Hebrew Text Database in LAF

Hebrew Text Database ETCBC4

VU University Amsterdam

0. Introduction
This dataset has been created by the SHEBANQ (System for HEBrew Text: ANnotations for Queries and Markup) project.
The project has converted the Hebrew Text Database of a team of theologians at the VU University Amsterdam into an archivable format: Linguistic Annotation Framework (LAF).
This database has existed for many years and is also known as the WIVU database (Werkgroep Informatica Vrije Universiteit).
It is an ongoing effort to code the Hebrew text of the Bible as a database, which increasing precision, completeness and scope.
With the retirement of Prof. Dr. Eep Talstra in 2011, the group is lead by Prof. Dr. Wido van Peursen, and its new name is Eep Talstra Centre for Bible and Computing (ETCBC).

A previous version has been archived in 2012 at DANS [9]. This dataset [10] differs from the previous one in many ways:
A. it contains a new version of the data, called ETCBC4. The content has been heavily updated, with new linguistic annotations and a better organisation of them, and lots of additions and corrections as well. However, this new version is a version in progress. Because of the deadline of SHEBANQ, we could not wait for the finalization of the ETCBC4. Yet we have made a thorough effort in ensuring that the text representation meets the highest standards, and that the linguistic features are consistent and as complete as possible.
B. the data format is now Linguistic Annotation Framework (LAF [3]). This contrasts with the previous version, which has been archived as a database dump in a specialised format: Emdros (see the link below).
C. a new tool, LAF-Fabric is added to process the ETCBC4 version directly from its LAF representation. LAF-Fabric has been instrumental in producing the LAF data from the ETCBC sources and in analyzing and checking the resulting data. See the image below for an impression.
D. Extensive documentation has been included. There are manpages from within the ETCBC department, there are quickref manuals, there are lots of on-line pages stored off-line in this dataset. Together they give a profound insight in the approach by which the ETCBC has developed their Hebrew Text Database.
Hebrew Text Database in LAF

GEN 01,03 words 6, 7, 8
monad 36, 37, 38, node 37, 38, 39

WA75-J:HIJ&>O75WR00

Psalm 6

שתשת כלות עני עותקה בצל-אודיה:
שמות מספרים אונג יראים אותם קהל בכם:
שמעו חיות המתים ו(Random יבשה)
שהם חשים את הקהל של الملك
ויבשו רמלה עד כל-אודיה משוב רעוה:
שהם חשים את הקהל של الملك
ונפשם בצל-אודיה על יום יראתם:
שמות הקהל של الملك יראם כל-אודיה
ונפשם בצל-אודיהрош בצל-אודיה:
שמות הקהל של الملك יראם כל-אודיה
ונפשם בצל-אודיהрош בצל-אודיה:

ETCBC4 2 of 9 urn:nbn:nl:ui:13-048i-71
1. Data

Conceptually, the Hebrew Text database is a collection of objects: books, chapters, verses, sentences, clauses, phrases, word and others. These objects carry features. It is in these features that the knowledge about the text is coded. For example, word objects carry a feature part of speech which tell whether the word is a noun, verb, etc. Moreover, word objects carry the way they are represented. A Hebrew word may be represented in a Latin transcription, in Hebrew characters (UNICODE) with or with vowels, or as lexeme. All these different ways are coded in features. Other kinds of objects have different features. In particular, dependency relationships between clauses and phrases are coded by the feature mother.

1a. QUEST

The ETCBC holds this data internally in a set of plain text files, in a format known as the QUEST database format. This format is not suitable for the purpose of analysis. Therefore, the ETCBC exports its data in a text database.

1b. EMDROS

The data model describe above is not a typical relation model, although it could be expressed in a purely relational database. As Doedens[1] has pointed out, there are better ways to store texts, where the properties of sequential order and embedding are naturally modelled. Ulrik Petersen has picked up these ideas and turned it into a working database system: Emdros [2]. The previous archived version of the (then) WIVU database consisted of an EMDROS database dump, which is an MQL file, a text file that resembles the structure of SQL.

1c. LAF

Since 2012 there is an ISO standard for the stand-off markup of language resources, Linguistic Annotation Framework (LAF) [3]. As a result of the SHEBANQ project, funded by CLARIN-NL and carried out by the ETCBC and DANS, we have a tool, LAF-Fabric, by which we can convert EMDROS databases into LAF and then do data analytic work by means of IPython notebooks. This has been used for the Hebrew Text Database in its ETCBC4 version, but it can also be applied to the Syriac text in the Peshitta.

The underlying model of LAF is a primary data source (plain text) with secondary annotations (linguistic markup). Between the text and the markup is a graph of nodes and edges, where some nodes can be linked to regions of text. In our case, the mapping between the data model as described above, and LAF, is: objects are turned into nodes, features are turned into annotations, relations between objects are turned into edges. See the graphical picture below.
1d. Condensed LAF

LAF is a verbose format. A LAF resource is a bunch of XML files adhering to the GrAF schemas [4]. The ETCBC4 is over 1.6 GB large. It is not obvious how to process this material efficiently and get targeted information out of it. The resource has to be loaded in its entirety, and the associated XML parsing takes 10 minutes on a present day laptop, and its consumes all of its memory. That is why we have built a tool for efficient LAF processing which is included here: LAF-Fabric. This tool compiles the data once and for all to a compact format that loads in seconds. This compiled format is basically a bunch of Python data structures, pickled and zipped to disk.

1e. Data summary

All of the data formats mentioned above, except QUEST, are present in this dataset. The LAF resource is the main archived source, present in the folder lat, while its schemas and feature declarations are present in the folder decl. However, there is also a zip file, laf-fabric-data.zip which contains the other representations. See the section laf-fabric-data.

1f. Metadata

The linguistic features in this dataset are registered in ISOcat [7,8] a concept registry.

2. LAF-Fabric

The tool needed to perform the conversion from EMDROS to LAF and to perform analytical tasks on LAF data are included in an open source software package called LAF-Fabric, a python package on Github [5]. A snapshot of the version used for this dataset is included in laf-fabric.zip. Off-line documentation and examples are included as well.
LAF-Fabric is bonus material. Its inclusion in this dataset serves two purposes: provenance and re-use.

**Provenance**
the source data is in the dataset, and LAF-Fabric contains the tool to convert the source data into LAF. LAF-Fabric is Open Source, so it is completely transparent how the LAF resource has been produced from more original sources.

**Re-use**
LAF-Fabric offers a platform by which researchers can explore and analyse the data. It is not a tool with predefined functions, but an application platform, aimed at humanities researchers that can program themselves of are working with programmers. The programming skill level needed to write your own LAF tasks on top of LAF-Fabric is moderate. Basic Python programming and an affinity with scientific computing in the IPython way [6]. IPython notebooks are documents that contain executable (python) code and documentation. LAF-Fabric is your lab to study the Hebrew Bible, and you can carry out and document your experiments in these IPython lab notebooks. You can also share these notebooks easily on the web.

When you unzip laf-fabric.zip you find the contents of a Github project, including documentation as used on the readthedocs documentation site. We refer to the documentation for information how to set up LAF-Fabric on your own system and use it. The documentation can be accessed by opening the following file in your browser: /laf-fabric/docs/_build/html/index.html.

In order to work with LAF-Fabric on the ETCBC4 data, you also need the data in compiled form. This dataset still contains the EMDROS version of the data, and you can also process the ETCBC4 data with the query tools of EMDROS (2).

### 3. LAF-Fabric Data

In order to get started easily with the ETCBC4 data under LAF-Fabric, you can download and unzip laf-fabric-data.zip in your home directory.

Alternatively, you could compute this data yourself. That is a good exercise in reproducibility. Here are instructions how you can do it on a unix-based system, such as Mac OS X or Linux. All steps are also possible on Windows, but you may have to install additional software and the command line instructions are different there.

#### Preparation:
- download and install the Emdros software[2], after this the command mql should work
- place sourcedata/etcbc4.mql.bz2 in a temporary directory (say ~/tmp)
- install LAF-Fabric (say in ~/laf-fabric)
- create a directory ~/laf-fabric-data/etcbc4 with subdirectories mql and config
- add to this config subdirectory the files main.cfg, Objects.txt, ObjectsFeatures.csv from sourcedata

#### Action:
```
cd ~/tmp
```
Hebrew Text Database in LAF
unpack (2 mins)

bunzip2 etcbc4.mql.bz2

import the mql database dump into an sqlite3 database (3 mins)

mql -b 3 etcbc4.mql
cp etcbc4 ~/laf-fabric-data/etcbc4/mql

use LAF-Fabric to generate LAF (10 mins)

cd ~/laf-fabric
python lf-convert.py --raw --validate --source etcbc4 --parts all

finally use LAF-Fabric, it will compile the data first (15 mins)

python lf-gallery.py full

4. Contents of the dataset

4.1 Overview
This dataset contains the data proper plus bonus materials. The data proper is the contents of the ETCBC4 Hebrew Text database, and the bonus materials consist of the tool LAF-Fabric and various intermediate representations of the data proper. The bonus materials serve to reproduce the generation of the data and to aid in the use of the data for analysis.

4.2 Listing
The dataset has the following folders and subfolders.

Data proper
- \textit{laf} - contains the data proper in LAF format. It consists of the following files
  - etcbc4.txt - the plain text of the Hebrew Bible in UNICODE
  - etcbc4.hdr - the LAF header (as prescribed by the LAF standard) (XML)
  - etcbc4.txt.hdr - the GrAF header (as prescribed by the LAF standard) (XML)
  - etcbc4_regions.xml - definition of the regions of the plain text as used by the remaining annotation files
  - etcbc4_sections.xml - all nodes, edges and annotations for sections (books, chapters, verses, half verses)
  - etcbc4_monads.xml - all nodes for the word objects (also known as monads), plus annotations specifying information about them that is particular to the Emdros database: id, monad numbers, object types
  - etcbc4_monads.lex.xml - all linguistic annotations at the word level
  - etcbc4_lingo.xml - all nodes for the linguistic objects (sentence, sentence_atom, clause, clause_atom, phrase, phrase_atom, subphrase), plus annotations specifying Emdros information about them
  - etcbc4_lingo.s.xml - all linguistic annotations at the sentence(_atom) level
  - etcbc4_lingo.c.xml - all linguistic annotations at the clause(_atom) level
  - etcbc4_lingo.p.xml - all linguistic annotations at the phrase level

urn:nbn:nl:ui:13-048i-71
- etcbc4.lingo.pa.xml - all linguistic annotations at the phrase_atom level
- etcbc4.lingo.sp.xml - all linguistic annotations at the subphrase level
- etcbc4.list - not part of the LAF resource. A convenience file listing for each monad (word): the monad number, the character position of the start of that word in etcbc4.txt, measured in unicode characters, the character position of the end of that word, and the character position of the end of the trailer of that word, the trailer being all the material after the word that does not belong to the following word (white space, punctuation and other marks). Tab-delimited plain text file.

• decl - contains the declarations involved in the LAF resource. It contains the following files
  - ft.xml - Declaration of the linguistic features used in the LAF resource. For each feature it contains a short description plus a hyperlink to ISOcat [7,8].
  - sft.xml - Declaration of the sectional features used in the LAF resource. For each feature it contains a short description plus a hyperlink to ISOcat [7,8].
  - db.xml - Declaration of the Emdros-specific features used in the LAF resource. These features are not registered with ISOcat [7].
  - graf-resource.xsd - XML schema to which laf/etcbc4.hdr conforms
  - graf-document.xsd - XML schema to which laf/etcbc4.txt.hdr conforms
  - graf-standoff.xsd - XML schema to which the laf/etcbc4_*.xml files conform
  - isofs_dcr.xsd - XML schema to which the feature declarations ft.xml, sft.xml, db.xml in decl conform. This is a customised version of the TEI module isofs.
  - dcr.xsd - schema needed to link the TEI isofs definitions with the Data Category Definitions as used by ISOcat [7].
  - xlink.xsd, xml.xsd, xml-isofs.xsd - public, generic, online schemas needed for XML validation. The inclusion of these files enables the off-line validation of the XML files in this dataset.

• sourcedata - contains intermediate representations of the ETCBC4 data, heavily compressed in bzzip2 format. There are also some configuration files. There are two data representations:
  - etcbc4.mql.bz2 - an Emdros database dump with unzipped size of 480 MB. Contains the complete contents of the Emdros database as produced by the ETCBC group. From this information the LAF data has been derived. However, in order to use this file, the Emdros software needs to be downloaded and installed. It is freely available from [2].
  - etcbc4.sql3.bz2 - an database file with unzipped size of 126 MB. It is the result of importing etcbc4.mql into an sqlite3 database. This is the source that LAF-Fabric operates on in order to produce the LAF version.
  - main.cfg - metadata of the ETCBC4 source in so far as needed for the conversion to LAF
  - Objects.txt - a listing of the object types in the Emdros data, and in what LAF files they end up
  - ObjectsFeatures.csv - specification of the Emdros features, with definitions for ISOcat [7] and identifiers in ISOcat. This is the basis to generate ft.xml and sft.xml in decl

• documentation - materials that help users to understand the meaning of the data in this dataset.
  Contains:
  - images - a folder containing several diagrams used in the description of this dataset on the jump-off page in EASY and in this very file
  - etcbc - documentation from within the ETCBC department. It describes in great detail the meaning of individual features and their possible values. This documentation is targeted at researchers within the ETCBC and researchers that work in close connection to them, and hence it contains many details about other representations of the data not present in this dataset. There are also slight discrepancies between the documentation and the actual data. Yet we trust that a dedicated user of this dataset will find here the answers to most of his questions as related to the interpretation of the data. Contains the following items:
    - qdf1.33-2014-03-26.pdf - a concise overview of nearly all the features that occur in the ETCBC4. This text refers to internal manual pages of the data creation workflow of the ETCBC. For that reason, all those manual pages will be included here.
  - man1, man3, man3hebrew, man3obl, man5, man5obl - these are folders with manpages from the ETCBC. Most of the material is not needed to understand the LAF data. Only the
Hebrew Text Database in LAF parts referred to by the document qdf1.33-2014-03-26.pdf maybe needed. Yet it offers a unique insight into the level of articulation and formalisation that the ETCBC has achieved in producing the data. All these man pages have been converted to pdf.

- **feature-doc** - exhaustive listing of all features and their values, with overview files. For every feature that exists in the data, there is a corresponding plain text file here that lists all its values, with the number of occurrences, ordered by decreasing prevalence. These files have been generated by LA-Fabric, using the module `etcbc.featuredoc` from the ETCBC4 data.

- **MQL-QuickRef.pdf** - Quick reference for writing MQL queries, including a concise listing of most features in the ETCBC4 data, consistent with the archived version of the data.

**Bonus material**

  - **dist** - packaged version of the python code in LAF-Fabric. Used to install LAF-Fabric on your system
  - **setup.py** - script to install LAF-Fabric within your python installation
  - **emdrros2laf** - the python package responsible for converting the Emdros data source of the ETCBC4 to LAF
  - **etcbc** - a collection of python modules that "know" the ETCBC4 data. They work on top of the `laf` package. By contrast, `laf` does not know anything about the specifics of the ETCBC4 data
  - **laf** - the core python package of LAF-Fabric. Understands LAF in a generic way, compiles it to an quick-load format, and supports analytic workflows
  - **example-data, example-output** - a toy LAF resource, used for (unit)-testing
  - **examples** - tutorial python scripts and notebooks, showing how LAF-Fabric can be used in the ETCBC4 context. Containing:
    - **gender.ipynb** - IPython notebook, plotting the frequencies of masculine and feminine words for each chapter in the Bible
    - **trees-r.ipynb** - IPython notebook, generating syntax trees for all sentences in the Bible
    - **mother.ipynb** - IPython notebook, showing syntactic dependency relations
    - **lingo.py** - python script that produces a representation of all linguistic objects in the Bible
    - **events.py** - python script that shows how to walk over the textual data and be triggered by the start and ending of linguistic objects.
    - **lf-convert.py** - top level python script to run an Emdros to LAF conversion
    - **lf-gallery.py** - python script to run a variety of tasks, mainly used to check whether a new installation of LAF-Fabric is well configured
    - **lf-unittest.py** - python script to run a variety of tests, mainly used to check whether a new version of LAF-Fabric does not have obvious bugs
    - **lf-patch.py** - python script to improve a beta version of the ETCBC4 Emdros dump. The file `sourcedata/etcbc4.mql.bz2` is the actual result of running this script on the Emdros dump that was obtained on Sunday 2014-07-13 from the Emdros database called `bhs4` at the server `jakob.wi.th.vu.nl`
    - **ship-data.sh** - shell script to export the data as LAF-Fabric needs it. Produces a distributable zip file
    - **LICENSE, MANIFEST, README.md** - files required by Github. The `README.md` just contains a logo and an hyperlink to the online documentation on readthedocs.
  - **laf-fabric-data.zip** - a compilation of the data such that LAF-Fabric can work with it right away. Contains:
    - **laf-fabric-data.cfg** - default configuration file for LAF-Fabric
    - **etcbc4** - compiled data of the ETCBC4, plus some extra annotations. Contains
      - **bin** - a bunch of files, all binary, which correspond to individual features in LAF

ETCBC4 8 of 9 urn:nbn:nl:ui:13-048i-71
• *config, decl, mql* - materials to perform the Emdros to LAF conversion
• *annotations* - extra annotation packages. LAF-Fabric can be instructed to combine the core annotations of the ETCBC with extra annotation packages. It is also possible to use LAF-Fabric to construct such packages out of additional data. Contains:
  • *testparticipants* - annotation package in LAF form, just to show the workflow of creating and using extra annotations
  • *px* - data coming from the internals of the ETCBC. This data has not yet made it into the official release, but is highly relevant for present research. It contains features at the paragraph level. It is added as an extra annotation package so that it can be used selectively
  • *px* - the raw data at the paragraph level, from which the extra annotation package *px* has been constructed.
  • *px_data* - Export of so-called px-data from deep within the ETCBC data creation workflow. Plain text with fixed-width columns.

References

(2) EMDROS. Text database engine for analyzed or annotated text. Open source software created by Petersen, Ulrik. Available from http://emdros.org.
(4) Graph Annotation Format: XML schemas for LAF. http://www.xces.org/ns/GrAF/1.0/.
(8) SHEBANQ data category selection at ISOcat [7], containing the registrations of the linguistic features used in the ETCBC4. https://catalog.clarin.eu/isocat/rest/dcs/739