LAF-Fabric: a data analysis tool for Linguistic Annotation Framework with an application to the Hebrew Bible

Dirk Roorda (Data Archiving and Networked Services (DANS) (The Hague) / The Language Archive (TLA) (Nijmegen))

The Hebrew Bible is an object of linguistic, literary and historical research. Its text has been entered in a database with careful linguistic markup. This text database has a user friendly query language called MQL, implemented in the EMDROS database system (Open Source, emdros.org). Recently the database has been exported to a modern, interoperable format, Linguistic Annotation Framework (LAF). This paper discusses LAF-Fabric (GitHub, laf-fabric.readthedocs.org), a new tool to analyse LAF resources in general with an extension to process the Hebrew Bible in particular.

The LAF representation of the text is essentially a graph, and hence the baseline mode of data access is to walk over node sets and use edges to explore the neighbourhood of nodes. Each node is loaded with features, from which the text and linguistic properties may be read off. LAF-Fabric is operated best in the IPython notebook, where you can populate tables, vectors, trees and graphs with the data obtained from walking the LAF data. You can use the facilities of the Python ecosystem to perform data analysis and visualisation. I wrote a tutorial notebook, gender, which shows all these steps in an exemplary way. More seriously, we use LAF-Fabric to extract tree structures from the Hebrew Bible, which are not encoded explicitly. We also facilitate issuing MQL queries and retrieving the query results as node sets inside LAF-Fabric. This gives us a powerful tree querying device on top of graph oriented processing.

Analytic study of a historical body of texts benefits from treating the text as data in a database. When the need arises to make that data interoperable, the most natural representations are those using stand-off markup. While there are no general purpose ways to perform analysis on such resources, there are attractive methods within reach: one with tree queries (MQL) and another with graph walks and the full power of a scripting language at hand (LAF). Both approaches, especially when combined, empower researchers in the humanities with the digital tools to analyse their data.

Representing data in LAF and processing it efficiently turned out not to be straightforward at first sight. But through the effort of experimenting with LAF-Fabric, we can now show a new, powerful scenario of working with stand-off data. It makes the methods of scientific computing applicable to the study of linguistic/historical resources.