## CONFERENCE REPORTS

### Workshop: Linked Data and Syriac Sources, Amsterdam, March 2018

RACHEL DRYDEN, UNIVERSITY OF CAMBRIDGE

Around 30 scholars from more than a dozen different countries met in Amsterdam in mid-March for two days of discussions and presentations on developments in digital humanities in Syriac language and literature.

*George Kiraz of the Beth Mardutho Institute presents an update on the SEDRA project.*

Participants were welcomed to the workshop and Amsterdam by Professor Joke van Saane, Vice-Dean of the Faculty of Theology at VU Amsterdam and Professor Wido van Peursen, the workshop organizer and host.

Following an overview of digital Syriac projects in the Netherlands (CALAP, Turgama, Polemics Visualized, Topic Visualizer for Syriac texts, LinkSyr: Linking Syriac Data

(CLARIAH), Linked Data and Syriac Sources, Electronic Peshitta Text, e-CSCO), Professor van Peursen explained the methodology behind the projects, which aim to produce more reliable versions of Syriac texts than are currently available.

Hannes Vlaardingerbroek (Leiden/Amsterdam), presented an overview of the LinkSyr project, which is using data in the form of tagged and untagged morphological terms from existing projects and materials and collating them into one dataset, with 160,000 items already tagged of what will eventually comprise more than one million terms. However, there is not enough data to train reliable HMM language models: existing tagging methods for Semitic languages, such as Hebrew and Arabic, use large corpora to train language models, which are not currently available for Syriac. Syromorph (BYU) claims high accuracy but is not yet compatible with the LinkSyr data. Mathias Coeckelbergs (Brussels and Leuven), discussed the nature of the data in more detail and longer-term plans, such as linking terms to the syriaca.org database, providing automatic reading tools for non-Syriac specialists, and more efficient search facilities. The dataset has some limits, as the method works by recognizing surface forms, which can have multiple translations. Eventually, it is hoped the classification of URIs will be more data-driven and searchable for specific collections of texts.

Following this, George Kiraz (Beth Mardutho) described the process for converting Syriac lexicons from image to text files, creating an on-line, searchable dictionary, as part of the SEDRA project. While SEDRA was designed specifically for Syriac, the project has the technical capability to be expanded to include other Semitic languages and is looking for funding to achieve this longer-term aim.

David A. Michelson (Vanderbilt) provided an update on the syriaca.org project, which has minted URIs for places, persons, primary source texts and citations (bibliographic items), and published them online. URIs relating to factoids (events), ontology (keyword classification) and manuscripts are available as raw data. The project is currently looking for someone to do the same for artifacts. Daniel L. Schwartz (Texas A&M) demonstrated the various features the site offers.

James Walters (Oxford-BYU Syriac Corpus) talked participants through the structure and functions of the Oxford-BYU website and the new edition of *Hugoye*, to be launched this summer.

*James Walters showcases the Oxford-BYU Syriac Corpus*

Daniel Stökl ben Ezra (EPEH Paris), demonstrated the interface and search functions offered by the ThALES lectionary database, which includes material in Syriac and Arabic.

In the afternoon, a number of working groups discussed Lexicography, Named Entities, Liturgy, Text Corpus Creation, Scholars’ Needs and Interests, How to Bridge Syriac Linked Data and the Syriac Community and Linking to Other

Traditions such as Arabic and Ethiopic, and brainstormed recommendations and suggestions for future projects.

The workshop continued on the second day with further project presentations and updates on the Galen Palimpsest and GREgORI projects and Hun@yannet. Natalia Smelova (Manchester) began by presenting the results of research on the Syriac Galen Palimpsest, which has been fully digitized using multispectral technology, with the resulting images available free online. The tools and software developed as a part of the project will also be made available to individuals and institutions dealing with multi-layered texts.

Grigory Kessel (Austrian Academy of Sciences/ Manchester) and Slavomír Čéplö (Austrian Academy of Sciences), showcased the ERC-funded Hun@yannet database, consisting of a Greek-Syriac-Arabic corpus of Classical scientific and philosophical literature. Hun@yannet will allow for a comprehensive comparison of Syriac and Arabic translations of Greek texts, through lexicographical analysis, for which a dedicated research tool is being developed. It will be capable of handling vertical texts in UTF-8 encoding, (with manual parallel alignment by sentence or clause units), via a TEI to HTML transformation reading interface. While there remain some issues with parallel corpora, search by REGEX of 15 texts with 50 translations in Syriac and Arabic is already possible.

Bastien Kindt (Leuven), also presented the GREgORI project, a free, multi-lingual corpus and fully lemmatized concordance, with parts of speech tagged, although Syriac is still to be completed.

In the afternoon, the working groups reported their findings from the previous day with the Lexicography working group stressing the need for both archival/reading and linguistic analysis versions of texts. A general consensus was reached that a TEI format of a Syriac text was useful for archival, canonical, and human readable formats and that texts would then need to be "unwrapped" or converted for use in text analysis tools such as Text Fabric. As Text Fabric uses the word as the basic entity, more work would be needed to handle idiomatic phrases, compound nouns and names. It was suggested that Syriaca.org and SEDRA might ensure that two types of tagging could be facilitated by dividing the NER into two stages: the identification of proper nouns as simply being proper nouns and the identification of specific names and concepts for persons, places, etc.

The Named Entities group discussed the desirability for current projects to be structured so as to permit a machine-generated model of Syriac scholarship in the future, which could be created by the analysis of named entities and their links; methods that diverse projects could use link to named entities, through using URIs minted by Syriaca.org, with the resulting benefit of creating a collective body of linked data and making various diverse projects discoverable through Syriaca.org. For projects linking to Syriaca.org URIs, the preferred method was to link to permanent Syriaca.org URIs (e.g. http://syriaca.org/place/78). Syriaca.org agreed to provide content negotiation for this purpose and was willing to link to projects in any way desired e.g. permanent URIs, RESTful API queries. The group also discussed how to generate a cloud of RDF data related to Syriac from various projects, concluding that more discussion was needed to work out how to create, host and query this.

These two groups concluded by agreeing on a protocol for connecting various digital texts online in various formats, considering it desirable for all projects to use CTS/DTS URNs. Syriaca.org offered to serve as a catalogue for various projects and conservator of standards. Through the New Handbook of Syriac Literature, Syriaca.org could also provide a standard form for the first half of a CTS URN, language, corpus, work family, work number. The second half, the local ids, would then be published by each individual project. The resulting URNs would then be combined, crosswalked, archived and catalogued by Syriaca.org. The end result would be to establish a standard method for citation and api calls to digital Syriac texts across all projects.

The Liturgy working group members agreed future cooperation between LinkSyr and ThALES, either through the exchange and/or linking of data. A liturgical app is being developed by George Kiraz, with Srophe potentially hosting it. The Text Corpus group looked at how synergy could best be established, recommending joining forces to avoid the same work being done twice and also suggested potential collaboration with commercial partners. It was noted that many projects have encoding at the word level, which is sufficient for a large corpus with a broad public audience, but linguists and exegetes may required more layers of analysis. The group also agreed to form a standing committee to follow-up on discussions at the workshop, identify issues and make their materials available on GitHub and hold a regular Google hangout seminar for Syriac studies.

The group on Other Traditions discussed the fact that, except for the connections with Greek loan words and the Bible, this area is otherwise generally neglected. The group felt the most important traditions to link to were, qur'ānic Arabic, Aramaic, medieval Arabic/Greek philosophy and non-Christian sources, and that there was a need for more inter-disciplinary tools. More video-based help guides instead of text-based explanations of how to make use of the various aspects of the databases available, were also requested and it was felt there was a need to somehow bridge the gap between those specialised in digital humanities and those wanting to develop a digital aspect to their research projects.

The working group looking at How to Bridge Developments in the Digital Humanities with the Syriac Community, felt that a focus on language, liturgy, tradition and history would be beneficial to members of the Syriac Orthodox community, and also provide a means of strengthening their sense of identity and heritage. Liturgical and Bible readings with accurate vocalization and options for comparison (e.g. how to deal with those Sundays for which there are no readings), and musical aspects were also considered desirable. They also supported the translation of lectionaries into Turoyo with a format comparable to the ThALES database with additional audio/visual recordings, which would be particularly useful for a non-academic audience, who could then actually *hear* the word of God. While there are large amounts of audio/visual resources available online, it was felt that it would be useful to somehow collect and organise these and link specific topics with other online resources. Regarding the public impact aspect of the LinkSyr project, it was agreed there was a need to create more awareness of what tools the University can offer the Community, perhaps through offering some sort of training about this. The University and Community could also facilitate more exchange in the form of visits, lectures etc. Addressing young people was considered one of the biggest challenges; while there is some material available for teenagers, there is not much for younger children. Since teenagers tend not to read books, digital tools that can be used on a tablet/smartphone, such as interactive apps, games, etc. were also recommended. Regarding the question of how to translate information for children, it was felt the NT2 method (*Dutch as a second language*) was considered a good guideline. Regarding the financial aspect of community interaction, while attracting more people was obviously a positive, it was felt important that they be paid for their contributions (e.g. transcribing manuscripts), either through financial recompense or, in the case of university students, perhaps in exchange for skills/experience. Crowdsourcing was also discussed as a possible means of obtaining knowledge/skills free of charge.

The Scholars' Interests group asked whether current projects really meet the needs of scholars and if not, what desiderata and caveats these projects should take into account, in order to develop things that are really useful. Amongst the issues they outlined were; the use of the techniques under discussion to make stylistic analyses or characterizations of texts, to identify authors, or compare texts and translations; how to publish texts, which are both in print with a publisher and in a digital format in the open domain, and how to refer to digital texts, that in time may change, or otherwise remain fluid in their presentation. It was felt that audiences could be successfully reached by means other than promotion e.g. by providing interfaces in Arabic, Persian and Turkish, and that there should be further campaigns for the implementation of Syriac word processing; the submission of additional characters to the Unicode Consortium; an update of the Meltho fonts; an approach made to Apple about the implementation of Syriac, and the development of a cross-platform Syriac language pack for LibreOffice.

The workshop provided a rare opportunity for face-to-face discussion and exchange amongst scholars working with Syriac in a variety of fields and it is to be hoped that the connections that were made at the workshop continue to develop to the benefit of current and future projects.