Automated Semantic Tagging of the Göttingen Septuagint Apparatus

Tuukka Kauhanen and Hannu Kalavainen

SVTG Electronic Text

Septuaginta: Vetus Testamentum Graecum (SVTG) is a series of comprehensive critical editions of the Septuagint, published by the Göttingen Academy of Sciences. The main text of the Göttingen Septuagint editions is an eclectic text that presents a scholarly reconstruction of the oldest attainable Greek text of each book. The apparatus seeks to report all the meaningful textual variants in the Greek witnesses and, in addition, noteworthy readings of the daughter versions and other secondary evidence. For most volumes published so far, the text and the apparatus can be found in an electronic format in Bible software such as Accordance™ and Logos™. However, the modules in these software programs provide little more than the apparatus with the appropriate formatting. While it is of great advantage that an electronic text for the Göttingen apparatuses is available, its usability could be greatly enhanced with a *semantic tagging*: meta-information about the textual elements.

We are developing an Automated Semantic Tagging of the Göttingen Septuagint Apparatus (ASTAGS). It will consist of two entities. First, it includes a comprehensive chart or schema for such tagging that names the various parts of the apparatus and parses the information. Second, it incorporates an algorithm that can recognize the apparatus elements, process them according to the type of information they contain, and automatically generate the necessary tagging.

An XML Schema for the SVTG Apparatus

The tagging will be developed by using Extensible Markup Language (XML), which is a standardized set of rules for human and computer-readable document format. XML allows for structuring data in customizable, highly complex hierarchies. An XML schema is a formal description of the elements that may appear in an XML document. The most widely used guidelines for representing text in XML are those of the Text Encoding Initiative (TEI) consortium.[1] Its derivatives include, for example, the EpiDoc standard for tagging epigraphy in tablets and manuscripts.

Our own schema attempts to capture the information in a Göttingen Septuagint apparatus. All apparatus text is encapsulated as content in the elements. When necessary, metadata is captured in attribute values. The schema is designed specifically to capture the information in the already existing Göttingen apparatuses. It is not meant to be an optimal presentation of the data.[2] Many of the elements and attributes needed for our schema are already defined in the TEI guidelines.

[1] Text Encoding Initiative Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 2019, https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

[2] Once the apparatus has the semantic tagging, as proposed here, it will be easy to import the information in a database. Once the data is in the database, it can be printed again into apparatus format. The XML Schema for that printout can be considerably less complex.

Use Cases for a Semantically Tagged SVTG Apparatus

With the full semantic markup in the apparatus, users can easily locate any desired information, for instance:

- Find variants featuring any form of the word κινέω.
- Find variants attested by MS 72 or the manuscript group *O*.
- To get a quick overlook on attestation patterns, one can query for the number of readings against the main text shared by two witnesses.
- Parse the continuous text of a given witness using the information in the apparatus.

The benefits are obvious: instead of manually searching for the information, information can be gained instantly and with a high degree of reliability. The search results can be visualized in a more readable format than the highly condensed apparatus layout, and they can be used as raw data for statistical inspection of linguistic phenomena or manuscript relations.[3] The possibility of searching and visualizing the data will make the text-critical information more accessible to nonspecialists. An inexperienced reader of the apparatus will benefit from the possibility of seeing the implicit attestation for the lemmas and the manuscripts embedded in the sometimes elusive notation *reliqui* ("rel" or "rell"). The ongoing or forthcoming projects for the Septuagint editions could benefit from the possibility of validating their provisional text and apparatus with the SVTG tagging.

Automated Tagging

The algorithm that will automatically create the necessary tagging for apparatus content needs four input documents: (1) the main text of the SVTG book to be tagged; (2) the electronic text of the apparatus;[4] (3) the SVTG XML schema document; and (4) an XML skeleton file, which contains all possible branches of how the apparatus may be structured. The schema document is used in parallel with the XML skeleton. Together they provide the necessary information for data organization, such as the number and order of elements that can be nested inside other elements. The output produces a single XML document containing the fully tagged apparatus with all relevant metainformation.

The algorithm workflow is broken down to several subtasks: (1) *Prepass filtering* standardizes, for example, the use of dash-like symbols and asterisks. (2) *Tokenization* separates characters to divide the source text into its smallest pieces of information. (3) After tokenization, the algorithm tracks the XML skeleton to find a node that best matches any given token of source text. (4) Finally, the algorithm tags the apparatus content and creates the necessary attributes with values.

Future Development

We have presented an automated tagging system that will hierarchically order the information contained in various apparatus notations. The resulting XML documents can be used for various purposes, including as source files for a future electronic edition of the Göttingen Septuagint. Derivations of the XML Schema presented here can be used in presenting the apparatuses

[3] We are developing ways to analyze such data using the Coherence-Based Genealogical Method of the Institut für neutestamentliche Textforschung (INTF) in Münster.

[4] Naturally, the electronic texts are under a copyright and will be used only in accordance with the copyright statements.

of those Göttingen Septuagint projects that already have their data in a database: 2 Samuel, Kings, and the soon-to-be-started Psalms project. The algorithm is designed to be primarily data driven; that is, the markers for tokenization, rules for element nesting, and tests for valid content can be changed to match other needs of automatic tagging. The algorithm is scheduled to be working in 2021, and when ready it will be made openly available.

© Copyright *TC: A Journal of Biblical Textual Criticism*, 2020