What's in a corpus? Utilizing metadata in Latin and Greek text collections

Neven Jovanović
University of Zagreb
neven.jovanovic@ffzg.hr

September 22, 2007

For quite a long time now we can use a good number of electronic collections of texts in Latin and ancient Greek [1]. These collections are commercial or not, open-source or not, prepared by specialists or put together by enthusiasts; but, regardless of their origin and aim, most of such collections share one feature: they are not linguistic corpora *sensu stricto*, because they do not — and cannot — meet the criteria for linguistic corpus as proposed e. g. by Sinclair [2]. Accordingly, collections of texts in Latin and Greek seem to be designed — and used — more as libraries or archives than as tools for the study of language. Here I wish to propose some ideas and examples of how to make such collections even more similar to libraries, in order to make them more accessible and searchable, and more widely used. In other words: if collections of texts in Latin and Greek are seen as libraries, let us think about how we design their catalogues.

Most users of Latin and Greek text collections fall into two groups: one comprises those who are still learning to know the cultures inside which the texts were composed, and the other consists of experts in one or more areas of classical studies. Members of the first group do not know inside and out the authors and the texts from the collection; members of the second group tend to ask questions that are complex, complicated, or simply hard.

Confronted with a query interface to a text collection, the learners find themselves looking for an unknown in a sea of unknown (and what they do find is difficult for them to interpret and contextualize). The researchers, on the other hand, find out that the questions they ask do not easily translate into a "brute force word search" of texts in the collection, that "a database organizes information in ways that will facilitate some queries but complicate or rule out others" [3]. As a result, both groups of users feel unsatisfied with the collection, stop using it, use it as a reading tool, or limit their searches to simplest single-word queries.

It seems to me that a collection of texts in Greek or Latin may be enhanced, or reorganized, in a way that will help both user groups at the same time. The learners do not know what is in all those texts; the researchers want to "harvest" very specific things from very specific sets of texts. What both groups need is metadata, and rich metadata at that: annotations presenting bibliographic and analytic description of texts, tagging them for periods, places, literary genres, authors and producers — as well as for summaries of contents and overviews of text structure. Both groups would find useful, for example, a collection of Greek poetry marked by meter (hexameter, iambic trimeter) and genre (epigram, elegy); or a collection of Cicero's, Jerome's, or Erasmus' letters annotated — and, ipso facto, searchable and regroupable — by sender and addressee, by dates, by subjects; or an early 16th century epic poem — such as *De morte Christi* by Damjan Beneša from Dubrovnik, in 10 books and more than 8300 verses (editio princeps 2006) [4] — annotated by narrative units and features (the narrator, the plot, characterization, scenery, aspects of time, speeches).

The task of annotating collections of Greek and Latin texts may seem daunting. A "collection of texts in ancient Greek" would, theoretically, contain everything written in Greek that has survived from the 8th century BC up to 1453, while for a "collection of texts in Latin" the upper age limit would be extended to 20th, even 21th century. Vast ranges of these writings are virtually unknown today; almost all of us are learners in one aspect or another with regard to texts from patristic, Byzantine, neo-Latin literatures. Finally, the community of users could disagree about annotations (e. g. in tagging a passage as being "about Rome", or as "laudatory").

All those problems, however, could be solved. First, we do not have to mark up an electronic *corpus totius Latinitatis* or *Graecitatis*, simply because it does not exist yet; we seem to be just laying groundwork for such collections.

Second, a large source of quasi-metadata is rapidly becoming more available: it is existing scholarship. Especially valuable may be old scholarly and school editions of Greek and Roman texts, with abundant notes and summaries of edited texts, often itself in Latin (e. g. Lemaire's 19th century *Bibliotheca classica Latina*). A move towards summarizing and indexing works of neo-Latin literature in digital editions has already been proposed six years ago by Schibel [5] — but today many useful old editions are digitized and readily obtainable through Google Book Search, as recognized and discussed by Schibel and Rydberg-Cox [6].

Third, basic infrastructure for such annotated collections is also already available: tools such as PhiloLogic; standards such as TEI XML; Greek and Latin texts with structured markup such as those from the Perseus Digital Library; strategies such as Just in Time Markup, which supports "conflicting logical and theoretical interpretations of the authenticated transcriptions of the original source documents while allowing for the continual development of additional editorial material" [7].

Finally, there are enough existing communities of users which could collaborate in tagging (what Mahoney [8] calls "social support" for a text system): such a community may be, for example, an university course on Cicero's letters. Reading a number of letters, students may add markup to them (converted e. g. from Lemaire's 1827 edition [9]), working together with a "corpus editor" — an expert who manages "a collection of materials that are thematically coherent and focused but are too large to be managed solely with the labor-intensive techniques of traditional editing" [10] — and using the Just In Time Markup system for standoff markup to avoid "markup pollution" of the text being tagged.

Several projects have already enriched Greek or Latin texts with rich metadata and semantic annotations: the Chicago Homer tags narratological features in the corpus of Early Greek epic; the Perseus under PhiloLogic and Peter Heslin's Diogenes enable "bibliographic searches" in their respective corpora; the Vindolanda Tablets Online and Inscriptions of Aphrodisias allow browsing or searching their documents collections by subject, category, type, places, historical and archaeological context (incidentally, these two projects are based on existing scholarship — they are digital conversions, or offshoots, of printed books); the CAMENA neo-Latin project carefully digitizes early modern textbook and reference works, which are in themselves sources of metadata.

As a starting point for further discussion, I would present a prototype of the Croatiae Auctores Latini database (CAuLa) — a digital collection of medieval and neo-Latin texts by Croatian authors, a collection currently being assembled by the Croatica Neolatina et Mediaevalia project at the University of Zagreb, Croatia. The CAuLa database will have an interface which will, by means of metadata from the CAuLa texts, help users both orientate themselves in the unfamiliar material, and reorganize or search this material according to their research questions.

References

[1] Digital Classicist contributors, "Greek and Latin texts in digital form," *The Digital Classicist.* 14:09, 21 June 2007, <http://wiki.digitalclassicist.org/index.php?title=Greek_and_Latin_texts_in_digital_form&oldid=2096> [accessed 22 September 2007]

[2] Sinclair, John, "Corpus and Text — Basic Principles", *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne. Oxford: Oxbow Books, 2005: 1-16. Available online from http://ahds.ac.uk/linguistic-corpora/ [accessed 22 September 2007].

[3] The Chicago Homer editors, "Understanding the Chicago Homer", *The Chicago Homer*. <http://www.library.northwestern.edu/homer/helpbotunderstandframeset.html> [accessed 22 September 2007].

[4] Beneša, Damjan, *De morte Christi*, ed. V. Rezar. Zagreb: Ex Libris, 2006.

[5] Schibel, Wolfgang, "Digitale Medien und editorische Strategien im Bereich der neulateinischen Literatur", *Neulateinisches Jahrbuch 3* (2001): 249–258.

[6] Schibel, Wolfgang and Jeffrey A. Rydberg-Cox, "Early Modern Culture in a Comprehensive Digital Library", *D-Lib Magazine*, March 2006, Volume 12 Number 3. `doi:10.1045/march2006-schibel`.

[7] Australian Scholarly Editions Centre, "Just In Time Markup", *Australian Scholarly Editions Centre Projects*, <`http://www.unsw.adfa.edu.au/ASEC/JITM/publications.html`> [accessed 22 September 2007].

[8] Mahoney, Anne, "Creating an Infrastructure for Scholarly Publication On Line", paper presented at Ancient Studies - New Technology. December, 2000. Available online from `http://www.perseus.tufts.edu/cgi-bin/ptext?doc=Perseus%3Atext%3A2000.06.0010` [accessed 22 September 2007].

[9] *M. T. Ciceronis (...) Epistularum omnium libri*, ed. N. E. Lemaire, Paris: Lemaire, 1827. Available online from `http://books.google.com/books?id=_lQQAAAAIAAJ` [accessed 22 September 2007].

[10] Crane, G. and J. A. Rydberg-Cox, "New Technology and New Roles: The Need for Corpus Editors", *The Fifth ACM Conference on Digital Libraries*. 2000. San Antonio: ACM. Available online from `http://www.perseus.tufts.edu/Articles/corpused.html` [accessed 22 September 2007].