THE *TLG* DATA BANK, THE *DGE* AND GREEK LEXICOGRAPHY *

The writers set out to explain from the perspective of lexicography both the great advantages and various disadvantages derived from the use of the *TLG* in the making of the *DGE*. Also, some proposals are put forward on the future development of the *TLG* project, now that the data bank is nearing completion.

I. Introduction

The *TLG* data bank has become already a work instrument of great importance for the study of antiquity on the various subjects which have as starting point ancient Greek texts and language. Some uses of the *TLG* were known or guessed years ago, when the project was in its first steps. Other uses have been discovered later, when the texts began to be distributed among scholars, first in magnetic tapes and then in CDROM.

With the help of proper programs, scholars (not always acknowledging it) use the *TLG* data bank not only to locate quotations, search

* Paper read at the International Conference *Thesaurus Linguae Graecae et Thesaurus Linguae Latinae: New Directions in Greek and Latin Lexicography*, Irvine 17-18 december 1993. We thank professors Theodore Brunner and William Johnson, members of the *TLG* team, and Patrick Sinclair, professor in the Dept. of Classics at UCI, for their invitation and for the interest shown in our project. Other bibliography on this subject by members of the team of the *DGE* includes the following articles: J. Rodríguez Somolinos and Ignacio Álvarez, «Informática y lexicografía: la experiencia del *Diccionario Griego-Español*», EMERITA 59, 1991, pp. 81-99; J. Rodríguez Somolinos, «La lexicografía griega en los últimos años», *EC* 33, 1991, pp. 83-118; F. R. Adrados and J. Rodríguez Somolinos, «The Diccionario Griego-Español and Byzantine Lexicography», *Jahrbuch für Österreichische Byzantinistik* 42, 1992, pp. 1-11 (paper read at the Round Table *Thesaurus et lexica linguae graecae medievalis* of the XVIII. Internationalen Byzantinistenkongress, Moscow 1991); E. Gangutia, «El volumen III del *DGE*. Previsiones futuras», *Actas del VIII Congreso de la Sociedad Española de Estudios Clásicos (Septiembre 1991)*, vol. I, Madrid 1994, pp. 109-113.

loci paralleli, identify literary papyri or to create a *corpus* of texts as a base of a philological, literary or linguistic study. Scholarly literature based on *TLG* covers also fields as thorny as textual criticism, statistical studies for purposes such as authenticity questions or chronology of literary works, morphological or syntactic studies with research or pedagogical aims, etc.

But lexicography is its most quoted area of use, even when there are not too many purely lexicographical works or projects which proclaim their debt to the data bank of the *TLG*. The *DGE* is one of them (another would be the *Lexicon zur Byzantinischen Literatur* directed by professor Trapp at Bonn), and here we are going to speak about it from our experience during the last years. One of the programmers of *software* for the *TLG* CDROM called us its most «heavy user», because our remarks to his programm showed an exhaustive use of the CDROM.

II. Our experience

1. *Before the CDROM*

Our experience started in the time when the texts of the *TLG* data bank were distributed in magnetic tapes. Certainly, it was not a very convenient support, because it required a great equipment and specific programs to which we didn't have access at that moment. With the help of the Centro de Cálculo (Computation Centre) of the Complutense University of Madrid, we started a line of work which made indexes of low frequency words of important authors quite forgotten in dictionaries, such as Galen and Pseudo Galen. These indexes were combed for our dictionary.

2. *The CDROM C*

2.1. Introduction

The incorporation of CDROM C to the tasks of the dictionary has been progressive. First we had to study it, make up our minds about its contents, get and adapt the software, quite experimental and rudimentary at the beginning. We had to evaluate the different ways of profiting by it, from the point of view of the organization of our work and the possibilities offered by the available programs. We also had to train our team for this new working tool for collecting lexicographical data. Nowadays, we use it more systematically, but we have not reached the optimum necessary.

There is a previous questions for us in orden to profit by the CDROM C. The editions selected by the *TLG*, following the advice of a committee of the American Philological Association, frequently differ from the ones adopted by the *DGE*. The criteria for selecting editions, even when similar, are not the same. Besides, the way of classifying authors and anonymous works may be different, even when the edition is the same; the Canon takes also into account works we do not, as for instance, authors after sixth century A.D. or editions consisting of not literal fragments. The Canon also introduces eventually more than one edition of the same work while we make a point of quoting always from a single edition. Obviously, the purposes of the Canon and our List are different.

On the other hand, the references differ in many cases even when the edition followed is the same. In relation to this question, it has been of great utility the Appendix to the third edition of the Canon, where there is a list of the reference systems used for each work in the data bank. On any case, the Canon of the *TLG* remains a very useful tool for us and for Greek philology in general.

All these considerations took us to create an internal concordance between both lists in relation to differences in editions, ascriptions, reference systems, etc. This task, which had been partially made for the CDROM C, will be taken anew for the CDROM D.

In addition to the task of collecting more or less systematically lexicographical data (we shall speak about this soon), another utility of the CDROM for our dictionary is to help us to find difficult and sometimes mistaken quotations, coming, for instance, from previous dictionaries or indexes with obsolete quotation systems or errors (such as Stephanus' lexicon or Xenophon's index). Also it is useful for consulting texts missing in our library.

At the time when we didn't have any means of using the *TLG* INDEX as a searching tool, we planned to make indexes and concordances of several authors or groups of authors who didn't have any. We also adapted the software then available to do massive searches by chronological periods, literary genres or combining both criteria. Our aim was to follow the line of work started when the texts were available in magnetic tapes, but to do all the work by ourselves in PCs.

2.2. The INDEX of CDROM C

This line of work was almost abandoned when there was available a proper software allowing us to acceed to the *TLG* INDEX. This index has become progressively our main tool for searching through the *TLG* data bank. The program we have been using is Searcher, by R. M. Smith and some other colleagues from California University at Santa Barbara. Maybe it is imperfect, but, as far as we know, is the only which allows to acceed to the *TLG* INDEX from PCs. We have solved some of its deficiencies. For instance, we have developed conversion programs from BETA format to our Greek format. This allows us to work with the results of the searches directly in our word processor and to print them in Greek. For CDROM D, we shall use a new program (by the moment called Scriptorium) developed by Randall M. Smith and D. J. Dumont in Windows environment. The first version of this program is still quite experimental in order to be used on a really effective way.

We started employing a single person who collected data from the *TLG* INDEX so that we may add more references to *DGE's* entries. This procedure was due mainly to the fact that we were at the point when the manuscript of the fourth volume (now published: Madrid, C.S.I.C., 1994) was almost edited. This person collected data for this volume and for the fifth one, which was in progress. This procedure was not useful in some way, because it favored the collection of rarities, few in number and sometimes without real interest.

Afterwards, we became persuaded that the direct consulting of the CDROM by the writer in the moment when he is writing up an entry is the most convenient procedure to work. This must be done as the last link of the chain, that is, once the writer has studied the materials he is going to introduce in the entry from other dictionaries, our files and other literature. In that way, his search is less laborious and at the same time becomes a powerful way of control. Of course, this implies the same task made by different people and sometimes applying different criteria. This is a problem that we are now studying in order to enhance the quality of the entries with a certain uniformity.

Now, each of our writers has in front of him or her the xeroxes of other dictionaries, such as *LSJ*, Stephanus, Lampe, etc. and other data from our own readings (in regular files and, from some years ago also in a data base developed by us). Besides, he has two new useful working tools.

The first one is a print of the *TLG* INDEX corresponding to the section of the dictionary he is working on. This index includes, as it is known, all the forms in the texts of the *TLG* data bank and its frequency. This printing was possible years ago thanks to the help of our colleagues of the Istituto di Linguistica Computazionale of Pisa, specially Andrea Bozzi and Antonio Sapuppo. At a certain moment, this task has been made by us as a parellel one to the progress of the dictionary. The writer studies in the CDROM the information associated to the forms in the *TLG* INDEX, selects what he is interested in, and asks the program to search for them. Working from the *TLG* INDEX means more speed and efficiency than any other way of working.

The second tool the writer has is another print, corresponding also to his section of dictionary, of what we have called ISCAPLIG (Spanish abbreviation of Índice Selectivo de los Cien Autores Principales de la Literatura Griega, that is, Selective Index of the Main Hundred Greek Literary Authors). This index, created by one of the members of our staff, through a program developed in Hypercard environment, starts with preposition **διά** (about the middle of fifth volume of *DGE*) and also is being done as a parallel task to the progress of the dictionary. It is based on one thousand hundred works by hundred authors in the data bank. So, it covers about twelve millions words, more than the fourth part of the CDROM. As it is based on the INDEX of the CDROM, it is not lemmatized and doesn't make difference between proper and common nouns. This means that we may have sometimes too much information on a word in a certain author and also that we have to look for certain flexive forms (aorists and perfects) in other sections.

It is a selective, not an exhaustive index. It gathers a maximum of six references of a same form in a single author. When a form appears more than six times, the index points out the number of occurrences and remits to the printed INDEX, lexicon or concordance, if any. Inside each entry, the authors are ordered with two criteria: first poetry and then prose. Inside each group, authors are ordered *grosso modo* chronologically.

This index is conceived as a mean of control of any omissions of important authors of the different periods of Greek literature. This flaw is more common in lexicographical tradition than is usually believed. As in this index appears the number of occurrences in each author, the writer is able to make a selection, taking into consideration what is rare and what is common. As far as possible, the quotations are given with our abbreviations. The writer checks every quotation he decides to include and he adapts them to our conventions.

In a certain way, ISCAPLIG means going back to the line of work aforementioned of making selective indexes. This index, and other similar that we could make in the future (for instance of medical writers or fathers of the Church), has the purpose of sparing work for the writer before consulting directly the *TLG* INDEX and allows its consultation being limited to what is not covered by any other mean.

2.3. Lexicographical interest of the *TLG* INDEX

The lexicographical interest of the *TLG* INDEX is enormous and we can always profit from its consultation. It not only allows to locate rarities (hapax, second quotations, new references of words with few quotations, interesting morphological forms, etc.), but, inversely, it avoids the lacking of things really important, as documentation of first rank authors for words of medium or high frequency. We are concerned with rare words but not obsessed. Rare authors and rare words are the condiments (and sometimes the ornament) of the main dish, whose main ingredients are words of medium and high frequency and important authors: the *TLG* data bank affords us basic ingredients and also condiments for this main dish which is our lexicon.

Our experience teaches us, as it was expected, that hapax and rare words come from rare, late and second rank authors, that is to say, authors not properly studied in greek lexicographical tradition. These authors are included in four great groups: 1. Medical writers and veterinary surgeons. 2. Neoplatonic philosophers and commentators on Aristotle and Plato. 3. Christian authors. 4. Grammarians and commentators of classical authors as Eustathius and Photius. Besides these four groups from which we collect many novelties, there is a great pleiad of authors who afford us from the CDROM lexicographical data of outstanding interest. We mention *exempli gratia* Arist., X., Thphr., Str., Luc., Plu., App., D. H., Arr., comic poets, Attic orators, etc. They are authors that, for several reasons, have not been properly treated in the dictionaries of ancient Greek, even when they have sometimes their own indexes. Consulting the CDROM allows us to grasp better the history of the word, documenting it in different periods and genres.

As we take into consideration more lexicographical data for their inclusion in the dictionary, we discover, apart from rarities, new senses, new uses and new constructions. The consultation of the *TLG* INDEX allows us to locate more easily special morphological forms which are interesting by themselves.

Finally, the CDROM allows us to avoid repeating the same quotations Greek lexica have been taking one from another since the seventeenth century. Now, usually, we can select better ones and in greater number, even if the meaning is the same. Twenty years ago, we talked of making a dictionary twice the *LSJ*. Nowadays, with the CDROM it would be possible theoretically to multiply it almost *ad infinitum*, but this is not our aim. We believe today that an ideal proportion is about three times the *LSJ*, including of course papyri and inscriptions.

The *TLG* INDEX helps to apply, even minimally, statistical criteria to fix relative extension of entries, regarding the total number of quotations in the CDROM. But we are aware that it is very difficult to apply them strictly. Other dictionaries which start from a previous data bank (as for instance the *Trésor de la Langue Française* at Nancy) apply these criteria quite strictly. On the other side, there is an almost unavoidable and quite logical tendency towards short entries being proportionally larger as to the number of quotations than long entries.

2.4. Some problems of the *TLG* data bank

But consulting the CDROM has also some problems, in general and from our own point of view. Some of them come from the existing software. Some programs are better and more sophisticated than others; in general all of them are quite experimental, even when their updates soften their inconveniences. We don't want to be critical with the CDROM because of the insufficiencies of the existing software. Reading the literature in relation to the *TLG*, we notice that some critics see the reality of things only through the program they use.

Certainly, we have sometimes made critics to the *TLG* for not providing any software with the CDROM. Even when this fact adds problems, specially for new users, today we are more moderate in our critic. This is not only because the *TLG* gives information on the existing programs, but because we have become aware of two facts. On one hand, this situation has encouraged the merging of different programs in research centers and commercial enterprises. This situation has revealed positive in our opinion. On the other hand, we must say that we will never be thankful enough to the *TLG* for not having distributed a hyperprotected CDROM, as the CETEDOC or the *Institut National de la Langue Française*, which distribute their disks with a proper software but the information included is impossible to handle out from this software.

Other inconveniences are intrinsic and sometimes of laborious solution. We have already talked about the selection of editions: some of them are not wholly adequate and others may become out of date in the future. This is a problem that we know well and we are aware that updating editions is a never-ending story. In the *DGE* we have accepted the challenge of updating, as far as possible, our canon list and quoting texts by the last adopted edition. Maybe the *TLG* and the APA should take into consideration this type of updating in some cases. For instance, the CETEDOC assumes this fact as something natural to be taken into account in future updates of its CDROM.

In Newsletter 10 (July 1986), the *TLG* noticed that they were studying, at the urging of the APA's Committee on the *TLG*, to consider adding *apparatus criticus* materials to the texts residing in the *TLG* data bank. Apparently, this is a let down project, because we have not read more about it. The adding of *apparatus criticus*, which probably has technical problems, should be considered, from our point of view, a secondary aim, but not to be forgotten.

Concerning the copy mistakes and misprints in editions, we must say only that they are statistically quite reduced in number. The problem for us is that, even when they are usually easy to detect, they make us loose our time and the unfortunate case can be that some ghost-words may enter the dictionary. Curiously, most of the mistakes come from the original edition, what makes manifest the extreme care when typing it into the computer, but also at a certain point it questions the way in which the checking was done. Both mistakes are easy to detect with the help of the *TLG* INDEX and it would be interesting to publish a list of misprints, as the authors of the CLCLT plan to do, after having corrected them in their CDROM, called by themselves an *opus semper perficiendum*.

We also have to say that sometimes the number of quotations that document a given word is a little misleading. We can find the same word quoted by a classical author and a source who quotes the passage or comments on it. This problem is not too big, but supposes an additional critical effort for a work such as ours.

As to the *TLG* INDEX, its first problem in its actual state (we talk later about more ambitious improvements) is the lack of differentiation between proper and common nouns. This is quite important and should have been solved from the very beginning.

Sometimes, the huge mass of materials at our disposal may overcome and disconcert the writers. This forces us to make an extra critical effort in order to discern what is relevant and what is of no value and takes us also to the conclusion that the consultation of the CDROM must be always considered an intermediate step before verifying and completing the information through the consulting of the editions and other printed literature. Very soon we understood that the *TLG* data bank (at least in relation to lexicography, but without doubt to other type of research) is not just a substitute of books. It is something else: it is an intermediate tool with multiple possibilities and the conscious user shall always have to go back to the books, looking for more and more precise information. We dont talk yet about inscriptions and papyri data banks. These are still more problematical texts and must be frequently analyzed in our work and in others taking into account their history, succesive editions, critical problems, notes, translation, etc. In a round table on «Epigraphy and Computers» which took place in the Xth. International Congress of Greek and Latin Epigraphy at Nîmes in France (October 1992) we had the chance of noticing the misgivings and misunderstandings still ruling in some circles about what data banks of ancient texts really are.

III. Future directions

Finally, let us say something briefly about several proposals for the *TLG* data bank. They are based on our experience as users of it and on our knowledge of other data banks such as those of the CETEDOC and the *Institut National de la Langue Française*. We told about updating editions and adding *apparatus criticus* to the texts. We are not going to insist on it.

In our opinion, the greatest possibilities of future stand on the development of the *TLG* INDEX through procedures partially not automatic as well as on a better integration between the *TLG* INDEX and the texts through carefully designed sofware. Our first proposal and most obvious one faces the possibility that the *TLG* INDEX becomes complete, that is, it should include references. As it is known, the *TLG* INDEX includes information about authors and works which document each form, but not the exact references, which have to be located through specific searches. This would mean that all searches would have been already made and getting to the text from the *TLG* INDEX would be done very quickly.

Our second proposal would be to undertake the lemmatization of the *TLG* INDEX and, at the same time or later, the morphological analysis of the forms. Even when lemmatization and parsing are complex processes, there are already many experiences in this sense in the USA and Europe. It is possible to profit by them and even by their programs through agreements of cooperation. Existing automatic dictionaries of ancient Greek feed themselves and much work is already done. The *TLG* INDEX could be passed through one of this dictionaries and then suffer a thorough revision.

These ideas take us to what would be the highest *desideratum* for this data bank. The *TLG* INDEX could actually be made into a great data base where each form would be associated to the following information: complete reference, lemma and morphological analysis and, on the other hand, the data of the canon concerning the date and genre of the authors and works. Another possibility would regard the inclusion also of the reverse form.

Starting from the assumption that in the *TLG* INDEX each form would be associated to all this complementary information, the way of working would be the following: there would exist the possibility of making different types of searches to generate subindexes of any type from which ask for the contexts. The possibilities open to the searcher through the development of proper software are unlimited. The simplest one would be to get in seconds the context of a single reference. The most complex would be to get a complete lemmatized concordance of all the CDROM internally organized in any way. This last possibility would be a great progress towards a whole *Thesaurus* of ancient Greek, if anybody feels the need to do such a thing. Other medium possibilities could be to create almost automatically indexes or concordances of authors, works, periods, genres with different possibilities of internal order (chronology, genre, morphology), as well as reverse or proper nouns indexes with the same possibilities.

IV. Conclusion

Going back to the *DGE,* we want to say as a conclusion that thanks to the *TLG* data bank we are now in a situation that enables us to enhance the quality of the dictionary with a greater uniformity, choosing the quotations with better criteria and reducing the unavoidable arbitrariness and the most evident gaps. We have the greatest hope that, as a parallel to their new ambitious project of conversion of the *Thesaurus Linguae Latinae* into electronic form, our American colleagues continue working in the Greek data bank and improving this genial initiative, without doubt pioneer in its genre.

The *TLG* data bank may put in our hands many materials quite elaborate, organized and refined and become in the future a still more magnificent tool. But we have to say, as a general statement, that what it never will do for us is to provide the interpretation of words (with all the bibliographic consultations and the thinking effort needed) nor the organization of entries, a basic question for our dictionary. We believe the *DGE* meets a different need, and also primordial: the updating of general dictionaries of ancient Greek for the benefit of classical scholarship and scholars on other subjects.

FRANCISCO R. ADRADOS
JUAN RODRÍGUEZ SOMOLINOS