Drawing Knowledge from Information: Early Modern Texts and Images on the TAPoR Platform

Claire Carlin
University of Victoria
ccarlin@uvic.ca

Abstract

This paper enumerates the technologically naïve humanist's wish list for the type of text analysis tools to be developed by the TAPoR project.

KEYWORDS: Annotation, anthology, navigability, searchability, usability.

Introduction

The CaSTA conference was for me an opportunity to learn more about text analysis in general and TAPoR in particular. I presented the following paper without insights to offer on technology; rather, I spoke as a scholar in the Humanities with a wish list.

1. An anthology of documents on Early Modern marriage

1.1 Beyond print

My interest in TAPoR has been keen since Michael Best and Peter Liddell first solicited expressions of interest from the Humanities faculty at the University of Victoria in 2002. I came to these discussions with what was immediately deemed a "traditional" project -- and with very little technical knowledge. Currently, I am finishing a book about representations of marriage in early modern France. When I applied for the SSHRC grant that supported the research for this book, one of the evaluators suggested that I consider preparing an anthology of the texts I was studying, since most of them have not been republished since the early modern era.[1] A print anthology could only contain a small sample of the numerous lengthy extracts that I had patiently keyboarded into my laptop while in Paris libraries, where scanners are not allowed. When TAPoR came along, I began to take the idea of an anthology much more seriously because the possibilities of the platform immediately expanded my conception of what

this sort of collection could become. Early on, my colleagues and I were shown examples of what TAPoR might mean for Humanities computing: the Graves diary project, the Lydgate manuscripts project, the Internet Shakespeare editions that Michael Best has developed, and now the Old English Project; all are inspirational. (See the list of Works Consulted at the end of this article.)

It was apparent that my material could benefit from this opportunity in several ways: I could (at least theoretically) include the hundreds of pages I have available, I could include images and music, there could be hypertextual links and a rich embedding of inter and intratextual materials which could be annotated, not just in the way a scholarly edition is annotated but in a manner that would allow users with varying goals to use it differently. Ray Siemens' plenary talk at the CaSTA conference, published in this volume as "Text analysis and the dynamic edition? Some concerns with an algorithmic approach in the electronic scholarly edition," explores the issue of "usability" and the way most of us read online, having been schooled in print culture. As the act of reading evolves, the sort of electronic anthology I envisage should offer different ways of reading and different ways of knowing, as exemplified by the now common experience of reading a website full of hypertextual bells and whistles.

1.2 Beyond the Web

TAPoR is of course intended to be more than a website. Indeed, when discussing the TAPoR project with colleagues, one question that often comes up is "Why not just create a website for your texts? What's so special about TAPoR?" Although the material developed for TAPoR could of course be placed on a website, what excites me is that the Text Analysis Portal has as its primary goal to develop new tools for working with texts. Despite the sophistication of what we find on the Internet, I have yet to see a website that does what I would like to do with the material I have collected, especially in the area of searchability.

1.3 Searchability across document types

Interactive word searches are a first step, but I would like to offer the option of a search linked to themes or concepts developed in my introduction to the collection. Now that the wonderful world of electronic text analysis allows us to create powerful concordances, it is possible to see

the steps leading to a level of performance I (at least) would have found difficult to imagine before TAPoR. From word frequency lists, to multiple ways of sorting (alphabetical order, frequency, location in text), to phrase identification, we are arriving at a point where we can indeed examine a word within a given context as well as develop rules for weighting words or phrases within said context.

I would like to be able to enrich background material with digital cross-referencing that would include scholarly notes on people, places and other information about the society in which these texts were created. My corpus includes satire, conduct manuals both religious and secular, medical treatises, legal decisions, as well as literary texts (fiction, drama and poetry), and each genre of text raises different questions because each was constructed according to the expectations of very different groups of early modern readers. Many of the books I have used have frontispieces that deserve analyses all their own since they were designed to summarize the contents of a three to four hundred-page volume on one leaf.[2] Not surprisingly, art historians have not worked on most of the rather obscure frontispieces I am studying; the interest lies in the juxtaposition of a good number of texts and illustrations. Well-known engravings of time should also be included, and annotated; many of them have verse captions. This sort of image is widely disseminated in current scholarly publications, unlike the frontispieces, but I have come across some engravings that do not appear to have been published before and whose provenance I will be able to research at the Cabinet des Estampes of the Bibliothèque Nationale de France where I found them originally on microfilm, largely unattributed. Facsimile pages from marriage manuals that contain printed marginal annotation, common at the time, could also be displayed in order to illustrate the seventeenth century's version of textual linking. Bringing these diverse document types together can create a resource that could serve students and other researchers interested not just in the topic of marriage, but in any of the genres of texts or images on the platform, in the authors or in the institutions they represent -- for example. Concepts, thematic threads or "habitus" could be traced across genres; possibilities include the notion of duty, problems related to procreation or issues in household management. New perspectives will be born from the enrichment of context.

Besides the different sorts of contextualization called for by these document types, I would also like to include the annotated database of scholarship on early modern marriage that my research assistant prepared for the book I am writing. This bibliography could be a wonderful

resource for scholars, along with all of the other elements I am imagining for my project. If we can create a setting where multiple windows allow for several perspectives on a document to be visible at one time, we will have created a whole new context, indeed a whole new document as its reception/interpretation by the reader/researcher shifts with each additional window. In a talk given at U Victoria in June 2003 to launch a series of summer Humanities Computing workshops, Ray Siemens described "melding dynamic textual analysis with hypertextual linking," an exciting prospect for novices in humanities computing. The possibilities do indeed seem endless.

2. Text Analysis Tools

2.1 Naïve dreams?

As I describe my ambitions, I realize that I may be letting my imagination run away with me; this could be a life-long enterprise. Ray also suggested in the talk he gave in June 2003 that providing too much material can create problems of navigability. But I prefer to think big. In my optimistic moments, I assume TAPoR will allow me to meet these challenges along with the ones I have not yet envisaged and that will inevitably come up. Most of the readers of this article are far more aware of the technical problems than I, for example the challenge of making a working interface among different types of mark-up on the same text. As I learn to do well-formed mark-up, I am also faced with its limitations as I am told that we may not be able to mix different mark-up schemas. Will TAPoR be able to develop multi-dimensional XML? I certainly hope so, but I have no idea how far in the future the realization of this dream might lie.

2.2 The learning curve

I will go out on a limb, but not a very long one, by suggesting that my obviously limited but growing technical knowledge is typical of Humanities faculty members. Another appealing aspect of TAPoR has been the technical support provided for naïve humanists like me. Peter Liddell and Michael Best have convinced us that we can be involved in a cutting edge metamorphosis of text analysis tools. Summer workshops at U Victoria in XML and Text Encoding were eye-openers! It was comforting to learn XML and to be told that HTML is on the way out; one more

technology I won't have to deal with... With this very basic knowledge comes many more questions, and one of the most fundamental is "Do I have to do this encoding myself or can I hire a research assistant to handle the drudgery?" There are various answers to this query floating around the Humanities Computing and Media Centre at U Victoria; students can do light encoding (marking the title, author, publication information, that sort of thing), but since editorial decisions are called for regularly, researchers will have to be capable of tagging their own material, especially as the meta-tagging of a document becomes more sophisticated. This is apparently a point of debate; the role of research assistants will become clearer as the several TAPoR projects progress, but I have come to believe that principal investigators must know the basics in order to have input into tool development, and to continue discoveries into just what "text analysis" means. One thing it has begun to mean is organic growth: that is, expansion in various directions, such as the ones I have outlined for my particular project, amid constant give-and-take between the process of tool development and research in the Humanities. Students can and should participate in these developments, and not just the "techies." When I pursue a SSHRC grant specifically related to TAPoR, I hope to interest some of my students to get involved in the project -- but that does not absolve me of the burden to understand the principles of markup and encoding.

In fall 2003, many of us were asked to participate in an electronic survey designed by Geoff Rockwell, Ray Siemens, Stefan Sinclair, Lynne Siemens and Elaine Toms to gauge levels of knowledge about humanities computing among colleagues in the humanities. This effort to involve humanities scholars at various levels of computing knowledge is greatly appreciated by many of us, not only because it is a gesture of outreach, but because in my case it reminded me that there are already text analysis tools that I need to learn more about. When I did the survey, I came upon a list of eleven text analysis tools most of which were completely unfamiliar (for example, OCP, Micro OCP, TUSTEP, PARA CONC, CONCORDANCER, HYPERPO). This is intimidating, but also a sign that the field is evolving rapidly.

2.3 First steps: an experimental cocoon

Theory has indeed begun to turn into encoded text: my project now exists as an experimental TAPoR cocoon site. Preliminary markup on 14 texts (all examples of misogamist satire) was done during the summer of 2004

by research assistant Annick Nkurunziza. My involvement in the process was constant, given the nature of the texts: even a basic file description called for judgments about the many anomalies that early modern documents present. Fictitious publishers, unclear attribution of authorship and how collections were constituted were all issues generating questions that a non-expert reader could not answer independently.

In September 2004, Martin Holmes of U Victoria's Humanities Computing and Media Centre reviewed and corrected the markup that had been done, after which he began construction of the site. Notes appear as pop-ups, an important step in the move toward the sort of functionality we hope to see develop.

During the spring of 2004, I selected and ordered in CD-ROM format from the Bibliothèque Nationale de France six engravings related to these satirical texts. Among our next challenges is experimentation with these images, which we intend eventually to post online (at a cost of 50 euros per image for permission from the BnF to do so).

Martin Holmes and I have a work plan in place, although while waiting for significant funding it is difficult to move rapidly. Increasingly proficient student assistance with markup will, I hope, be available when this project is fully supported.

Conclusion

The keyword for me as we get further into this exciting exploration is potential. We are now at a threshold, and the inclusion of a variety of research projects is essential if we are going to develop tools that will serve a large segment of the population of Humanities students and scholars. To quote U Victoria's Martin Holmes, the motto of TAPoR could be "Find yourself an obsession and run with it" -- an extremely stimulating proposition. Thanks to TAPoR, I am seeing my material differently, and most certainly from a broader perspective. The question of editorial choice is central, as it is in print culture, but in an environment of growing interactivity, we are also creating a new twist on reader-response theory: the reader's role as creator of the text(s) s/he reads takes on new meaning in a multi-layered setting.

We will continue to learn in new ways from the mass of information it is possible to present electronically, but only to the extent that new instruments will allow us to do so. Gratitude is owed to our colleagues working on text analysis tools, both for the work already done and for the

progress being made daily toward the developments that make it possible to draw knowledge from information.

Notes

[1] This despite the fact that, Gallica<http://gallica.bnf.fr>, a site of the Bibliothèque Nationale de France, has made available hundreds of Early Modern texts.
[2] The best example of such a frontispice is found in *Claude Maillard's Le Bon Mariage ou le moyen d'estre heureux et faire son salut en estat de mariage* of 1643, where the allegorical figure of Marriage is surrounding by legions of the faithful as well as religious symbols such as the Lamb of God.

Works Consulted

De Casu Cizaris Dutis Regis Iabin. An Episode from John Lydgate's Fall of Princes. <http://web.uvic.ca/hrd/lydgate/>.

Extracts from the Diary of Robert Graves. <http://web.uvic.ca/hrd/graves/>.

The First Voyage of Othere. <http://web.uvic.ca/hrd/worldcall_2003/oldenglish >.

Gallica. *Bibliothèque Nationale de France*. 5 Dec. 2004 <http://gallica.bnf.fr/>.

Holmes, Martin. Personal Conversation. 7 Nov. 2003.

The Internet Shakespeare. <http://web.uvic.ca/shakespeare/>.

Maillard, le R. P. Claude (1643). *Le Bon Mariage ou le moyen d'estre heureux et faire son salut en estat de mariage avec un Traité des Vefves : Livre tres-utile à ceux qui sont mariez, & à ceux qui aspirent au mariage, ou qui ne sont encor determinez à aucun estat, & condition de vie*. Douay: Jean Serrurier.

Siemens, Ray (2004). "Text analysis and the dynamic edition? Some concerns with an algorithmic approach in the electronic scholarly edition." *Text Technology/Computing in the Humanities Working Papers.* [Forthcoming.]

Siemens, Ray, Andrew Mactavish, Susan Schreibman and Stéfan Sinclair. "Welcome and Introductory Lecture: What is Humanities Computing?" 2003 Local Workshops in Humanities Computing. U. Victoria, Victoria, BC. 23 June 2003.

Toms, Elaine, Geoffrey Rockwell, Ray Siemens and Stéfan Sinclair. "Text Analysis Needs of Humanists: A Survey." *Canadian Symposium on Text Analysis (CaSTA)*. Nov. 2003, U. Victoria, Victoria, BC. Abstract. 5 Dec. 2004 <http://web.uvic.ca/hrd/casta/pages/abstracts.htm>. Subsequently re-presented as "Text Analysis Research: What is Being Done and What is Needed." *Consortium for Computers in the Humanities* at the 2004 *Congress of the Canadian Federation of Humanities and Social Sciences*, May 2004. U Manitoba, Winnipeg; with Lynne Siemens as: "The Humanities Scholar in the Twenty-first Century: How Research is Done and What Support is Needed." Joint International Conference of the *Association for Computers and the Humanities* and the *Association for Literary & Linguistic Computing*. June 2004. Göteborg U, Göteborg. Abstract. 6 Dec. 2004. <http://www.hum.gu.se/allcach2004/AP/html/prop139.html>; "Modelling the Humanities Scholar at Work." The Face of Text: Computer Assisted Text Analysis in the Humanities. Nov. 2004. McMaster U, ON. 5 Dec. 2004 <http://tapor1.mcmaster.ca/~faceoftext/abstracts.htm#toms>.