*, 2012, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=37326, 22 June 2017.

[7] F. de Vree, 'Using social co-occurrence networks to analyze biblical narrative', 2016, https://github.com/Fred-Erik/social-biblical-networks, 22 June 2017.

[8] See K. Elliger and W. R. Rudolpfh, eds., *Biblia Hebraica Stuttgartensia*, 5th corrected edition (Stuttgart, 1997), www.bibelwissenschaft.de/online-bibeln/biblia-hebraica-stuttgartensia-bhs/lesen-im-bibeltext/, 22 June 2017.

[9] See Roorda and van den Heuvel for an early formulation of the idea of queries-as-annotations; D. Roorda and C. M. J. M. van den Heuvel, 'Annotation as a New Paradigm in Research Archiving', *Proceedings of ASIS&T 2012 Annual Meeting. Final Papers, Panels and Posters*, 2012, http://arxiv.org/abs/1412.6069, 22 June 2017.

[10] G. J. Kalkman, *Verbal Forms in Biblical Hebrew Poetry: Poetical Freedom or Linguistic System?* PhD thesis, VU University (Amsterdam, 2015), https://shebanq.ancient-data.org/tools?goto=verbsystem.

[11] J. W. Dyk, O. Glanz and R. Oosting, 'Analysing Valence Patterns in Biblical Hebrew: Theoretical Questions and Analytic Frameworks', *Journal of Northwest Semitic Languages*, 40 (2014), 43–62, https://shebanq.ancient-data.org/shebanq/static/docs/methods/2014_Dyk_jnsl.pdf.

[12] See Roorda, Naaijer, Kalkman, & van Cranenburgh for initial examples; D. Roorda, M. Naaijer, G. J. Kalkman and A. van Cranenburgh, 'LAF-Fabric: a data analysis tool for Linguistic Annotation Framework with an application to the Hebrew Bible', *Computational Linguistics in the Netherlands Journal*, 4.4 (2015), preprint http://arxiv.org/abs/1410.0286.

[13] Indeed, using LAF-Fabric requires programming skills. It is a Python package that gives streamlined access to the Hebrew Text Database. A beginner's course in Python is enough to get started. Another, even more computationally intensive, example is the quest for parallel passages in the Bible. This is part of the Syntactic Variation project, carried out by a team of (PhD) researchers at the ETCBC. To see what is at stake here, see R. Rezetko and M. Naaijer, 'An Alternative Approach to the Lexicon of Late Biblical Hebrew', *Journal of Hebrew Scriptures*, 16.1 (2016), www.jhsonline.org/Articles/article_213.pdf.

[14] N. Winther-Nielsen and C. Tøndering, *Bible Online Learner*, n.d., http://www.bibleol.3bmoodle.dk/, 22 June 2017.

[15] SHEBANQ is meant as a service to publish queries for the academic record. DANS, as a national research data archive, is capable of archiving the database as a whole. It is also possible to store the data on Github, and preserve a snapshot of the repository to Zenodo, a service of CERN to preserve repositories for the academic record.

[16] E. Talstra, C. J. Sikkel, O. Glanz, R. Oosting and J. W. Dyk, *Text Database of the Hebrew Bible*, 2012, http://www.persistent-identifier.nl/?identifier=urn:nbn:nl:ui:13-ukhm-eb; W. T. van Peursen and D. Roorda, *Hebrew Text Database in Linguistic Annotation Framework*, 2014, PID: urn:nbn:nl:ui:13–048i-71, http://www.persistent-identifier.nl/?identifier=urn:nbn:nl:ui:13-048i-71; W. T. van Peursen and D. Roorda, *Hebrew text database ETCBC4b. Dataset available online at Data Archiving and Networked services*, Den Haag, 2015, dx.doi.org/10.17026/dans-z6y-skyh.

[17] S. Clay, *Here Comes Everybody: The Power of Organizing Without Organizations* (London, 2012).

[18] F. Pérez and B. E. Granger, 'IPython: a System for Interactive Scientific Computing', *Computing in Science and Engineering*, 9.3 (2007), 21–29, http://ipython.org, ISSN: 1521-9615, DOI: 10.1109/MCSE.2007.53.

[19] Text-Fabric: Data model, file format and processing tool for annotated texts. https://github.com/ETCBC/text-fabric/wiki.

[20] Text-Fabric-Data: Text and Annotations of the Hebrew Bible and the Greek New Testament. Includes documentation of the annotation features. https://etcbc.github.io/text-fabric-data/.

AQ: Please provide missing abstract for this article.