Beyond *LSJ*: An Open-Ended Database on Classical Greek

GREGORY CRANE

Tufts University

There is nothing especially new about the application of computing technology to the study of Greek. Theodore Brunner began work on the *Thesaurus Linguae Graecae* a quarter century ago. My own efforts to bend the technology to the needs of our field began in 1985, and I began planning for what would become the *Perseus* Project twelve years ago, writing my first grant proposal in the summer before I became an assistant professor.[1] A lot of work has been done, but quantitative changes –in the amount of data available, in the speed of machines, the type of software– have now begun to have qualitative effects not only on what we, specialists in various aspects of Greek, can do, but also on who can work effectively with Greek. The latter issue is, from my perspective, far more important: it is unclear what the real benefits are from more efficient or wide-ranging research in the classics; it is, however, entirely clear that we can never have too many people engaged in the study of classical Greek. There are still more than 16,000 people enrolled in classical Greek language courses in the United States, but the percentage of language students studying classical Greek, though always tiny, is down by more than a third since 1980 (from 2,4% to 1,4%).[2] If we can use modern technology to expand the number of those who learn the language of the ancient Greeks, then we will help to solidify that foundation on which everything else rests. Nevertheless, I do not wish to strengthen the often unnecessary and almost always destructive dichotomy between "teaching" and "research". Focusing only on the one or on the other is equally harmful.

Electronic publication has allowed us, for the first time since modern print reference tools took their current form, to redefine in a fundamental way the relationship between Greek texts and their readers. Electronic publication opens up two basic possibilities.

First, we can begin to represent in electronic form some of that expertise that could, before now, only be exploited through the labor of a human brain. We are,

[1] A substantial portion of the Perseus Digital Library is currently available on the world wide web at http://www.perseus.tufts.edu. For a survey of work being done at Perseus, see Crane 1998.

[2] See Huber 1998 and Brod & Huber 1997, 55-61.

of course, a long way from making genuinely intelligent machines, but there are more modest subsets of our knowledge that have been rendered dynamic and given a life separate from the “wetware” of the human brain. Much of our work over the past decade has focused on developing a system that could analyze the complexities of Greek morphology and then using the resulting analyses to create links between different documents. A computer system can, by applying a complex set of rules and drawing on a vast database of stems and inflections, automatically recognize that *οἴσετε* is the 2nd person plural of *φέρω*, thus generating a link between the inflected form in a text and a dictionary headword. In classical Greek, of course, the morphology is daunting: not only can classical Greek verbs, when preverbs are taken into consideration, appear in millions of permutations, but the corpus of Greek includes numerous dialects and extends over more than a millenium. Nevertheless, our automatic system can now handle Greek morphology relatively well. In a world of hypertext, the significance of this system lies in its ability to generate millions of links automatically. Those doing research can ask for *φέρω* and automatically retrieve *οἴσετε*, *συνοίσετε* or any other inflected form, while those struggling through a Greek text can go from inflected form *οἴσετε* to the morphological analysis (second person plural future indicative active) to its dictionary entry (*φέρω*). Simply put, the same links, used in different ways, serve scholars and students alike. The core technology both opens up new research possibilities and allows students to read Greek in new ways. In the past fifteen months, we have served up approximately 371,000 pages of classical Greek, 94,000 searches and 558,000 morphological lookups. Traffic continues to grow: on Wednesday November 5, for example, (the day before I finalized this paper) we served 2,300 pages of Greek, 366 word searches, and 3,731 morphological lookups.

The numbers aside, the long term implications of this ubiquitous Web of electronic resources are profound. The study of classical Greek has been, in some measure, an exercise in futility because virtually none of our students in the English speaking world are able to maintain their hard won skills long after they graduate. When they have finally become partners in their law firms, senior vice presidents in their companies or have otherwise reached a stage of maturity, many of our former students will look to broaden their interests, but, given print tools, virtually none of our students have been able to return to the language of Homer, Plato or the New Testament. The language is so complex that almost no one can pick it up again unaided and caught up in the rush of a busy adult life. We can, however, already reach virtually every office and a growing number of homes via the net. We can now realistically establish a support system that will allow people to return to their Greek in later years, querying Data Bases for linguistic information or dictionary entries and exchanging questions in on-line discussion lists. The prospect of reading Greek in one’s forties and fifties fundamentally enhances the value of learning Greek in one's teens and twenties. We face an opportunity of historic proportions to reestablish the study of classical Greek, unrivalled since printing allowed knowledge of Greek literature to reappear in the West.

Second, we do not need elaborate expert system software to have a major impact on what people can do. The simple translation from print to electronic form can, if properly done, have many positive effects as well. The *Thesaurus Linguae Graecae* was established, for example, because classicists all over the world understood that on-line texts could be searched automatically for strings of letters. You could not search for *φέϱω per se*, but you could look for all words that contained the patterns -*φεϱ*- or -*οισ*- or -*ηνεγϰ*- and thus find various inflected forms of this word. Ultimately, the *TLG* came to be used not only as a source for word searches but as a real digital library, in which one could call up texts by chapter and verse, displaying them on the screen to be read on-line. Those of us with ready access to large research libraries didn't need a CD ROM to read Greek texts in the *TLG*, but very few professors of classics, much less students, have such a luxury and the *TLG* CD ROM gave them ready access to many texts that are not in their local libraries. This means that on a day to day basis students of Greek could work with a wider range of texts than was heretofore possible – by changing "what we can do" (i.e., we can use a CD ROM to look up texts) we change "who can do what" (i.e., people outside research institutions can call up all of Galen, for example).

But if source texts are our primary material, they benefit from a range of other resources. Our own work over the past decade at the *Perseus* Project has centered in large measure on developing a network of resources that could help people work with Greek texts that were on-line. Central to this effort has been producing electronic versions of the standard print Greek-English lexica on which we all rely. Both the (soon to be superseded) *Intermediate Liddell-Scott Greek-English Lexicon* and the monumental ninth edition of the *Liddell-Scott-Jones Lexicon* are in electronic form and currently available at our web site.

The drawbacks of the large research lexicon are obvious to anyone who has used it. The codex was not designed to hold 35,000,000 characters worth of data. The book is too heavy and the print is too small. *LSJ* lexicon entries are notoriously hard to read. There are very few clues such as highlighting or blank spaces whereby a reader can parse out the clean, hierarchical outline of the entry or scan for citations to Homer or whatever author the reader is studying. The research lexicon is so cumbersome that virtually all students work with either the *little Liddell* or the *middle Liddell*, essentially abridgments of nineteenth-century editions of *Liddell-Scott*. The *middle Liddell* (which is on-line on our server) was first published more than a century ago – the classics-hating Winston Churchill might have seen one of the first copies as a schoolboy at Eton. Virtually all students of Greek have relied on these embarrassingly out of date lexica ever since. The research lexicon has not been reedited in fifty years – in part because the market for this resource is so small that it no longer justifies the expense. The student lexicons have languished for more than a century at least in part because few classicists are willing to dedicate much of their lives to a pedagogical lexicon – this is thankless work and not sufficiently well rewarded.

When we put the research lexicon on-line, we planned to make it easier to use. We intended (and still intend) to let people filter information, allowing them, in effect, to generate abridged lexicon entries on the fly. We did not anticipate, however, precisely how much difference the simple shift from print to electronic format would make. When readers look at a lexicon entry on the web site, space is not nearly as crucial as with the book. The font in the electronic version can be larger. We can put blank lines between definitions and can even usually represent the hierarchical structure of the lexicon by means of indentation. We can use bold, italics, and even colors to highlight different categories of information. And readers can use any functional[3] web browser to search for strings such as “Hom.” or “Soph.” to find definitions relevant to the author at hand.

The consequences of this simple change have been far greater than we had anticipated. In the past fifteen months (July 9, 1996 through November 5, 1997), the on-line research and student lexica have been accessed 407,475 and 280,377 times for a total of 687,852 dictionary lookups. Although, insofar as we can determine, the numerical majority of our users are intermediate students of Greek, the research lexicon is 45% more heavily used than the student lexicon. Although we have made only the most rudimentary efforts to make the research lexicon more usable, it already has more than overtaken its abridged version.

The consequences could not be more profound. We could now have a single lexicon that could serve both researchers and students. If we go to the trouble of rigorously tagging the structure of the dictionary entries, then we could generate abridged versions on the fly so that those interested in an overview of the word could view a synopsis of its meanings or could see only definitions appropriate to a given period or style of Greek. There are, of course, good arguments to have a separate lexicon that is geared to the needs of students and not simply an abridgment of a scholarly lexicon, and I am delighted to hear that we may get such a tool. Nevertheless, having a single, flexible resource that serves everyone from researchers to second year Greek students has tremendous possible benefits. Students can get access to the same Data Base as do scholars, and if the lexicon is kept up-to-date (as can be done much more readily in electronic form), the stu-

[3] At the time of this writing, bugs in some versions of both Internet Explorer and Netscape Navigator made it impossible to search for words on some web pages.

dents will also benefit – at the very least, they won't be reading century old definitions. Scholars, though, will benefit at least as much: it is obviously much easier to maintain a reference work with a potential audience of (in the U.S.) 17,000 students and faculty than simply 1,000 faculty by themselves.

Many of the citations in *LSJ* are now dynamic links: click on "Hom. Il. 5.303" and you will call up the Greek text of Homer, *Iliad*, book 5, line 303. Now these links point to the 3.4 million words of Greek in *Perseus*, but they could also point to texts in the *TLG* if that is available to the system. We can now look up the passages that supposedly illustrate a given meaning, seeing not just the quoted Greek in the lexicon entry but the full context as well. We thus facilitate a task that we have all done by hand.

But what happens if you are reading Homer *Iliad* 5.303? There is no way to determine, given conventional print technology, that the *LSJ* article on *φέρω* has something to say about the usage of *φέροιεν* in this particular line. One can, of course, produce a print index of the citations in *LSJ*, but this would result in an unwieldy book that relatively few people would ever see. In the electronic environment, however, we can automatically generate the "back-links", allowing the reader looking at Homer *Iliad* 5.303 to see that *LSJ* comments on *φέροιεν* and providing a link back to the precise section of the dauntingly large entry on *φέρω*. Simply creating these links from the texts to *LSJ* will help those reading the standard Greek authors: 40% of the 524,000 citations in *LSJ* point to 5% of the surviving corpus of Greek literature: *LSJ* specifically comments on one out of seventeen words in those commonly read texts included in *Perseus*. In effect, when these links have been added, *LSJ* becomes not just a lexicon but a lexical commentary to these texts.

But, of course, we can do the same things with citations in any on-line publication, not just *LSJ*. Thus, if you are reading book five of the *Iliad*, you could see not only which passages were cited in *LSJ* but which passages were cited in Smyth's Grammar, recent issues of *AJP*, or any other properly tagged electronic publication.

There are obvious problems with such automatic links. First, how do you filter the links that will rapidly accumulate around often read passages? This is a special case of the general information filtering problem that we all face, whether on the Web or in a library. While there are no perfect solutions, there are strategies that we can adopt: e.g., show me only links from the following six journals between 1990 and 1995 or do not show me citations listed only in footnotes etc.

Things get much more complex, though, if we start wondering how this increased connectivity affects the way in which we write. Traditionally, if I write a commentary on Sophocles' *Oedipus Tyrannos*, most people are going to get to that commentary via the text. If they are reading line 238 of the text, they will look for notes relevant to that particular line. I can assume that my audience will consist primarily of *Oedipus* readers, and I can write accordingly. But now anyone reading any passage that I cite can find my comment. If I can reach a more general audience, should I change the way I structure my information? Imagine the increased usefulness of scholarly notes, now tucked away in journals, if readers of a *TLG* text, for example, could call them up directly.

Consider the problem of reference works. We can update an electronic *LSJ* much more readily than we can its print counterpart, but the role of the lexicon itself changes. A reader curious about the meaning of *αἰδώς* should find attached to the *LSJ* entry a link to Douglas L. Cairns' monograph on this term and should then be able to call up the text of that monograph. I can imagine a genre of "lexicon entries" separate from *LSJ* or any monolithic reference work, published in a range of journals or monograph series, that would be designed to update, augment or even supplant the central *LSJ* entry. Even given the relatively primitive technology of the web as it stands today, we can envision a "virtual lexicon", with entries stored in many different sites and judged by many different editorial groups. There is an inherent untidiness to such an arrangement – but libraries are themselves inherently untidy. The fundamental shift is this: in a digital world, the hermetic closure of the codex diminishes and the distinction between book and library begins to blur.

But what should dictionary entries look like? Since electronic publications can readily accommodate images and even interactive 3D reconstructions of places and things, a modern Greek lexicon should certainly have many links to visual information. Nevertheless, it is by no means clear how we should organize dictionary entries. The hierarchical dictionary entry, with its neat outline format, is easy to read, but this structure does not reflect the way that our brains store lexical information. Advances in our understanding of words in the brain may allow us to radically redesign, if not to replace altogether, the form of individual dictionary entries. Certainly, we need to do a great deal more to link semantically related words within the dictionary: contemporary tools such as George Miller's *WordNet* and traditional lexical resources such as Pollux's *Ονομαστικόν* (essentially, a list of semantic fields) point in the same direction.

Our work now stands at a transitional phase. We have spent more than a decade developing the core technologies for handling Greek morphology. We have entered Greek texts, translations into modern languages, lexica, grammars, commentaries and other categories of information – embedding Greek texts in a heterogeneous digital library. But, of course, progress raises questions of its own and we have ever more work to do.

Our currently funded efforts include:

First, we are enlarging our Data Base of resources. We have entered Smyth's *Greek Grammar* and may enter the massive *Kühner-Gerth Grammar* as well. We are entering a series of commentaries both because of their intrinsic value and because integrating these commentaries with the text, grammars and lexica will pose important questions of document design. We will learn what happens when a dense set of commentaries point to a small set of texts, concentrating on Pindar and Sophocles.

Second, we are collaborating with other Data Bases so that electronic resources, developed by different projects and available at widely disparate points, work smoothly together. We have already created a webversion of the *Duke Databank of Documentary Papyri*, and we are collaborating with our colleagues at the *TLG* as they prepare to create a web server for their own materials. In adding citation and morphological links to books published by Johns Hopkins Press and now available on their web site, we have made the first step towards adding new functionality to the growing number of scholarly publications now available on the web from publishers like Johns Hopkins, the University of California Press, and others.

Third, we are beginning to prototype new types of publication. Simply replicating print documents in electronic form is clearly just a first step. Converting citations to links, tying Greek words to our morphological Data Base and similar tasks carry us a bit further, but they only overlay the surface of an existing structure. Archaeological publication clearly stands on the brink of revolution, as sites gain the capability to publish Data Bases, CAD drawings, or interactive VRML reconstructions, but even publications about words and language will surely evolve as well, embedding interactive links to source texts or searches or drawing upon advances in the cognitive sciences.

Fourth, we are trying to develop a new editorial process aimed at supporting the creation of electronic resources. A major goal of this effort will be to bridge the gap between research and general publications: we want to develop new publications that can, in different ways, serve both scholars and a more general audience – again, not just “what we can do” but “who can do what”. The world wide web allows us to reach millions of machines instead of thousands of libraries, but we need to rethink the way we write if we are to take proper advantage of these new possibilities. We need abstract editorial standards, exemplary documents, reasonable policies for the use and reuse of intellectual property, and, above all, a growing community of collaborators. We have been fortunate to receive support from the U.S. Department of Education to address this specific set of issues and my colleague, Ross Scaife, editor of the Diotima web site on “Gender in Antiquity”, and I are collaborating on this particular effort.

To sum up, the modern (or perhaps postmodern) world offers many challenges to those of us who are dedicated to keeping alive the study of classical Greek. The origins of the West are not fashionable in many intellectual circles, at least in the English speaking world, and the demands of late twentieth century life do not make it easy for us to convince students to dedicate years to the study of an ancient language – why not study Spanish or Chinese? Nevertheless, the electronic world offers opportunities as well as challenges. If we can create a network of independent, but interconnected and mutually reinforcing electronic publications, available for free or at some nominal cost to everyone attached to the net, we have an opportunity to reach out and expand our audience far beyond its current limitations, while revolutionizing our research practices as well.

References

BROD, R. & B.J. HUBER. 1997. Foreign Language Enrollments in United States Institutions of Higher Education, Fall 1995. *ADFL Bulletin* 28: 55-61.

CRANE, G. 1998. The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. *D-Lib Magazine* January.
http://www.dlib.org/dlib/january98/01crane.html

HUBER, B.J. 1998. Variations in Language Enrollments through Time. *ADFL Bulletin* 27.