

System for HEBrew Text: ANnotations for Queries and Markup

Dirk Roorda

DANS

<http://dans.knaw.nl/en/content/contact/staff-members/dirk-roorda>

The languages in which the Hebrew Bible are written, Hebrew and Aramaic, are dead languages. Written manuscript sources are heavily relied upon when studying these languages. The study of such a historical body of texts involves research questions from different disciplines. Linguistic analysis is a stepping stone which must be followed by questions at higher levels of abstraction, such as literary questions: how did authors use the system of the language to craft their design: i.e. style, literary effect, focus, and all those features of the text that are not dictated by the language system? Another line of questions falls into historical linguistics: systematically charting linguistic variation in the biblical linguistic corpus can help addressing the question as to whether the variation reflects diachronic development.

The central data source in this discipline is the Biblia Hebraica Stuttgartensia version of the text of the Hebrew Bible. This text has been enriched with features that primarily result from linguistic analysis. It has taken decades of manual and automated work to arrive at a database that is completely marked up with linguistic information up to the syntactical level. The work is on going: now features at the discourse level are being added to the database.

This database is not just an SQL database.

In his Ph.D. thesis, Doedens (1994) explored ways to represent language data so that they can easily be queried in linguistically significant ways. The central concept is that of objects with features, where each object can carry unlimited features, and where objects can be aggregated arbitrarily into new objects. Based on this object-feature model, Petersen (2004, 2006) developed an actual query language implementation. The data resides in an ordinary relational database conforming to the relational translation of an object-feature model, and the queries are executed in a front end that interacts with the relational database engine. This query language is called MQL [7], and, as in SQL, a dump of the data can be represented in an MQL file. A dump of the Hebrew Text database in MQL format is archived at DANS [5].

One of the salient points made by Crist-Jan Doedens in [4] was the concept of "topographicity". The idea is that there must exist an isomorphism between the structure of the query and the structure of the objects found. This isomorphism, in particular, is to be construed as holding with respect to two "textually topological" relations which are central to text: "sequence" and "embedding". An MQL database is optimized to query for patterns of embedding and sequence. The data can carry multiple structures of embedding, but only a single linear order, although there are developments to relax that. MQL is not just an idea in a Ph.D. thesis, it has been implemented as well, by Ulrik Petersen [8].

The SHEBANQ project is aimed at preserving this database, and moreover, to turn it into a hub for linguistic and literary research. Some of it can be seen in a demo application [1,2]. First of all, a representation of the database in the Linguistic Annotation Framework (LAF, an ISO standard) has been made, and secondly, a query saver will be made. The query saver is a tool by which researchers can share their queries, in particular those that form the basis for published interpretations. Saving a

query involves creating a persistent reference to the query instruction, the query results and metadata about the query. That can be published as an OpenAnnotation, hence the queries-as-annotations slogan. See [3] about the role of archives with regards to annotation.

LAF was chosen as a preservation format, because its model of stand-off markup fits very well with the concepts modeled in the database: objects and features. The annotations in LAF contain feature structures with features that are linked to ISOcat. The linking is done by means of a TEI feature declaration document, for which we created a customized TEI schema. Currently we are not sure that this is the optimal way of linking LAF features to ISOcat definitions.

SHEBANQ is a cooperation between the VU University Amsterdam and DANS (Data Archiving and Networked Services), funded by CLARIN-NL.

References

- [1] Queries-As-Annotations demo application: <http://demo.datanetworkservice.nl/qaa>
- [2] Queries-As-Annotations Wiki:
http://demo.datanetworkservice.nl/mediawiki/index.php/Queries_As_Annotations
- [3] Dirk Roorda, Charles van den Heuvel: Annotation as a New Paradigm in Research Archiving. ASIS&T 2012 Annual Meeting Proceedings of Final Papers, Panels and Posters. <https://www.asis.org/asist2012/proceedings/Submissions/84.pdf> (author's version: <http://demo.datanetworkservice.nl/mediawiki/images/8/84/ASIST2012-Annot-DR-ChvdH-final-submission.pdf>)
- [4] Doedens, C.F.J. (1994). Text Databases. One Database Model and Several Retrieval Languages. Language and Computers, Number 14. Editions Rodopi Amsterdam. Amsterdam and Atlanta, GA. ISBN: 90-5183-729-1.
- [5] Talstra, E., Sikkel, C., Glanz, O., Oosting, R., Dyk, J.W. (2012). Text Database of the Hebrew Bible. Dataset available from Data Archiving and Networked services after permission of the depositor. <http://www.persistent-identifier.nl/?identifier=urn%3Anbn%3Anl%3Aui%3A13-ukhm-eb>
- [6] Petersen, U. (2004). Emdros - a text database engine for analyzed or annotated text. Proceedings of COLING 2004. 1190–1193. <http://emdros.org/petersen-emdros-COLING-2004.pdf>.
- [7] Quick intro MQL: <http://emdros.org/foundation.html#WhatIsMdf>
- [8] Petersen, U. (2006). Principles, Implementation Strategies, and Evaluation of a Corpus Query System. Lecture Notes in Computer Science Vol 4002, 2006, pp. 215-226. http://link.springer.com/chapter/10.1007%2F11780885_21