Greek into Arabic
ERC *Ideas* Advanced Grant 249431
Workshop in Pisa, October 3-5, 2011

Andrea Bozzi

A Philological Computational System for Graeco-Arabic Texts Editing and Processing: Methodology and Samples


Andrea Bozzi and Cristina D'Ancona

Preliminary remarks

I would like to present an open source web application for scholarly editing and processing digital documents to be disseminated either in paper format or through on-line servers. It consists in a computational philology system taking into particular account the joint editorial activities to be performed by scholars in a collaborative way. It is subdivided into two parts, if necessary interacting with each other: the former makes it possible to perform a number of special indexes and concordances on texts and on critical apparatuses, when available; the latter is more oriented towards the computer assisted textual criticism.

This application **(Fig. 1)** is nowadays far from being a finished product, but it can be considered reliable because it is already used in a number of national and international projects.

As far as the first ones, I quote: - the ***Electronic edition of the manuscripts of F. De Saussure***; - the management of the ***Historical Archive of the "Pontificia Università Gregoriana"*** in Rome; - the Archive of ***Galileo Galilei's Opera Omnia***). As far as the international projects, I would like to focus your attention to "Greek into Arabic", the ERC Advanced Grant 249431 coordinated by Cristina D'Ancona.

Its use on so many different projects demonstrates its flexibility, because it has been designed at supporting many different sectors, from epigraphy to papyrology, from classical to medieval philology, from the philology of modern and contemporary works to the philology of ancient printed books.

My contribution today is concerned with only two main components, while some Natural Language Processing (NLP) modules under development for the projects mentioned above are not outlined. I would like only to quote here, without details or specific slides: - some interesting experiments carried out with the use of a special module which allows scholars to edit lexical entries using a specific formalism, following the positive results obtained for the De Saussure's project. It is based on the theory of computational lexicons and makes it possible to enhance the semantic information retrieval operations; - I am not going to speak about the modules for automatic morphological processing and lemmatization of texts already performed for Arabic wordforms. All samples in this presentation are related to the Arabic text of the so-called pseudo-*Theology* of Aristotle realized within the al-Kindi circle.

The two components I intend to examine in detail are referred to two activities strictly connected: - one special indexation system; - and a text criticism module aimed at browsing digital images of collated witnesses and allowing scholars to find variants.

Indexation component.

The creation of indices follows rules which depend on many different phaenomena, for example the alphabet in which the texts are edited, and the aims of the editorial project. For these reasons, the computational system should be designed so as to include a series of essential indexation tools, and an additional number of modules meeting the search requirements and the different types of texts. I will simply describe the case on which we are currently working: each linguistic form must be accompanied by two parallel contexts referred to two works, one of which is the translation of the other. Parallelism is essential for two reasons: firstly, it is necessary to check any semantic differences between the two texts and, secondly, to highlight whether and to what extent there are particular form types responsible for this divergence, so as to be able to produce a "contrastive" lexicon. The examples are drawn from the work – still in course – within the framework of the project "Greek into Arabic", where we have to process and compare the Greek text of Plotinus' *Enneads* (**fig. 2**) with its Arabic translation (**fig. 3**). The translation often deviates from the original not only because it modifies the sequence of the chapters, but also because it re-processes the concepts, generally of an abstract type, on the basis of a still relatively unknown hermeneutic-interpretative structure which is one of the main goals of the research. For these reasons, the computational tool has been studied in order to produce results which make it easier the work of Arabists and lexicographers working in Pisa and Bochum.
The two texts have been organized in corresponding numbered pericopes (**fig. 4** and **fig. 5**: the progressive number is visible on the top of each pericope).
(fig. 6) shows the xml mark-up system of a given pericope.
(fig. 7) They have been stored so that it is possible to read the Greek text in its original sequence and the corresponding Arabic version. In the green box it is immediately evident that the Arabic version expanded the original text. In the same slide, we notice that there is a jump in the Arabic text, highlighted in red. It is really easy to find where this part is located, and to what section of the original Greek it corresponds: in fact (**fig. 8**), browsing now through the Arabic text in its natural original flow, it is possible to read the Greek text corresponding to the lines jumped in the Arabic version. (**fig. 9**) The line-jumping (i.e. interruption of the sequence) previously observed is now clear, and can be analyzed. The Arab adaptor has broken up the Greek text recomposing (e.g. expanding) and re-assembling it.
The indexation system produces two indices, Greek and Arabic, which can be queried using the Arabic term as search key (**fig. 10**), with the production of the results in the

form of Arabic-Greek pericopes, or also using the Greek term (**fig. 11**), a procedure which produces the Greek-Arabic pericopes, in which both Arabic and Greek forms are attested.

(**fig. 12**) The scholar has at his disposal a robust, formal, flexible and expressive tool designed to analyze the text, and to query the electronic pericope using logical/Boolean expressions. When querying the system, the user can see the pericopes attesting the requested form in Greek, or (i.e. inclusive "or") a possible translation in Arabic. What follows is the possibility of performing a contrastive and detailed analysis of the lexicon, displaying the different translations according to the grade of emphasis given by the translator/adaptor.

(**fig. 13**) Users are allowed to include their personal notes and/or comments to the results produced by the system.

I think that what has been shown so far might demonstrate the importance of this procedure where we can see how the Arabic text has expanded part of the original Greek work with autonomous insertions and where the contexts are provided, so that the underlying reasons can be understood.

Text criticism component

(**fig. 15**) The main functions of the computational philology system are shown in this slide, simulating a page of the graphical interface. In particular: variants can be easily found comparing the text of the document considered as the base of collation (W0) and, each time, the images and/or transcriptions (if available) of further witnesses (W1, W2, W3, etc.). The slide shows the way in which the system stores the information retrieved in the transcription files and in those related to digital representation of the documents: as for the former, it records the page (or sheet) parameters, line and position of each word in the line; as for the latter (**fig. 16**) it stores the coordinates of each portion of image in which a variant reading or error has been selected by the user. This aspect could be underestimated, but we think it is extremely useful, because the system stores all the necessary information to display the "graphical" concordances of a specific variant form. Such a result might lead to rethinking and suggesting alternative readings, especially in the case of words which are poorly attested, or in the case of terms for which the copyist has expressed some doubts.

(**fig. 17**) A field is made available to the philologist, who can add a comment on a single variant reading.

Finally, I would like to present a particular type of annotation – the semantic-ontological annotation - that the system allows to perform on texts and images.

An ontological system is based on contents classification, according to a predefined logical-conceptual order, which is shared by the community of the scholars involved. This need has strongly emerged within the framework of the electronic edition of

Ferdinand de Saussure's manuscripts, and we could suggest the same strategy also for "Greek into Arabic".

(**fig. 18**) As an example, I would like to show the image of a manuscript of Tehran, Madrasa-i Sipāhsālār 1296. This slide allows me to briefly outline and clarify which needs are emerging in terms of the organization of the textual material, and which solutions we can offer by our application. The study of the documents by the scholars, and the proposal made by the information system designers to organize such documents according to a logical scheme, have made it possible to proceed gradually. The scheme, originally empty, becomes even more rich with explicit classes, subclasses and relations, organizing the conceptual elements in a hierarchical net-like structure at the top of which the more generic and omni-comprehensive class is placed. Therefore, the indexation procedures are much more expressive and flexible as they can produce results which are not only consistent with the restrictions defined by the researcher, but also derive from the activities of "reasoning" typical of the ontological systems. It should also be pointed out that it is often essential to comparatively consult those parts of the documents which are dealing with the same or with very similar and conceptually-connected subjects. In this respect, no data cataloguing system would make possible a conceptually-oriented navigation of textual material. **(fig. 19**) Annotations can be organized at different levels: (a) technical-structural; (b) content (re-transcribed in clearly readable Arabic); (c) literal translation; (d) external intervention. The figure highlights the portions of text containing the technical annotations in black; the transcriptions in red; the translations in blue; and the free comments made by the scholar in green.

(fig. 20) Here the annotation relative to the rotated word in red is shown.

Conclusion

(**fig. 21**) In order to develop a complex system like the one described above, it was necessary to distribute the different tasks: the Institute for Computational Linguistics in Pisa, general technological coordinator of the project, has made available the components of indexation and concordances, the data structure of the critical apparatus, the morphological analysis and the formalized description of the entries following the computational lexicons theory. The Fondazione Rinascimento Digitale in Florence is responsible for the ambience design related to the presentation of data on the web, and to the management of the ontological components. Important positive results have been achieved thanks to some projects that have considered this application very innovative and useful. I hope this presentation has aroused your interest and the wish to join us for further experimentation.

Andrea Bozzi
(Director of the ILC/CNR, Pisa)

European Research Council