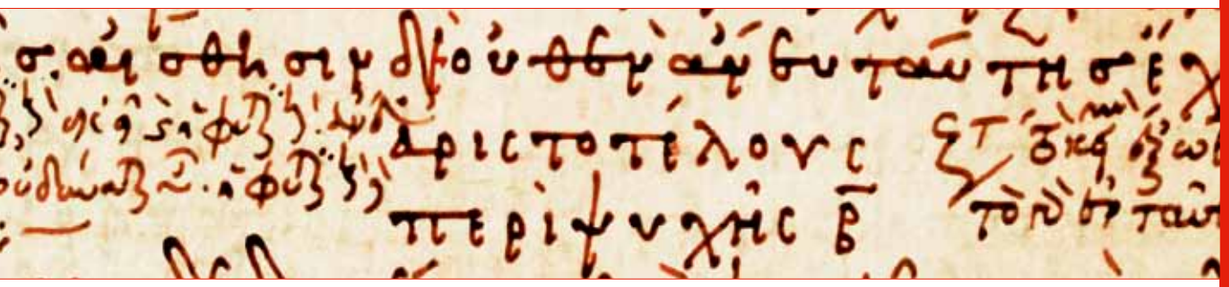
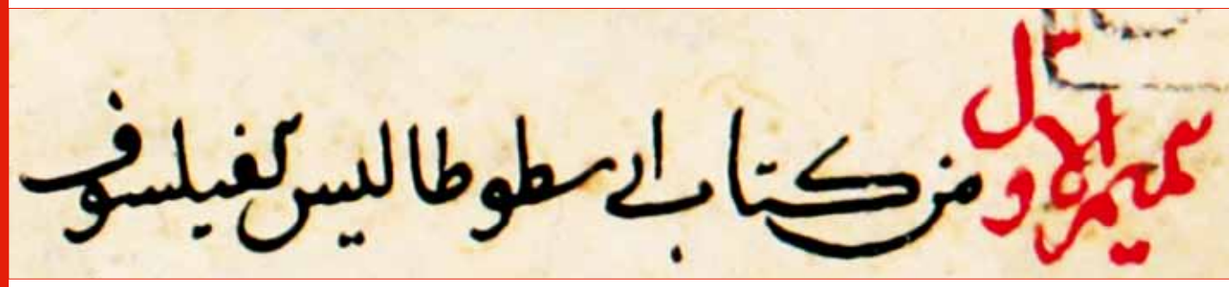
ISSN 2239-012X



With the support of the European Research Council

Studia graeco-arabica

Studia graeco-arabica



3

2013

Studia graeco-arabica

The Journal of the Project
Greek into Arabic
Philosophical Concepts and Linguistic Bridges
European Research Council Advanced Grant 249431

3

2013

Published by
ERC Greek into Arabic
Philosophical Concepts and Linguistic Bridges
European Research Council Advanced Grant 249431

Web site: http://www.greekintoarabic.eu
Service Provider: Università di Pisa, Area Serra - Servizi di Rete di Ateneo

ISSN 2239-012X

Registered at the law court of Pisa, 18/12, November 23, 2012.
Editor in chief Cristina D'Ancona.

Cover
Mašhad, Kitābḫāna-i Āsitān-i Quds-i Raḍawī 300, f. 1v
Paris, Bibliothèque Nationale de France, grec 1853, f. 186v

The Publisher remains at the disposal of the rightholders, and is ready to make up for unintentional omissions.

Publisher and Graphic Design
Pacini Editore
Via A. Gherardesca
56121 Ospedaletto (Pisa) - Italy

Printing
Industrie Grafiche Pacini

Studia graeco-arabica

3

2013

G2A Web Application

Istituto di Linguistica Computazionale “Antonio Zampolli”
Consiglio Nazionale delle Ricerche - Area della Ricerca di Pisa

G2A: a Web application to study, annotate and scholarly edit ancient texts and their aligned translations

Part I. General model of the computational philology application

Andrea Bozzi

Abstract

This paper presents the general model of a Web application for computational philology and describes the modules implemented by ILC-CNR in Pisa for the ERC project *Ideas* "Greek into Arabic. Philosophical Concepts and Linguistic Bridges" ADG 249431 (acronym: *Greek into Arabic*). The main principles on which the model is based are modularity, flexibility and development of the software according to open source criteria. These elements make it possible to include additional components in the modular structure, as well as components essential to the *Greek into Arabic* project (modularity), thus allowing the application to extend its functions to many other philological fields, from classical and medieval philology to genetic criticism and philology of ancient printed texts (flexibility). Dissemination of this application, especially in the research and academic fields, is guaranteed by the fact that its development is performed using internationally acknowledged systems of standard mark-up language and tools with no copyright restrictions (open source). In Part II a preliminary version of the user manual of G2A Web application is provided.

1. General model of the computational philology application

1.1. Background

The design of a Web application for a comparative and collaborative study of a text and of its translation involves a number of factors which are not only technical but also philological and linguistic. This is particularly true for ancient philosophical texts, in order to allow specialists to perform hermeneutic activities by means of efficient and simple-to-use investigation and research tools.

The activities which for some years have been carried out in research centres specialized in the creation of information systems for Linguistics and Philology have hardly treated this issue, even though interesting results have been obtained, processing single texts or textual corpora and producing indexes, concordances or lexicons. However, it is not easy to use the same text processing programs designed for the treatment of a single work or corpus of texts belonging to the same linguistic field so that, with only a few changes, they can also be employed for the comparison of a text and its translation in a second, third, or nth language.

In this paper I will examine the methodological features at the basis of a Web application specifically studied for "Greek into Arabic. Philosophical concepts and linguistic bridges" (*Greek into Arabic*), a research financed by the European Research Council (ERC), coordinated by Cristina D'Ancona of the Department of "Civiltà e forme del sapere" (University of Pisa), with the participation of the Institute for Computational Linguistics "A. Zampolli" (National Research Council, Pisa), for both the technological and philological computational aspects, and of the Seminar für Orientalistik (University of Bochum). What follows is only a brief description of a rather complex project aimed at connecting the chapters of the *Enneads* of Plotinus with the Arabic translation performed around

the 9th century and known as Aristotle's *Theology*. Comparison and knowledge of objective data are indispensable if one is to be acquainted with the method adopted by the translator, especially when the original text presents difficulties of interpretation for those who have lived in a very different cultural ambience and have devoted themselves to transmitting ideas, concepts, moral and religious principles from one culture to another. In the next phase, the work will consist in extracting meaningful terms that have an important role in both texts, and that will be included in the dictionary of Greek-Arabic philosophy produced at the Seminar für Orientalistik (University of Bochum), the other partner of the *Greek into Arabic* project.

As we can see from this concise list of tasks, the computational component does not only provide simple text processing programs similar to the ones that produce the alphabetical indexes of occurrences accompanied by all the contexts in which they occur (the so-called concordances). This and other past projects have led to study a more general method of computer-assisted management of translated texts, ancient works with their ancient translations on the one hand,[1] and ancient works with their modern translations on the other. It is worth underlining that the work, currently in an advanced stage of development, is referred to texts of particular historical and cultural value. For a comparative study of these texts, it is useful to dispose of an application shared by multiple specialists collaborating in the activities of study and interpretation. For this reason the computational Greek into Arabic Web application (G2A) should be provided with features like flexibility, use of open source tools and, above all, modularity.

• *Flexibility*. Firstly it must be flexible, and adaptable also to other projects, so that the efforts may be distributed over multiple works.

• *Open source tools.* Secondly, it is necessary to follow criteria of software development which do not use tools that oblige the project designers and final users to pay for the rights or to subscribe user licences to private bodies. The added value of systems implemented in academic Institutions and public research centres consists in the free circulation of software developed for scientific purposes with no commercial implications. Only the intellectual property rights of the designers and authors is safeguarded. It will also be possible for scholars to modify the source code in order to reach personal aims and to share the realized adaptations with the community of digital philologists.

• *Modularity*. Thirdly, the application should respond to modularity criteria strictly connected to the flexibility of the system: the architecture and technical structure should envisage the link of different modules, each responsible for specific tasks. According to this principle, the modules are activated on the basis of the goals of the project; if necessary, the general system should include and activate new modules not considered previously. G2A application is in any case a multivalent and multifunctional ambience whose basic structure is already designed to host the following annotation management tools, indexes, lexicons and apparatuses.

- *Module for annotations and variants*. If a text is transmitted by different witnesses and it is necessary to record variants and errors in their critical apparatuses, a specific 'Apparatus' module is activated. This module requires the collaboration of experts working in the field of textual criticism,

[1] One of the first experiments for the development of a computer-assisted program able to manage ancient texts and their translations was performed at the "Informatique et Bible" Centre in Maredsous (Belgium) during the '70s by Ferdinand Poswick: see *Actes du Premier Colloque international Bible et informatique: le texte (Louvain-la-Neuve-Belgique, 2-3-4 septembre 1985)*, Slatkine, Genève 1986. See also A. Bozzi, *Il trattato ippocratico sulle arie, le acque e i luoghi e la sua traduzione latina tardo-antica. Concordanze contrastive con il calcolatore elettronico e commento linguistico-filologico al lessico tecnico latino*, Giardini Editori e Stampatori, Pisa 1981.

and it would be an extension of a more general module for annotations and comments, in the general structure of the system, without having to modify any of the pre-existing components, but expanding, in this way, the fields of application of G2A to ever wider areas of the philological disciplines.[2]

- *Indexing module: morphological analysis*. Another component already installed and running in the multi-modular structure of the application is represented by a program specifically designed to assist the scholar in the production of lexical indexes. This is a component of Natural Language Processing (NLP), which uses specific linguistic rules able to perform the automatic morphological analysis of an inflected language.[3] As we have already said, one of the principal scopes of the G2A application consists in extracting Greek and Arabic philosophical terminology. A fundamental step towards the achievement of this goal is obtained by methods of automatic processing of occurrences, each of which is thus associated with its lexical entry (lemmas). With regard to Ancient Greek, various programs have been made available over the years, which well perform automatic lemmatization by segmentation of the forms, assigning corresponding codes to the grammar values (POS, or parts of speech); for the Arabic language, a similar tool has been appropriately designed and implemented within G2A application, making it easy for specialists, in our case the team of the University of Pisa and the editors of the *Glossarium Graeco-Arabicum* (Bochum), to semantically evaluate the philosophical terminology.[4]

- *Module for description of the lexicon*. Other additional modules have been considered and implemented experimentally, but have not yet been integrated in the general system. One of the most innovative of these modules is capable of describing the semantic value of the lemmas, according to a defined structure (metalanguage) based on standardized criteria. This method, based on principles developed within the philosophy of language and its applications in the NLP field, could be effective to highlight the conceptual relations existing between particular terms (moral, religious and philosophical) contained in contexts with strong abstract connotation. Moreover, it could make consultation more articulated with respect to what is obtained by submitting single forms or lemmas to the system.[5] This method could, instead, query lexical data as well as parts of the text in which they are located, using semantic features or semantic relationships as keys for querying.

- *Module for management of authorial changes.* The implementation of another module mainly addressed to scholars specialized in genetic criticism and to those operating in the sector of the philology

---

[2] This aspect is discussed more exhaustively from a methodological point of view in section 5 below and in further detail in the technical contribution by Angelo Mario Del Grosso.

[3] By NLP we intend Natural Language Processing systems which, thanks to the use of Computational Linguistics programs (morphological and morphosyntactic analyzers, natural language parsers, extractors of meaning from a text, etc.) contribute to enhancing the value of information on a text, increasing the possibility to respond to complex queries. The standard value is represented by the fact that the annotations introduced by NLP systems follow criteria which are shared at an international level making it possible for archives and corpora implemented at different sites to interoperate.

[4] Ouafae Nahli deals with this aspect and with the problems correlated with the conversion in digital format of Arabic texts in her contribution to this volume. As concerns the more general aspect of automatic processing and lemmatization of Latin language, see A. Bozzi - G. Cappelli (eds.), "A project for Latin Lexicography: 2. A Latin Morphological Analyzer", *Computers and the Humanities* 24 (1990), p. 421-6.

[5] This system implemented by Nilda Ruimy (ILC-CNR) and Silvia Piccini (Università di Pisa; ILC-CNR) was carried out on a vast number of terms of Saussurian linguistics, in view of the creation of digital electronic editions of manuscripts written by the great Genevian linguist; detailed information on a similar topic is available on http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=917/vers=ing. The set of Greek and Arabic words chosen by Ruimy and Piccini, which has been semantically described in structured form using the computational lexicon theory, is yet too small. A major quantity of data is necessary to test the validity of this approach and this is the reason why the experiment does not appear in this issue.

of ancient printed texts is envisioned. Such a module is an extension of the G2A application: even if it is not requested by this specific project, it can contribute to become a robust research infrastructure for the computer-assisted philological disciplines. Hypertextual and multimedia technologies, so far and very often adopted in these two situations, have given results only partially satisfactory and, from a methodological point of view, do not seem to have made real progress.[6] Basically, it consists in a technological approach – often based on previous works carried out with traditional philological methods – designed to tag a text so as to facilitate non-sequential reading. The resulting hypertext has allowed multidirectional reading, from that of the avantexts up to the corresponding printed version, which has, often erroneously, been considered the final one approved and licensed by the author himself. The technological choice adopted in G2A application is not hypertext-like: in the program that manages annotations and comments, it has been decided to insert a second sub-module in addition to the one that handles variants and errors transmitted from different witnesses of the same text. It looks like an editing tool for *avant-textes*, offering the possibility of creating a 'genetic apparatus', the functions of which will be described later.

1.2. General model

For a system based on modular architecture to be able to respond to the many requirements of the *Greek into Arabic* project and to assume the role of a research infrastructure in the field of both textual and lexical ancient philosophy, it is necessary to take into account some of the elements constituting the formal modeling. A study of the technological components and of the tools most suitable to software development is then performed on that model.

I will now present the essential parts of the model, without entering into technical details which are discussed by Del Grosso in his contribution, and partially examined in the *Appendix* illustrating the G2A user manual.[7]

• *Standardisation*. Firstly, G2A adopts the XML language; in the specific case of text labeling and of its associated features, it uses the XML version of the Text Encoding Initiative (TEI), a standard level internationally shared in the fields of text processing and digitized textual archives. This choice is practically compulsory, since the value of the codes employed is unique and known, being described in easily accessible guidelines on the Web.[8] The rules described are certainly exhaustive, but it is necessary to establish beforehand which elements in the text must be codified, in order to facilitate the preparatory work, before loading the digital document on the Web. As often occurs for these studies and for the update of data storage methods, it is necessary to find a compromise between the expected results (major or minor elements of the text to codify), and the invested resources in terms of times and costs. For example, the machine can produce an index of words found in citations, as long as an appropriate code is included at the beginning and the end of each citation, as indicated in the TEI guidelines. It should be reminded that there are a number of Centres dealing with the problem of text digitization and of the production of NLP tools which for various reasons do not follow the TEI instructions, but which can at any time produce tables of correspondences between

---

[6] Some interesting examples of this kind are represented by the HyperNietzsche project (see P. D'Iorio, *HyperNietzsche*, PUF, Paris 2000) and by the Samuel Beckett genetic edition project (see http://www.beckettarchive.org/introduction.jsp). Also see P. Delany - G.P. Landow (eds.), *Hypermedia and Literary Studies*, MIT Press, Cambridge 1991.

[7] See Marchi, "Towards a user manual", p. 173-83.

[8] The aims description and the most recent version of the manual with the complete set of the markup elements to encode documents are available at the following address: http://www.tei-c.org/Guidelines/

the mark-up criterion they consider most appropriate and the values expressed by the coding imposed by TEI. Therefore, the decisive element is the respect of a criterion which, once established, should be followed clearly, to avoid the loss of reusability of one's data by the users.

• *Processing: from primary to secondary sources*. The second component of the model is represented by two different paths followed by the documents once they have been digitized (for example, by means of a system of optical character recognition in the case of printed volumes) and encoded.

First path: the documents can be directly loaded onto a Web page, with no additional interventions; second path: they can be submitted to computational processing through different programs designed to perform specific operations like indexing, concordances, morphological and syntactic analyses, etc. In this case, any element added to the primary source (the text digitized and loaded on the server for consultation via Web) must in turn be linked to the linguistic form to which it is referred, and encoded according to a standard mark-up language, as described in the previous paragraph. In this case, the text is transformed from 'primary' to 'secondary' source, i.e. a document enriched with information it did not have before.

The number of processing tools is practically unlimited, and often depends not only on the flexibility and tailoring one intends to assign to the application, but also on the type of document to be processed. For example, a study on the severely damaged fragments of a Greek papyrus requires the use of processing programs which are certainly different from those necessary for the study of texts published on easy-to-read 19th century printed books, although some valid programs may be applied to most linguistic and philological works.

• *Production (on paper vs on the Web).* Going back to the scheme of the model I am describing, it is now possible to insert the digital text on the server, this time accompanied by additional information introduced by electronic systems, eventually revised and corrected by a specialist. The document can be consulted in two different ways, either in Web format or in traditional paper format. Both the data of the text and those of the information added by software components, are always accompanied – as we have already seen – by encoding systems which, in turn, are interpreted by a page layout software. Any apparatus notes, indexes, annotations, bibliography and other information will be included according to the author's and publisher's choices.

I will now consider the specific methodological aspects of G2A, obviously based on the model described briefly above, and consistent with the specific needs of the study and of the goals it intends to achieve which, among others, will be evaluated by a panel of experts appointed by ERC.

1.3. Textual analysis method

A crucial factor for the success of the project depends on the implementation of an innovative method allowing a comparison of parallel texts. In this way it will be possible to read the original work alongside its corresponding translation. In the specific case of the *Greek into Arabic* project, one will need to dispose of a file containing the treatises of the *Enneads* accompanied by the text of their Arabic translations.[9]

Despite the great development of text processing programs realized over the last decades, there is a considerable lack of products able to manage texts and their translations for scholarly philological studies. There is also a lack of similar tools to assist a single translator (or a collaborative team of translators working simultaneously), who can better evaluate the sentences and meanings in order to

[9] The text adopted is ʿA. Badawī, *Aflūṭīn ʿinda l-ʿArab*, Dār al-Nahḍa al-ʿarabiyya, Cairo 1966.

render them consistent and uniform. These tasks require strategies more complex than those offered by a graphical interface, facilitating the distribution of two files (opened by a wordprocessor) on the same screen of the computer; the original text on one side (e.g. left or top) and the facing translated text on the other (e.g. right or bottom).[10]

The design of our system takes into account the specific needs of the *Greek into Arabic* project but, in compliance with the criterion of maximum flexibility and adaptability, it analyzes the phenomenon of translation from a more general point of view in order to extend its potential applications. Indeed, G2A provides useful research tools, not only to scholars who investigate Medieval translations of ancient works, as in the case presented here, but it also meets the needs of translators currently involved in ancient or modern works.

This is not only a theoretical prospective predicted by the model, but it is based on an active involvement of ILC in other projects with similar elements. The text alignment and comparison, the creation of indexes organized alphabetically by inflected form or by lemma, as well as the possibility to perform annotations related to the particularly difficult interpretation of some passages are all valid actions not only for *Greek into Arabic* researchers but also for other communities of scholars.

• *Preliminary alignment and comparative reading*. Let us now further analyze this issue in relation to the aims of G2A, facing the problem of how the Greek and Arabic texts must be structured, within the framework of a computer-assisted application. This will facilitate the interpretative process concerning the semantic value that some keywords of Plotinian philosophy have conveyed in the *Enneads* and, in particular, the semantic value that the corresponding Arabic terms have retained or modified. The method is based on the comparison of perfectly corresponding sentences in the two texts involved. Even in the case of omission of a passage, an empty context must in any case be present, so that parallelism may be guaranteed.

On the basis of some experiments performed at ILC-CNR, the segmentation of texts in semantically corresponding parts was successfully created some years ago by an off-line program,[11] which consists in synchronizing two texts and visualizing the corresponding sentences. The program allows the scholar to validate the results of the automatic procedure or to intervene by introducing corrections, adjustments or changes. An approach of this type would not have been possible for *Greek into Arabic* project, since comparison, due to the peculiar features of the two texts in question, must be performed using manual alignment by the scholar. Only the specialist is in fact fully aware of the problems involved, and is able to establish the exact boundaries of the Arabic contexts which translate, either in literal or interpretative form, the corresponding contexts of the original Greek text.

In brief, two alternatives were possible when dealing with the methodological aspects of text alignment: the former, based on a semi-automatic procedure of synchronization, if necessary with human intervention to correct any mistakes; the latter, by which an operator is responsible for manual alignment in the form of parallel pericopes to be done when the texts are recorded in the memory of the computer.[12]

---

[10] A simple consultation of automatic aligned Greek and Arabic texts, without the possibility to perform special queries or to use navigation tools, is available in the Alpheios project at http://www.beckettarchive.org/introduction.jsp

[11] C. Peters - E.Picchi, "Bilingual Lexicons, Parallel and Comparable Corpora: Creating the Basis for Cross Language Information Retrieval", *Linguistica Computazionale* 18-19 (2003), p. 573-96.

[12] Broadly speaking, by 'pericope' we refer to a portion of text that makes up a unit of thought, and that the specialist has labeled with appropriate delimitors.

The first alternative proved to be unacceptable for the reasons I have described, whereas the second was a useful preparatory work, which made comparative reading easy, and represented a valid starting point for the specialist. In order to evaluate and interpret Greek and Arabic, the scholar can confirm the previously traced boundaries of the pericopes, but the system also allows to expand or reduce the text they contain.

The user interface allows scholars to intervene easily when they want to associate semantic interpretative annotations to an entire Greek-Arabic pericope, or to parts of them.

The software was implemented on the basis of the principles described above and makes it very easy to read the two works following the original sequence of each one: the graphical interface allows the user to position himself on the Greek text and to see the corresponding Arabic translation or, vice-versa, to position himself on the Arabic translation and browse through the corresponding text in Greek. Immediate visualization of the parts that have been transposed, re-elaborated or removed by the Arab translator is facilitated by the fact that the correspondences appear in the form of pericopes. As we have already said, this result constitutes the first stage of a process of semantic assessment, followed by two successive strictly correlated phases.

1.4. Methodology for annotations and comments

It should be recalled that one of the aims of the G2A application is to put at the disposal of the editors of the *Glossarium Graeco-Arabicum* at the University of Bochum a computer-assisted lexicographical ambience in which the lemmas extracted from the Arabic Plotinus are appropriately documented by the contexts in which they occur. The lexicographical component is made up of two elements: the choice of contexts in the form of parallel pericopes (alignment) and the annotation/comments made by scholars.

• *Pondered alignment and annotation of the pericopes.* The former is represented by the pericopes in which each linguistic form is attested: it means that the pericopes constitute the context. Generally speaking, the electronic processing systems of linguistic and literary data use contextualization rules based on punctuation; instead, in our case, we have avoided the arbitrariness of this criterion and have created concordances in which the contexts are represented by the pericopes defined by those who have been able to evaluate their meaning and to mark their boarders. Once again, the application is distinguished from the widespread practices of automatic generation of the contextual concordances which do not sufficiently take into account the fact that there can be cases, like the one in *Greek into Arabic* and in many other projects of computational philology, which require the contexts to be cut manually. Delegating this operation to the computer could prevent a correct vision of the role played by a linguistic expression in the exact position in which it is found. This is particularly valid in the case of philosophical words which have been translated in an ambience very different from the one in which they were conceived and then transmitted in the course of centuries.[13] It is for this reason that the system, as already anticipated, makes available a function thanks to which the limitation of parallel contexts and, therefore, of their alignment, is the exclusive responsibility of the scholar who is making an in-depth interpretation of the text.

The second element appears as an easy option to select in the graphical interface of the application; it allows to select the entire pericope (or of one of its parts), so that it can be associated with an

[13] An example of a work performed with the aid of an automatic indexing system, but with manual delimitation of the contexts in concordance, can be found in M.S. Corradini, *Concordanze delle biografie trovadoriche*, I-II, Pacini Editore, Pisa 1982.

annotation.[14] The information is recorded in a database, and represents valuable materials for the dictionaries of Graeco-Arabic translations, especially for our partners in Bochum. These types of annotations are generally referred to a lemma or to linguistic expressions formerly recorded in that lexicon; however, there may be cases in which the linguistic form present in the *Greek into Arabic* texts is not recorded, but it can be introduced by means of useful documentation. The annotations are not expressed in structured, but in discursive form, as if they were short or very short monographic essays about an expression, word, particular or unusual use of a word compared to its common use.

• *Classification of thematic annotation.* G2A has also envisioned a further possibility of navigating within the textual data, using a classification of the annotations as key of access. Offering to the scholars the possibility of organizing their comments and annotations according to a typology represented by key words (a type of subject catalogue) is a considerable advantage especially in the phase of information retrieval. Thanks to this tool which is obviously optional and only used in cases of necessity or utility, the system easily shows all the pericopes which present common traits or refer to the same theme in the annotations. The implementation of this procedure is not technically complex and does not imply an excess of work and time. In other words, it consists in introducing a drop-down menu, a series of key-words or subjects (for example, 'misrepresentation of sense', 'sense extension', 'sense reduction', 'radical transformation', etc.), which synthetically classify the annotation associated with the pericopes by the scholars themselves. The result produced by the system after a specific 'query' has been made provides a synoptic vision of the set of Greek-Arabic pericopes responding to the type of annotation denoting them all. The scholar can thus confirm his/her observations on the basis of better selected and larger amounts of data, or go back and change the comments previously provided, this time in different manner, more convenient and conformant to the texts.

Summing up, the application provides two different ways of inserting the annotations: a free form and a controlled form. Free annotation implies the selection of an entire pericope or one of its parts (up to the point of considering, if necessary, a single word) and opening of a portion of the interface in which non structured and unlimited comments can be included. Instead, controlled annotation allows the specialist to list different types of judgement. Once these classes have been included in the menu, they will be recalled whenever the scholar intends to mark a pericope or a particularly significant expression which should be connected and evaluated comparatively with the others classified in the same way. G2A does not only offer indices of words present in the Greek and Arabic text, but also automatically presents all the parallel contexts annotated with the same classification.

The system has been designed with requisites of flexibility, adaptability and reusability, and therefore the structure of the annotations described above is able to respond to the needs that could appear to be distant from those required by the *Greek into Arabic* project, while the diversities are related to the interface, so that it can conform to other types of users. The method is well consistent with the requirements of digital philology works on translation texts. It is extremely useful for the translators to dispose of a system function able to cut out portions of text of different size and content. Therefore, this is an excellent approach if one is to render any text in a language different from the original, especially in the case of ancient texts dealing with complex themes, and object of successive reflections which may have modified or altered the original value. Such a value must be recovered at the moment of its first translation or when a previous or an old translation is updated.

---

[14] A particular and technical aspect relative to annotations is described in the contribution by F. Boschetti in this volume.

• *Annotations: sub-module for avant-textes.* The study of the model and the modes of design concerning the re-usability of the system envisions an extension of performance, so that the system can be adapted to other fields of philological studies such as genetic criticism.[15] After a phase of essential controls currently in course, the application could contribute to the production of digital editions of manuscripts where it is possible to read multiple interventions by the authors themselves or by second hand sources. Generally speaking, we intend to check whether, thanks to the structure of the module managing the annotations, it is possible to appropriately organize and treat the many *avant-textes* which can overlap in the same page of a document and which, in many cases, have preceded the publication of a work. This theme falls outside those included in the *Greek into Arabic* project and those we have spoken about so far, but they have in any case been considered in the design stage of the system, because a solution suitable and appropriately checked on concrete data could allow its application to other sectors of digital philology studies like genetic philology. The difficulties in treating author variations by hypertextual-multimedia methods and languages are well known. Satisfactory results can sometimes be obtained, but at the price of excessive work for text encoding.[16] On the contrary, the choice of organizing the data by parallel pericopes that can be annotated with comments is more efficient and less demanding: it consists in considering the *avant-textes* as different 'versions' of the same texts. These different 'versions' are subdivided into pericopes by the philologist using the above-described procedure, analyzed by the system and presented in tabular form, where the columns (from left to right) represent the successive phases of re-manipulation, while the cells include the text of the pericopes. A portion of text cancelled by the author will appear within two aligned cells belonging to two different columns: the first cell will contain the text, while the corresponding text will be empty. On the other hand, any additional data will appear as two facing cells the first of which will be empty, while the other will contain the part of text added at a later stage. There will be as many columns and therefore cells as the phases of re-writing performed by the author.

According to the limits imposed by legibility of what has been cancelled, it will be possible to perform sequential readings of different *avant-textes*, on the basis of the column used by the system, at the request of the scholar, in order to sum the pericopes contained in its cells. The system is able to regenerate the *avant-textes*, clumping the sequence of the pericopes and facilitating the study of the genetic process which has led to the creation of a literary work. The scholar of genetic philology disposes of exhaustive and well-structured elements to investigate stylistic, linguistic, psychological reasons, which have led the author to intervene on his work with cancellations, interlinear additions along the margin, footnotes, etc.

---

[15] Similar phenomena occur in ancient printed texts such as Vico's *Scienza Nuova*. A total of 63 of these copies are known, and are rich in autograph notes. The modern critical edition reports these annotations in the apparatus, but the paper support limits many forms of consultation, which only the electronic support is able to guarantee. In this case the main task consists in facilitating parallel reading of the printed text and the corresponding manuscript notes, not always identical in the different copies, and consisting either in interlinear interventions or actual glosses of variable length. Some simulations have shown that they can be treated with the same tagging module of the pericopes and relative annotations described above.

[16] The example of the HyperNietzsche project has already been quoted above. For other projects see, for example, D. van Hulle, "Compositional variants in modern manuscripts", *Linguistica Computazionale* 20-21, p. 513-27, Istituti Editoriali e Poligrafici Internazionali, Pisa - Roma 2004; H.W. Gabler, "Computer-aided critical edition of *Ulysses*", *Bulletin of the Association for Literary and Linguistic Computing*, 8 (1980), p. 232-48. Again about Joyce, see also D. Ferrer, "Représenter les manuscrits de Joyce: pour une édition hypertextuelle", in P. D'Iorio - A. Petrucci - A. Stussi (eds.), *Genesi, critica, edizione (Pisa, 11-13 aprile 1996)*, Scuola Normale Superiore, Pisa 1998, p. 227-32.

In other words, author variations and *avant-textes* can be structurally conceived as 'translations' of parts of a text in the corresponding parts of a second text, and a comparison of these 'translations' allows to introduce evaluations, as well as critical, semantic and interpretative annotations.

1.5. Method for the critical apparatus of variants

The *Greek into Arabic* project works on two manually aligned texts, indexed so that the scholar can analyze them comparatively and have an immediate visualization of the pericopes, which show the re-elaboration, in some cases quite considerable, performed by the Arabic translator. The text of the *Theologia* that is currently used requires a thorough study of the witnesses. A critical edition is being prepared by the scholars at work in the *Greek into Arabic* project: the G2A application therefore also takes into account a textual criticism component so that the comparison of the texts transmitted by multiple sources, partly collated, can produce an edition with an appropriate variant apparatus. Therefore, the design and development of a module for the storage and management of variants and errors has been envisioned.[17] These can concomitantly allow the editor to establish a new text, much safer than the one currently inserted in the system.

At this point, a series of problems arise, some more general, others specifically related to the project in question. Let us now analyze them in order. Firstly, it should once again be underlined that the application is based on criteria of flexibility and re-usability, but without overlooking the particular features of *Greek into Arabic*. We must remember that the borders of each pericope, the attribution of codes to parts of speech and the annotations or comments by scholars have already been made and linked to a text that will be re-edited in the next future. Therefore, the problem consists in translating the past work into the new Arabic text critically established by the editor in a simple and straightforward manner.

• *Critical apparatus tools.* The main module components of the critical apparatus are two: – the component for the management and treatment of digital images of the sources, with the possibility of transcribing the one that has been ecdotically considered the most reliable; – the component of registration of the variants in a database. As we shall see, the two subsets have been designed so as to be able to work independently from one another, collaborating so that the actions performed on one of the two can affect the other.

– *Treatment of the digital images of the sources.* The first subsystem allows the scholar to consult each time the text transmitted by the witnesses, browsing through the images and showing them on the screen. This phase simulates the traditional lecture of a text performed directly on the originals or on photographic reproductions or, again, with the aid of microfilm readers. However, G2A can also work in the absence of digital images, in other words it is capable of processing textual archives without images. It is essential that this philological research infrastructure is organized in modular form, where the graphical manipulation module of the images is only one of the functions available.

---

[17] The first experiment of a computer assisted philological workstation at ILC-CNR was realized for the European project BAMBI, see A. Bozzi (ed.), *Better Access to Manuscripts and Browsing of Images. Aims and results of an European Research Project in the field of Digital Libraries (BAMBI LIB-3114)*, CLUEB, Bologna 1997. Further methodological information can be found in A. Bozzi, "New trends in philology: a computational application for textual criticism", in A. Zampolli - L. Cignoni (eds.), *Linguistica Computazionale* 16-17 Special Issue, Istituti Editoriali e Poligrafici Internazionali, Pisa - Roma 2003, p. 47-77. The description of an application for classical philology and, in particular, for Greek papyrology can be found in A. Bozzi, "Digital documents and computational philology: the Digital Philology System (Diphilos)", in M. Veneziani (ed.), *Informatica e Scienze Umane. Mezzo secolo di studi e ricerche*, Leo S. Olschki Editore, Firenze 2003, p. 175-201.

If the data archive is without images, or if the user is not interested in using it, the module will remain inactive without compromising the performance of the others. Instead, if the images are present and have been loaded in the application, the most common and useful image handling functions (enlargements, brightness, contrast or chromatic levels) are made available, in order to facilitate the reading of the text they contain.

– *Critical apparatus*. If the reading of the sources implies the detection of variants with respect to the text chosen as the basis for collation, the system opens up an area of the working ambience in which there are as many variant-fields as there are collated witnesses. The set of these fields takes on the features of an omni-comprehensive apparatus in which there is also a field in which the editor can write the choices he has made. In this way, the user disposes of a tool necessary to the creation of a positive apparatus in which to insert all the readings of the witnesses, including those adopted in the critical text. Therefore, even from this point of view, the application assisting the editor during the creation of a critical edition simulates the traditional procedure. The editor can obviously ignore petty, insignificant readings (for instance, orthographic variants), useless to demonstrate relations between witnesses which of course will not be inserted in the apparatus. However, at least in the early stages of the work, it is advisable to record all types of discrepancies faithfully: both the errors that could contribute to determining the *stemma codicum* (for example, transcription errors such as haplography, homoioteleuton, dittography, etc.), and the readings that should later be cancelled, like the trivial errors of the amanuenses, useless to highlight particular *scripta*. In order to help the editor in this operation, the user interface should make available a section in which all variants can be supplied with necessary annotations (see above for the description of how to add annotations to the pericopes). An appropriate indexation system of annotations will, for example, help to retrieve all those readings considered trivial errors, to make an overall assessment and to remove them with greater awareness.

Since each reading included in the apparatus may have been analyzed in previous editions or specific critical studies, we also envision a space where to insert bibliographic information or a Web site where to retrieve data worthy of attention.

As previously underlined, the system disposes of a search engine for indexation, activated on the sections subject to study and assessment, as requested by the scholar. The system will thus produce the alphabetical indices of the readings of the critical texts, of the variants of each witness, able to connect the former (readings of the critical text) with the others (the variants) and viceversa.

A positive apparatus presents a quantity of information that can be processed by the application, offering various advantages for the philologist who has performed the transcription of the text from the only source considered the best, according to internal and external evaluations. Unlike other programs of computer-assisted textual criticism, this application does not oblige the user to produce the transcription of all the collated sources, but imposes the transcription of only one witness and is able to automatically generate the text transmitted by all the others.[18] Using the portions of transcribed text that are not associated with variants and integrating them, from time to time, with those attested in other sources found in the apparatus fields, the program can reconstruct the text of all the witnesses. Apart from those situations in which the procedure might not have a

[18] Very interesting and updated information about digital tools for literary studies and philological activities can be found in the Huygens Instituut KNAW, a research institute for text edition and textual scholarship of the Royal Netherlands Academy of Arts and Sciences (http://www.e-laborate.nl/en/).

full justification,[19] there might be the condition of sequential lecture of sources considered totally or scarcely reliable *a priori* but which, re-examined and observed as a whole, could suggest, even partially, to modify the reasons of the previous perplexities. It is evident, for example, that the search engine operating on the text of each witness generated automatically, can better evidence graphical, phonetic or linguistic features that are less evident when detected separately in the critical apparatus.

Moreover, this *modus operandi* meets the needs arising at the time of consolidation of the digital culture; the text assumes a growing mobile dimension, without the rigidity imposed by the paper support, at least for a certain period of time, until the text is re-edited and printed. The possibility of generating the text via a computer program does not actually produce a thorough change in the activity of the scholar who has studied the history and tradition of a text, suggesting a version as close as possible to the one created by its author, the original version of which has gone lost. What is instead increasingly changing is the possibility of exploiting the numerical condition that the text has assumed, owing to its conversion into digital format, and of producing quantitatively useful analysis data: however, these elements tend to diminish the uncertainty of a decision (the *divinatio* made famous by Lachmann), so that it could be increasingly based on objective and exhaustive phenomena.

It might be observed that an expert philologist does not need numerically considerable data to take a decision: with only a few certain, fundamental elements, he is able to operate a distinction between the sources to be collated and those to be rejected or only partially considered. Whether or not the scholar has hypothesized a stemma, the evaluation tools offered by the computer could appear very useful, in particular if this does not imply increased working times.The function of generating the text of single witnesses starting from the apparatus information can be naturally activated to associate the editorial choices with the readings of the other witnesses accepted by the editor: the established text can be inserted in a Web-connected server, producing a printed paper volume at the same time.

• *New and old: how to transfer the annotations to the newly edited text.* In the previous paragraph we have described the functions previously experimented thanks to which the following taggers are inserted: taggers to delimit parallel pericopes between Greek and Arabic texts; taggers for eventual annotations; taggers indicating lemmas for each Greek and Arabic occurrence, and taggers related to information added manually or automatically, for example the parts of speech of the words of the text. If the text containing the tagged elements changes, even slightly, following a re-editing process, there is certainly the problem of transferring the tagged codes and contents from the preceding to the new-edited text and this can present problems of considerable technical complexity.

The phases to consider on the basis of a simulation can be listed as follows:[20]

– subdivision in parallel pericopes of the re-edited Arabic text, while the ones included in the Greek text remain identical;

– the taggers associated with parts of text or single words which have not changed are re-utilized and associated with the corresponding elements of the new edition;

– the markers associated with parts of text or single words which, instead, do not correspond, are assigned by the system to the pericope they belong to;

---

[19] These phenomena include, for example, extensively contaminated sources sometimes leading to widely complex textual traditions, with considerable difficulty in reconstructing dynamically the text of all the collated witnesses, on the basis of a single text, using the functions of the software described above.

[20] These simulations based on real tests are however performed on parts of text formerly collated by the scholar.

– repositioning of the markers at the level of the single elements internal to the pericope is performed manually by the scholar, with the aid of a specific function of the graphical interface facilitating its work.

This solution seems to be the only one possible, also because some of the sample tests allow to hypothesize relatively simple and scarce eliminations.

1.6. Conclusive remarks

The *Greek into Arabic* project is a particularly interesting case study to see how a Web application for the treatment of digital documents (usable in collaborative form) can be used by a scientific community of scholars belonging to different disciplinary sectors. On the one hand, it is essential to respond to the specific needs of this project, on the other hand the structure of the application has been organized to perform more complex and integrated tasks. Modular architecture at the basis of single component production has proven essential to prefixed goals. The experiences made at ILC-CNR in the sector of computational philology, the prototypes realized over the past years, the specific modules for linguistic studies, demonstrate the erroneous opinion of inconsistency of the hardware and software tools, that characterize the digital technology of our times. Adaptation and re-writing of the code according to the current Web languages involves expenses and working times often inferior to those required in the pre-digital society, resulting in more reliable outcomes. G2A is a further example of how the user base of a highly specialized application can be increased. By means of appropriate devices and at moderate costs, the resulting system could be used for the production of scholarly editions and for the teaching of classical, medieval and modern philology, starting from digital documents and texts. Collaborative work between teachers and learners could consist in browsing and navigating textual corpora in digital libraries and archives, as well as writing of personalized annotations and comments subject to evaluation. We intend to propose the infrastructural dimension of applications like G2A, which is at the same time specialistic and general, according to whether the functions of the modules are activated or disactivated.

The components of subdivision of the texts into pericopes, currently performed by a specialist for the reasons explained above, can become increasingly automated so that the texts, either original or translated, are perfectly aligned. This process will need language control systems capable of convalidating centred parallelism on some support terms: multilingual dictionaries in electronic format are already used to this purpose for modern languages, but much work is still needed so that the same results can also be achieved for non-western ancient languages.

Finally, G2A constitutes a valid working tool for all specialists working on Arabic texts, translating works of philosophical or scientific Greek classical antiquity, examples of which are important treatises on medicine, botany and mathematics. Management of images, classification of annotations, realization of editions accompanied by critical and bibliographical apparatuses open prospects of research on entire textual corpora, using a method compatible with the increasing availability of library documents in digital format.

Finito di stampare nel mese di settembre 2013
presso le Industrie Grafiche della Pacini Editore S.p.A.
Via A. Gherardesca • 56121 Ospedaletto • Pisa
Tel. 050 313011 • Fax 050 3130300
www.pacinieditore. it