THE *THESAURUS LINGUAE GRAECAE* PROJECT: LOOKING TOWARDS THE 21ST CENTURY*

MARIA PANTELIA
University of California, Irvine

I would like to begin with two preliminary observations. First, looking at the programme of this conference, it appears that I am the only speaker who, strictly speaking, does not represent a lexicographical project. Having said this, however, I should also add that the *Thesaurus Linguae Graecae* (*TLG*™) Project, although not a lexicon *per se*, is a major lexicographical resource and in this sense, what I am about to say, should be of some value to this meeting. And this brings me to my second observation: considering the audience of this conference, there is no need for a long introduction to the *TLG*. I will, therefore, limit myself to only a few words about the history of the project and focus my presentation on the present state of its data bank and its future challenges and directions.

Efforts to create a Thesaurus of the Greek language go back to the late Renaissance when Stephanus published his 1572 Thesaurus, a lexicon based on 140 or so known ancient Greek authors.[1] At the beginning of our century, classical scholars decided that it was time for a new thesaurus, one that would accurately reflect the corpus of Greek literature as it was known by that time. An international *ad hoc* committee was appointed in Europe to consider the feasibility of this new Thesaurus. Hermann Diels, a member of the committee, argued that such a Thesaurus was doomed from the outset since extant Greek literature was estimated to be ten times the size of the Latin corpus, that is approximately 90 million words, and was simply too vast to be excerpted and semantically analyzed. Diels called the lexicon, if one could be created, "a monstrosity", and compared the task to an effort to bring "Nous into Chaos". Almost 100 years later the creation of a comprehensive historical dictionary of the Greek language would still be an extraordinary undertaking. At the same time, there is no doubt that the advent of technology has made what used to be a Herculean or an almost impossible task,

---

* a) All URLs were accurate at the time this was written.
b) Special thanks are due to Professor Brunner for having provided copies of several papers he delivered to conferences and scholarly societies over the 25 years of his directorship.

1. For a history of books and dictionaries see Schottenloher 1989.

much easier. One could even argue that technology has not only provided us with the means to compile such a dictionary but has also given us alternative ways, new media or methodologies in approaching and dealing with lexicographical issues. First and foremost one has to wonder whether a huge printed dictionary of the Greek language such as the one envisioned in the early 1900s is feasible or even necessary[2] or whether an electronic lexicographical database, one that could be continuously updated and reorganized would be more preferable today. But, this is an issue that, I hope, this meeting will further explore today.

1. THE *TLG* PROJECT

Established at the University of California, Irvine in 1972 with a generous gift from Marianne McDonald and technical support from David Packard, the *TLG* constitutes the first effort to create a comprehensive computerized collection of electronic texts in the Humanities. From its inception the *TLG* was not designed as a lexicon, in a fragmented lexical arrangement, as previous thesauri have been or intended to be, but as a collection of complete texts, each of which would be a "mirror image" –as far as this can be accomplished– of the source edition. The objective of the project was to identify and collect all known ancient Greek literature and create a permanent electronic record of it, thus providing a resource that would facilitate the development of other research tools, including but not limited to lexica. Even in the 70s and early 80s, producing such a collection of electronic texts required a special computer (a modified HP-1000), a series of programmes to verify and correct the texts, and a special character-encoding, first Alpha and later Beta code, to overcome the problem of the Greek script.[3]

Twenty-five years later, this electronic databank is a reality. In fact, classicists have been very fortunate to have several major computerized resources available to them. In addition to the *TLG*, the papyrological and epigraphical collection produced by Duke,[4] Cornell, and the Ohio State University under the aegis of the Packard Humanities Institute, and more recently the *Perseus* Project,[5] have allowed scholars in our field to research questions that would have been too long and tedious to pursue before. Since classicists no longer have to spend time compiling raw data they can concentrate on more interpretive and analytical subjects. Sorting and alphabetizing vocabulary is no longer a time-consuming and labor intensive task. Materials are now easily portable and accessible. Instead of searching around the world for obscure texts, scholars today can carry the entire corpus of ancient Greek

2. See Johnson 1994, 253-258 and Crane 1998, 471-501.

3. For Beta and coding convention for Greece see MacKay 1996, 221-229 and Rusten 1996, 204-215.

4. For a description of the Duke Data Bank of Documentary Papyri see Oates in Solomon 1993, 62-72.

5. The bibliography on the *Perseus* Project both in electronic and printed form is rather extensive. Up-to-date information can be found at the *Perseus* web site at URL: http://www.perseus.tufts.edu.

literature on a compact disk. What is even more important they can search the contents of that disk often in seconds. For example, St. John Chrysostom, the most prolific author represented in the *TLG* databank, has more than 4.5 million words in edited homilies, letters and commentaries. Locating all the occurrences of a given word in his works could take a lifetime. With the use of a computer, a full search can be done in one to two minutes depending on the speed of the computer one uses.

Another important accomplishment of the *TLG* is the *Canon,*[6] a comprehensive list of all extant ancient authors and works compiled by Luci Berkowitz and Karl Squitier. The *Canon* was originally compiled as an electronic aid to the *TLG* staff, as a way of keeping a record of the authors and works that were being converted to machine-readable format. Soon, however, the *Canon* became an invaluable resource, the first truly comprehensive list of all known texts. Today the *Canon* is available in both printed and electronic form and contains more than 3,300 authors and in excess of 10,000 works providing information about the names, dates, geographical origins and descriptive epithets for each author together with detailed bibliographical information about existing text editions for each work.

The *TLG* is now preparing for the fifth edition of its CD ROM which will contain in excess of 73 million words representing the entire corpus of Greek Literature up to 600 A.D. plus the Scholia, and most historiographical and lexicographical works up to 1453 A.D. The *TLG* CD ROM contains only the texts in the so-called Beta code whereas the tools, i.e. the software to search the disk, are provided by various software developers. At an early stage a decision was made that the *TLG* would not engage in the development of search software. As a result a number of software packages were produced by independent developers. This policy had some good and some not so good aspects. On the good side, the multiplicity of software has given users a wide variety of choices and platform independence. On the other hand, technical support for these programmes is not always available and what is even more important, the type of searches one can do varies from one programme to another. At the same time, without hands-on experience in software development, the *TLG* project did not have the opportunity to evaluate or reconsider its encoding or citation system, issues that would certainly have been the subject of some discussion, if the project had been involved in software development. As we look into the future, we face a number of challenges but also exciting prospects:

1.1 DATA ENTRY AND THE CONTENTS OF THE DATABANK

The original scope of the project, as determined by the 1972 International *TLG* Planning Conference was to cover ancient literary works. By the mid-80s the original plan was modified to allow the expansion of the databank into the Byzantine

6. Berkowitz & Squitier 1990[3]. See also Berkowitz 1993, 34-61 and Squitier 1987, 15-20.

period. It was decided that at the first stage only the Scholia and selected historiographical and lexicographical works up to 1453 A.D. would be included. These three areas will essentially be complete with the next release of the *TLG* CD ROM. Obviously, such divisions and selections are completely arbitrary and make no sense, if one wants to create a comprehensive and permanent electronic collection of Greek texts. We have, therefore, decided to continue data entry beyond the genre of lexicography and historiography and include all known Byzantine texts. Once that period has been completed, we will have to consider our next step, that is, how far down to proceed into the post-Byzantine period. It is our hope that modern technology will allow us to improve our methods of data processing so that we can move at a much faster pace. This will be very important because Byzantine literature is vast and far more unexplored compared to the classical period. Text identification and procurement along with the complex literary and historical research activities involved in expanding the *TLG Canon* will require time and reallocation of resources. Continuous interaction between the *TLG* and specialists in the various aspects of Byzantine literature will also be necessary since this is very much a new area for us.

In addition to digitizing new text editions, substantial retrofitting or updating of our present holdings will be necessary. As new editions are published, the project will have to invest considerable resources in updating its texts. This process has already started. For the time being, we are actively looking at the works represented in the databank, replacing outdated editions with more recent ones. Needless to say we are always working under strict copyright restrictions. In some cases we are not able to use the most recent or most widely accepted editions because permission to include them could not be obtained. We believe that copyright will become an even more pressing issue in the future as publishers gradually switch to electronic publishing.

1.2 SOFTWARE DEVELOPMENT

The *TLG* will have to concentrate on developing software tools and creative ways to access its databank. Although we believe that the texts in the CD ROM should remain in standard and independent text format so that other software can be used to search them, future releases of the CD ROM (or any new medium used for the dissemination of the texts) should include search software. For this purpose, we are presently increasing our technical staff and exploring new technologies. It has been a general observation that the quality of retrieval tools, that is, the existing software to search the *TLG*, has not taken advantage of the rich potential of the databank itself. All retrieval tools rely on simple string and substring searches and do not even make use of existing indices in the CD ROM. In all fairness, the problem is also embedded in the databank itself, as it does not contain a more complex

type of text encoding, such as SGML (Standard Generalized Markup Language)[7] or the currently much talked about XML, an expanded version of HTML, supported by the W-3 Consortium.[8] Text-encoding would facilitate or enable the creation of programmes to perform more complex and detailed searches. This type of retrofitting will require enormous resources and may or may not become a priority in the near future depending on how many of these functions and searches can be helped by full lemmatization and use of the *TLG* word index. In addition to text-encoding, we hope that an international standard for Polytonic Greek (possibly the Unicode)[9] will soon be adapted and fully supported so that data entry can be done in Greek. Such a standard will eliminate a plethora of technical difficulties and most importantly the need for different keyboard utilities.[10]

There is also the issue of the critical apparatus, which has been the subject of numerous discussions over the years. Although we all agree that the incorporation of the critical apparatus into the databank would be a worthwhile undertaking, the technical realities are such that I cannot foresee such a project starting in the near future.

Technological trends will determine whether there will be need for a new compact disk beyond CD ROM E. What we are experiencing at this stage, at least in the U.S., is an emphasis on fast networks and web access of information. For this reason, we are placing more emphasis on web technology and, thanks to the *Perseus* Project, we have in place an experimental or what we usually refer to as a "beta", web site. By accessing this site –which is now limited to a handful of institutions– one can retrieve and search the *TLG* using a standard web browser and software developed by the *Perseus* Project. Our goal from now on will be to expand the searches presently possible, improve our user interface, and work out all the technical and security/copyright issues associated with web dissemination so that we can offer the *TLG* to all our CD ROM users via the Internet.

I have already mentioned the collaboration between the *TLG* and *Perseus* to provide web access to the *TLG* databank. We can see a day in the near future when researchers and educators will be able to combine the resources of these two projects with all the other information that continues to be made available over the electronic network. A researcher or student of Greek will be able to search the *Canon* of Greek authors, browse and search the texts of the *TLG*, see an im-

7. Information about SGML and XML can be found at URL: http://www.sil.org/sgml/sgml.html. For SGML and the Text Encoding Initiative see Sperberg-McQueen & Burnard 1990.

8. For XML see URL: http://www.sil.org/sgml/xml.html.

9. For the Unicode, see URL: http://charts.unicode.org/Unicode.charts.normal/U1F00.html. HYPERLINK http://charts.unicode.org/Unicode.charts.

10. See Jeffrey Rusten's review on Palatino Unicode with Polytonic Greek, BMCR 98/11/11 (gopher://gopher.lib.virginia.edu:70/00/alpha/bmcr/v98/98-1-11).

age of the original papyrus or papyri that preserved the work, ask for a morphological analysis of each form in the text, look up words in the electronic dictionary or read a translation of the text. All this comes together with the wealth of visual information offered by *Perseus* and all the other web accessible projects, including the ability to search the holdings of remote libraries and on-line bibliographies and electronic publications. The use of these resources is already having an enormous impact on the teaching of classics since students are more likely to develop an interest in ancient languages, history, and archaeology now that, thanks to computers, we can bring the ancient world to life.

Let me conclude by saying that the goals of the *TLG* project today are not much different compared to those set by the participants of the first *TLG* Planning Conference 25 years ago. The *TLG* was established for the purpose of creating the electronic materials that would facilitate a number of other research pursuits, whether of philological, linguistic, historical or lexicographical nature. I hope we can continue to do so by expanding our databank and collaborating with other projects while seeking more resourceful adaptations of technology to explore the potential of our texts. I suggest that technology has given us a way to bring "Nous into Chaos", if I may paraphrase Diels words, but it will be up to us from now on to make the best use of it.

References

BERKOWITZ, L. 1993. Ancilla to the Thesaurus Linguae Graecae: The TLG Canon. In *Accessing Antiquity: The Computerization of Classical Studies*, ed. J. Solomon, 34-61. Tucson: The University of Arizona Press.

BERKOWITZ, L. & K. SQUITIER. 1990[3]. *Canon of Greek Authors and Works.* New York: Oxford University Press.

CRANE, G. 1998. New Technologies for Reading: The Lexicon and the Digital Library. *Classical World* 92: 471-501.

JOHNSON, W. 1994. Towards an Electronic Greek Historical Lexicon. *EM* LXII 2: 253-258.

MACKAY, P.A. 1996. The Greek Typeface Ibycus for TEX. In *Greek Letters: From Tablets to Pixels*, ed. M.S. Macrakis, 221-229. NewCastle, Del.: Oak Knoll Press.

OATES, J.F. 1993. The Duke Data Bank of Documentari Papyri. In *Accessing Antiquity: The Computerization of Classical Studies*, ed. J. Solomon, 62-72. Tucson: The University of Arizona Press.

RUSTEN, J. 1996. Greek Fonts and Keyboards in the United States. In *Greek Letters: From Tablets to Pixels*, ed. M.S. Macrakis, 204-215. NewCastle, Del.: Oak Knoll Press.

——. 1998. Palatino Unicode with Polytonic Greek. *BMCR* 1998.1.11.

SCHOTTENLOHER, K. 1989. *Books and the Western World: A Cultural History*, translated by W.D. Boyd & I.H. Wolfe. Jefferson, N.C.: McFarland.

SPERBERG-MCQUEEN, M. & L. BURNARD, eds, 1990. *Text Encoding Initiative: Guidelines for the Encoding and Interchange of Machine-readable Texts.* Chicago & Oxford.

SQUITIER, K. 1987. The TLG Canon: Genesis of an Electronic Data Base. *Favonius* 1(suppl.): 15-20.