New Technology and New Roles: The Need for "Corpus Editors"

Gregory Crane, Jeffrey A. Rydberg-Cox

The Perseus Project
Tufts University
Medford MA 02155, USA
E-mail: {gcrane, jrydberg}@perseus.tufts.edu

ABSTRACT

Digital libraries challenge humanists and other academics to rethink the relationship between technology and their work. At the Perseus Project, we have seen the rise of a new combination of skills. The "Corpus Editor" 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. The corpus editor must possess a degree of domain specific knowledge and technical expertise that virtually no established graduate training provides. This new position poses a challenge to humanists as they train and support members of the field pursuing new, but necessary tasks.

KEYWORDS: Editing, hypertext, corpus linguistics.

INTRODUCTION

A continuous tradition of editing has existed in Western culture since the alphabet emerged and Homeric epics were committed to writing more than 2,500 years ago. Editors prepare materials for dissemination to a wider audience. Their aims may range from accurate transcription of a preexisting source to a massive edition, with, for example, hundreds of pages of information about a single short play. The rise of electronic publication has allowed editors to expand their traditional goals, e.g., creating dynamic texts [1, 12], publishing color facsimiles with unprecedented detail and accuracy [6] and designing publications of time based media such as music and film (e.g., [2, 7, 16]). In fact, the literature on editing has just begun to grapple with the issues raised by electronic publication [11]. Editions of this sort make innovative use of technology, but they are also traditional in scope; one or more scholars carefully work over each page and word, tagging and structuring the data but doing so for the most part by hand.

The electronic environment also allows for the creation of very large collections of materials. These projects (like *Project Gutenberg* [8], the *Making of America* [15], the *Thesaurus Linguae Graecae* [9] and, to some extent, the Perseus Project [5]) have concentrated on building up large, but lightly tagged corpora with little annotation or commentary. (See, for example, the discussion of a "clean corpus policy" in [13]). Given the grand task that we in the humanities face — converting as much of the human record as possible into digital form — such massive efforts, with an emphasis on bulk, accurate transcription, and basic metadata are crucial.

In our own work building and extending the Perseus digital library on the ancient Greco-Roman world we have begun to encounter a third type of project, midway between "handcrafted" editions and large corpora. These "corpus editions" are thematically coherent collections of materials that are too large to allow for the minute scrutiny normally expected in a print work. These collections, however, are small enough to allow for the tagging of some material beyond the simple structure of a document. An editor who combines domain specific expertise and technical skill can in many cases produce substantial bodies of materials that are immediately useful, that lay the foundations for future "handcrafted" editions, and that contain more scholarly apparatus than is feasible in a large corpus.

THE CORPUS EDITION

The corpus edition begins at the point where an editor must rely on mechanized processes that can no longer be manually proofread. Every well-planned traditional edition has a set of goals. An editor can compare various manuscripts against one another and record the variants. An editor may gloss difficult words or provide a definition for every single term in the text. An editor may even provide "variorum" coverage, striving to read through and summarize all scholarship relevant to a given work. Many of these tasks involve judgement calls on which experts can differ (e.g., which scholarly ideas merit citation in a variorum edition?), but the scrupulous editor will traditionally review and ponder each such judgement individually. In a corpus edition, clearly defined goals are equally necessary, but such individual judgements are not feasible; the editor must establish a reasonable level of precision that balances scholarly standards with the need to quickly digitize large quantities of materials. It is also incumbent upon the corpus editor to clearly document the level of precision employed so that users know what they can expect when reading documents in that corpus.

The corpus edition ends at the point where the editing does

not require specialized knowledge of a field. Determining which phenomena to tag in a text is a major task and only those committed to the domain will be able to determine a widely accepted balance between utility and effort. An editor may — after considerable study and extensive discussions with readers and scholars — develop stable, well-defined categories of information that should be tagged. In a corpus edition, this methodology must be generalized so that it can be applied to all of the texts in the corpus in a scalable way. This first level tagging can be performed programmatically with a text processing language such as PERL or by the data entry firm that is entering the text.

In practice, most texts are complex and individual decisions have to be made about some tags. A program or data entry firm can identify elements such as chapter headings or speakers in a play if they are printed in a consistent way (i.e. speaker names are always listed at the beginning of a line and followed by a colon). However, this combination of formats may also appear as a regular feature of a line within a play and, thus, be mistakenly identified as a speaker. Corpus editors must proofread their texts to eliminate this type of problem. However, it is not possible or desirable for the corpus editor to check every tag in the document. The corpus editor will choose the types of tags that should be checked and corrected by hand and the tags that should be left "as is" so that this large body of material can be published quickly and support a great deal of present study. After the corpus has been established, it is then possible to make the collection of documents available so that other scholars (or even the original corpus editor) can refine and handcraft parts of this corpus according to their scholarly interests.

This does not mean that corpus editors must confine themselves to tagging only the surface structure of a document while leaving other (perhaps more interesting) elements for later editors. Automatic tagging of elements in the corpus that reflect the needs of its scholarly field is one of the key characteristics of the corpus edition. A great deal of recent research has been done in the area of knowledge extraction from unstructured text. The same systems that provide good results for Wall Street Journal articles, however, provide less satisfactory results for many types of texts in the humanities. The corpus editor must, therefore, know enough about both computational algorithms and the documents in the corpus to adapt these techniques for the corpus in question (see, for example, [10]). The corpus editor may ultimately do research in some area of knowledge management and feature extraction but the focus would be on adapting these techniques for the corpus and the scholarly concerns of her or his discipline.

Corpus editions are, thus, cumulative and dynamic, sharply distinguishing between tags that are hand-edited and tags that are dynamically applied. Corpus editors might, as a matter of course, check to make sure that each chapter and section within an electronic document is properly formatted. They may also develop systems to automatically tag features that are important for the study of these documents and even hand tag some elements of a text. The corpus editor does not, however, consider or proofread every tag in every text in the corpus. This allows for the creation of substantial bodies of materials that are both immediately useful and important foundations for future "handcrafted" editions.

EXAMPLES OF CORPUS EDITIONS

The distinction between handcrafted editions, corpus editions, and large corpora reflects our experiences building and extending the Perseus digital library over more than ten years. In some cases, we have enjoyed the luxury of being true corpus editors: experts in the field of classics with the technical skills to accomplish particular domain specific goals. The Greek and Latin texts in Perseus constitute a corpus edition tied together by software that analyzes complex inflected forms, mapping them onto linguistic analyses and dictionary entries [3,4]. This corpus is part of a widely used digital library that is available on both CD ROM and the World Wide Web. This system could not have been created without both scalable methods of tagging and specialized knowledge of classical languages and literature.

In the past four years, we have begun to move beyond classics to study more generally the problems of digital libraries in the humanities. Current projects include Archimedes (a digital library on the history of mechanics developed in conjunction with the Max Planck Institute for the History of Science in Berlin), an electronic edition for the New Variorum Shakespeare Series (in conjunction with the Modern Language Association), and a digital library on the City of London (in conjunction with the Tufts University Archives) [14]. In these projects, we have played the role of technical experts who understand the demands of scholarship in the humanities. Our experiences with our colleagues outside of classics have, however, driven home to us the need for corpus editors whose intellectual centers of gravity lie firmly within their fields of academic expertise.

CONCLUSION

Our underlying argument is hardly new: technical and academic expertise need to be brought together. Nevertheless, the humanities has few mechanisms to train corpus editors and support their work. The corpus editor requires a combination of technical and traditional humanistic expertise for which existing graduate programs and professional pathways are not well prepared. To cope with the rising demand for such scholars, we have begun supporting postdoctoral scholars — a practice common in the sciences and rare in the humanities. Postdoctoral positions will not alone fill the gap. Humanists need to consider more formally the best ways to train and support scholars who work on corpus editions. While much of our work at the Perseus Project concentrates on the design of digital libraries for the humanities, we have found ourselves increasingly contemplating the design of humanists and their training.

REFERENCES:

1. Binda, H., *Hell and Hypertext Hath No Limits: Electronic Texts and the Crises in Criticism.* Early Modern Literary Studies, 2000. **5.3**: p. 1-29.

2. Braunmuller, A.R., *Macbeth*. 1994, Santa Monica: Voyager Company.

3. Crane, G., *Generating and Parsing Classical Greek.* Literary and Linguistic Computing, 1991. **6**: p. 243-245.

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

5. Crane, G., *The Perseus Project: An Evolving Digital Library.* February, 2000. (date accessed). http://www.perseus.tufts.edu.

6. Eaves, M., R. Essick, and J. Viscomi, *The William Blake Archive.* February, 2000. (date accessed). http://jefferson.village.virginia.edu/blake/index.html.

7. Goodrum, A., B. O'Connor, and J. Turner, eds. *Computers and The Humanities: Special Issue on Digital Images* . 2000, Kluwer.

8. Hart, M., *What is Project Gutenberg?* Febraury, 2000. (date accessed). http://promo.net/pg/history.html.

9. Pantelia, M., *The Thesaurus Linguae Graecae.* February, 2000. (date accessed). http://www.tlg.uci.edu/~tlg/.

10. Rydberg-Cox, J.A., *Lexical Acquisition and Word Collocation in Ancient Greek Texts.* Literary and Linguistic Computing, 2000. **15**(2).

11. Shillingsburg, P.L., *Scholarly Editing in the Computer Age : Theory and Practice*. 1996, Ann Arbor: University of Michigan Press.

12. Siemens, R.G., *Disparate Structures: Electronic and Otherwise: Concepts of Textual Organisation in the Electronic Medium, with Reference to Electronic Editions of Shakespeare and the Internet.* Early Modern Literary Studies, 1998. **3.3**: p. 1-29.

13. Sinclair, J., *Corpus, Concordance, Collocation.* 1991, Oxford ; New York: Oxford University Press.

14. Smith, D. and J.A. Rydberg-Cox, *The Perseus Project: A Digital Library for the Humanities.* Literary and Linguistic Computing, 2000. **15**(1).

15. *The Making of America Collection.* February, 2000. (date accessed). http://moa.cit.cornell.edu/MOA/moa-mission.html.

16. Winter, R., *Multimedia Beethoven: The Ninth Symphony, An Illustrated Interactive Musical Exploration.* 1991, Santa Monica: The Voyager Company.