Coding and Encoding: Towards a New Approach to the Study of Syriac and Arabic Translations of Greek Scientific and Philosophical Texts*

Rüdiger Arnzen, Ruhr University Bochum,
Yury Arzhanov, Nicolás Bamballi, Slavomír Čéplö, and
Grigory Kessel, Austrian Academy of Sciences, Vienna

This essay presents the ERC project 'Transmission of Classical Scientific and Philosophical Literature from Greek into Syriac and Arabic' (HUNAYNNET) based at the Institute for Medieval Research of the Austrian Academy of Sciences. The main research question leading the project addresses the contribution of the Syriac tradition in the transfer of Greek scientific literature to the Arabic-speaking world. To fulfill this goal the project is going to provide digital editions of the Syriac and Arabic versions and tools for linguistic corpus-based analysis. The digital Greek–Syriac–Arabic corpus will offer a novel approach for research into the translation techniques and in the history of the transmission of classical Greek literature in Late Antiquity and the Middle Ages.

Introduction

Between the sixth and the tenth centuries CE, a vast body of medical, mathematical, philosophical and other scientific and technical texts were translated from the Greek into Syriac and Arabic. Apart from the social, cultural and historical aspects of the translation activity (sometimes called 'translation movements') as such, these translations contain valuable information for a number of linguistic and philological issues, to mention but a few:

- the development of the vocabulary, syntax and scientific terminology of Classical Syriac and Classical and Middle Arabic;
- the history and development of the translation movements from Greek into Syriac and Arabic;
- the diachronic semantic shifts from Classical to Middle and Byzantine Greek as evidenced by Syriac and Arabic interpretations of the Greek works;
- the critical establishment of the Greek texts, which are often preserved in manuscripts that are much younger than their Syriac and Arabic translations.

Whereas it is well known that mediaeval Europe received ancient Greek science and philosophy through Latin translations—many of which were pre-

* The research is being supported by the European Research Council under Grant Agreement no. 679083 (ERC Starting Grant 2016–2021, PI Grigory Kessel).

pared on the basis of the Arabic and Hebrew versions of these works—, it is less well known that the Arabic translations from the Greek produced between the eighth to the tenth centuries were sometimes directly based on a Syriac intermediary or depended indirectly on available Syriac translation.

The impact of the Syriac tradition upon the Arabic translations has been acknowledged but not thoroughly explored. Compared with the extant body of Graeco-Arabic translation literature, the available Graeco-Syriac translations constitute just a small fraction. Nevertheless, it is our firm conviction that the very availability of at least that relatively small group of texts urgently requires comparative examination. A sound study of those translations can inform us about the contribution of the Syriac tradition not only on the relevant Arabic translations but also on other Graeco-Arabic texts for which a Syriac version is wanting.

The present project aims to contribute to the study of that transmission process and, more importantly, to trace the role of the Syriac tradition through the creation of a digital trilingual (Greek, Syriac, Arabic) text corpus. Named after Ḥunayn ibn Isḥāq (*c.*808–877), arguably the most prominent figure in the history of scientific translations from Greek into Syriac and Arabic, the digital corpus HUNAYNNET is going to serve as an open-access platform for research in the transmission of Greek scientific and philosophical literature into Syriac and Arabic. It is the first attempt to present together the Syriac and Arabic translations along with the preserved Greek originals. The digital corpus is intended not only for specialists working on Greek, Syriac, and Arabic texts, but also for a broad spectrum of scholars and students interested in the history of culture, philosophy and translation studies, more generally.

1. Textual Basis of the Greek-Syriac-Arabic Corpus HUNAYNNET

The core group of the texts consists of those scientific and philosophical works that are available in all three languages, Greek, Syriac, and Arabic. We include straightforward translations (preserved in both complete form and in fragments) but exclude quotations embedded in other works. One of the things that makes the study fascinating is that for some texts there is more than one translation in a given target language. Table 1 below presents the principal sources of the corpus.

2. HUNAYNNET *corpus: Introduction*

Corpus linguistics in general and parallel corpora in particular are an invaluable aid for studying the translation process that may cover such aspects as translation techniques and style. The use of corpora in recent decades revolutionized the study of languages. Being first developed for modern languages the corpus-based approach has only recently been applied also to ancient languages and the major discoveries are yet to be made.

Table 1. Texts included in the Greek-Syriac-Arabic corpus HUNAYNNET.

Author, title	**Translations**
Porphyry, *Isagoge*	Syriac (two versions) Arabic
Aristotle, *Categoriae*	Syriac (three versions) Arabic
Aristotle, *De interpretatione*	Syriac (two versions) Arabic (two versions)
Aristotle, *Analytica priora*	Syriac (two versions) Arabic
Pseudo-Aristotle, *De mundo*	Syriac Arabic (three versions)
Galen, *Ars medica* (fragment)	Syriac Arabic
Galen, *De alimentorum facultatibus* (fragment)	Syriac Arabic
Galen, *De simplicium medicamentorum temperamentis ac facultatibus vi-viii*	Syriac Arabic (two versions)
Hippocrates, *Aphorismi*	Syriac Arabic
Hippocrates, *Prognosticon*	Syriac Arabic
Ptolemy, *Tetrabiblos*	Syriac Arabic

Even though we are fully aware of the advantages offered by corpus linguistics for a study of monolingual and multilingual corpora, our approach to the corpus needs to take into consideration the specific character of the material we deal with. Namely, the corpus is made of the ancient Greek texts that belong to classical antiquity. Each of them had its own transmission history and got to be translated into Syriac and Arabic. It is an axiom of classical philology that there is no ancient text that reaches us in its authentic and original form. For this very reason, philologists painstakingly labour to provide a reliable text paying minute attention to any detail. The optimal presentation of the philological work remains an edition.

To put it somewhat differently, the texts can be presented digitally either in the form of a digital edition or as a linguistic (parallel) corpus, whereas the former is 'text-driven' the latter is 'data-driven'. Both solutions have their own pros and cons, and offer different functionality and usage scenarios.

To have the best of both worlds, we decided to provide two platforms for data visualization in order to satisfy the requirements of both groups of poten-

tial users as best as possible. At the same time, however, both solutions are to be integrated as closely and as seamlessly as possible by taking advantage of all capabilities of the selected technical implementation. The two solutions are referred to as the *reading interface* and *parallel corpus* and in what follows, we offer a brief description of both.

3 Reading interface

3.1 General remarks

The primary purpose of the reading interface is to provide the user with a way to simply read the texts, whether in parallel or separately, in a way that is most suitable for digital consumption, while also preserving all the relevant scholarly information and providing access to tools, which aid them in their research.

3.2 Implementation

The reading interface is implemented as a minimalist cross-operative system and cross-browser static website with minimal AJAX elements, all built on open standards (XHTML, CSS/Flexbox and JQuery). This ensures maximum usability in the present, as well as the project's longevity: the entire reading interface can be downloaded and copied anywhere or even run offline as opposed to a solution with a database backend. The combination of Flexbox and JQuery, while providing the necessary functionality, is much more resilient to the ever-present problem of inter-version obsolescence than, say, Angular which serves as the framework for the excellent and powerful EVT platform, an open-source tool designed for publishing digital editions using TEI XML.[1]

In terms of process, the texts edited and aligned in *Classical Text Editor* (CTE, created by Stefan Hagel from the Austrian Academy of Sciences, see below) are exported to TEI XML using the functionality built into CTE. The TEI XML files are then minimally post-processed (largely to ensure compliance with the TEI standard) and, using XSLT transformation, converted to HTML. The HTML files are then read into the reading interface using AJAX and arranged visually using CSS.

In the final version, the reading interface will include (1) permanent links with Uniform Resource Names (URN), (2) downloadable TEI compliant XML files with CSS stylesheets, (3) downloadable embeddable HMTL files and viewable HTML files, (4) downloadable PDF files, and (5) basic text search.

1 Edition Visualization Technology, retrieved on 3 August 2018: <http://evt.labcd.unipi.it/>.

It should be noted that the text search implemented in the reading interface will truly be only rudimentary; the bulk of the search capabilities will be outsourced to the parallel corpus.

3.3 Enrichment of data

The reading interface will provide two major forms of enrichment, which aim to aid the user in reading and studying the texts:

a. Sentence synchronization.

This is added to the texts using the respective functionality in CTE (see below, section 5.a) and implemented in the interface as (1) sequential number (per chapter or per text, selectable by user), and (2) as a hover-initiated highlight of the synchronized sentence and its equivalents in all displayed texts.

b. Lexical and morphological information.

This functionality is implemented in a manner similar to that used in *Thesaurus Linguae Graecae*:[2] each word in each version is responsive, and upon calling, a selection of sources with lexical and morphological information is displayed where the user can select which of the sources they wish to access. This information is retrieved by API from existing publicly available online lexicons—provisionally Perseus for Greek, ElixirFM[3] for Arabic, and Syriac Electronic Data Research Archive dictionary (SEDRA)[4] for Syriac—as well as from the *Glossarium Graeco-Arabicum*.[5]

In addition to linking to lexical and morphological information, functionality will also be provided to search for the selected word in the parallel corpus and continue the analytical work there. This is the first major step towards integrating the two platforms, namely the reading interface and the parallel corpus.

4 Parallel corpus

4.1 General remarks

The purpose of the parallel corpus is to provide an interface to conduct standard corpus research involving, for example, collocation extraction, lexicographic analysis, n-gram analysis and parallel lexical analysis.

4.2 Implementation

The texts are imported into an open-source corpus management system, for the moment the *NoSketch Eingine* (updated in the future with a customized

2 *TLG*, retrieved on 3 August 2018: <http://stephanus.tlg.uci.edu/>.

3 *ElixirFM*, retrieved on 3 August 2018: <http://quest.ms.mff.cuni.cz/cgi-bin/elixir/index.fcgi>.

4 *SEDRA*, retrieved on 3 August 2018: <https://sedra.bethmardutho.org/>.

5 *GlossGA*, retrieved on 3 August 2018: <http://telota.bbaw.de/glossga/>.

KonText interface)[6] as parallel corpus; this is made possible by the sentence synchronization encoded in the files. Both these tools use the identical vertical text format as input in encoding the corpus and the vertical text is also provided as a downloadable resource.

4.3 Enrichment of data

The texts will be enriched with the standard set of linguistic annotation, i.e. (1) tokenization; (2) lemmatization; (3) part-of-speech tagging; (4) morphological analysis.

In the cases of Greek and Arabic, a number of tools exist to provide a reasonably accurate annotation for all of the above. These are, for example, for Greek, Morpheus[7] and the Classical Language Toolkit[8] and, for Arabic, the Stanford CoreNLP,[9] the aforementioned ElixirFM and Farasa.[10] With Syriac, however, what was said above in reference to lexical resources is doubly true of natural language processing tools, despite some recent progress.[11] One of the ancillary goals of the project is to use the project data and the experience of the project members to expand the existing range of computational tools for the processing of Syriac. Most annotation shall be carried out in CTE with additional post-processing directly in XML (see below).

Finally, a functionality will be provided to link the results of the search in the parallel corpus back to the reading interface using the identification of aligned synchronization units, so that a result of the query can be immediately consulted in the reading interface.

5 Making of the corpus

The primary goal of the project is to create an aligned multilingual corpus. The process entails (a) preparation of digital editions and (b) parallel alignment.

Both tasks are accomplished using the CTE software, which has been widely used for the preparation of critical editions in the field of classical philology and beyond and as such it provides a number of facilities required

6 *Czech National Corpus*, retrieved on 3 August 2018: <https://kontext.korpus.cz/first_form>.

7 *Perseus Digital Library*, retrieved on 3 August 2018: <http://www.perseus.tufts.edu/hopper/morph?lang=greek>.

8 *Classical Language Toolkit*, retrieved on 3 August 2018: <https://github.com/cltk/>, <DOI:10.5281/zenodo.593336>.

9 *The Stanford Natural Language Processing Group*, retrieved on 3 August 2018: <https://nlp.stanford.edu/projects/arabic.shtml>.

10 *Farasa*, retrieved on 3 August 2018: <http://qatsdemo.cloudapp.net/farasa/>.

11 E.g. Kindt et al. 2018 and the *LinkSyr* (Linking Syriac Data) project, retrieved on 3 August 2018: <https://github.com/ETCBC/linksyr>.

for critical text editing (e.g. unlimited number of critical apparatus, automatic insertion of line and chapter numbers, marginal references, freely definable sigla, etc.). Additionally, it also offers two crucial advantages for the purposes of our project: first, it enables the export of the text, the apparatus, any other notes and structural division markers into TEI XML, an open standard for the representation of texts in digital form.[12] Secondly and crucially, CTE allows easy and accurate handling of languages with RTL direction such as Syriac and Arabic, something that is generally not achievable by means of standard XML editors. All these and additional features (such as the export of a print-ready critical edition in PDF) make CTE the perfect tool for the kind of work entailed in the HUNAYNNET project.

In detail, the two aforementioned steps are carried out as follows:

a. Preparation of digital editions.

We have adopted the following general policy for the retrieval of the texts: to use available editions or manuscripts if the text is not edited. In the simplest scenario, we use an available edition and collate it against the manuscript(s). If there is more than one edition and none of those is superior, we use both editions which are collated against the manuscript(s) and against each other. In both cases, the errors are corrected and the editorial interventions are documented by means of an apparatus. For the unedited texts, we prepare minor editions based on a selected group of witnesses. Hence, we are going to offer improved editions for already edited texts and the very first editions for those that have never been edited.[13] To achieve the uniformity of the text corpus and thereby to guarantee better search results, all the texts are being normalized following established editorial guidelines (abbreviations resolved, homographs disambiguated; shaddas, hamzas and other orthographic features of Arabic supplied; all *seyame* in Syriac included, etc.). All the texts are provided with structural information referencing to the standard editions of the Greek texts (e.g. page, column, and line of Bekker's edition for the *Corpus Aristotelicum*), editions of the translations and manuscript witnesses.

b. Parallel alignment.

This is achieved by annotating the text with boundaries of minimally extensive syntactic and semantic units, roughly equivalent to simple or compound sentences, henceforth referred to as 'synchronization (or sync) units'. The

12 TEI P5: Guidelines for Electronic Text Encoding and Interchange, retrieved on 9 August 2018: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/>.

13 We would like to thank here the project 'The Syriac Galen Palimpsest: Galen's On Simple Drugs and the Recovery of Lost Texts through Sophisticated Imaging Techniques' (University of Manchester) that has kindly provided us with a transcription of the Syriac version of Galen's *De simplicium medicamentorum facultatibus.*

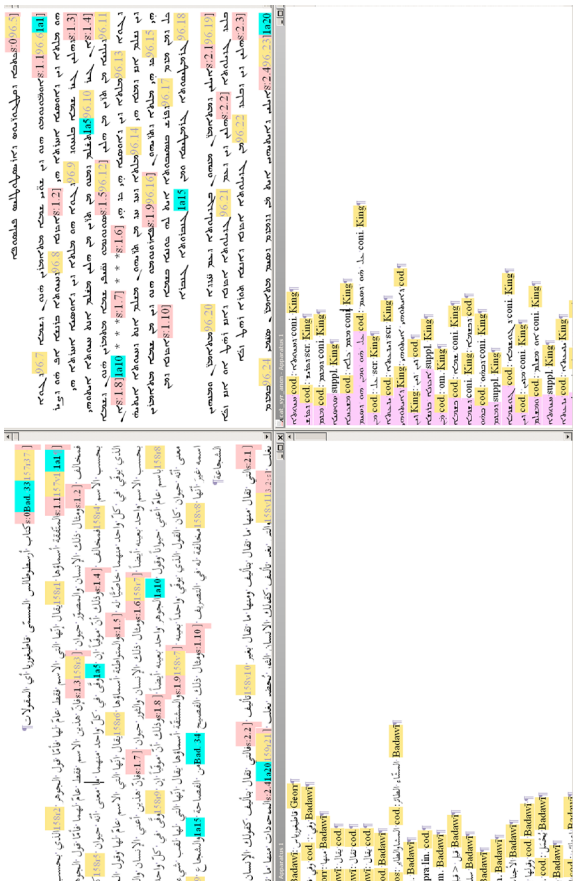

Fig. 1 Example of text encoding (Aristotle, *Categoriae*) in CTE

three languages have different structures and the translations do not always follow a *verbum e verbo* approach. In order to make the comparison more convenient, we present the texts aligned at the level of (sub- or coordinate) clauses or sentences.

In practical terms, this is done by introducing a so-called sync (or synchronization) mark, a special symbol in CTE, at the beginning of each of these minimal syntactic and semantic units. In the XML export, this symbol is converted into the TEI element <anchor> with a sync attribute;[14] this then allows to match the sync units as necessary.

In this context, the Greek text serves as the immutable fundament and so the division of sync units—defined broadly in semantic and syntactic terms—follows that of the Greek text and its semantic division. The automated processing necessary for the creation of parallel linguistic corpora (i.e. tokenization, lemmatization, and sentence alignment, see below) requires that the number of synchronization units (a term preferred to 'sentence' and thus used henceforth in the context of parallel alignment) be the same in all texts. Consequently, in cases where the translation lacks a passage, an empty synchronization unit is added to the translation; in cases where the translation adds a passage, the aforementioned principle of immutability of the Greek original requires that the added text be joined with the preceding synchronization unit, resulting in a translation that may be much longer than the original.

Fig. 1 illustrates a typical example of encoded Greek, Syriac and Arabic texts and the critical apparatus for two versions. In the main text, the sentence synchronization marks are highlighted in pink, chapter identifiers (Bekker numbers in the Greek, folio and line numbers of the manuscript Paris, Bibliothèque nationale de France, Ar. 2346 in the Arabic, and page and line numbers of King's edition of the Anonymous Syriac version) are in yellow, and Bekker numbers (every fifth line) in the Arabic and Syriac are in turquoise.

And finally, in addition to aligning the Greek original and the Arabic and Syriac translations, the project also envisages the addition of an English translation to those texts that have been translated into English and for which an out-of-copyright English translation exists. This would allow for the project output to be used as a tool for the study of the languages and the texts involved and thus expand its usability beyond scholarly study to classroom and self-study use.

References

Kindt, B., J.-Cl. Haelewyck, A. Schmidt, and N. Atas 2018. 'La concordance bilingue grecque-syriaque des Discours de Grégoire de Nazianze', *BABELAO*, 7 (2018), 51–80.

14 <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SASY>.