Preprint of book chapter to be published in 2007

Philology in an Electronic Age

Gregory Crane, David Bamman, Alison Babeu

from Anne Thompson and Bruce Fraser (eds.), Greek Lexicography after Liddell and Scott (forthcoming)

This paper considers two questions, one broad and thus more a series of further questions, the other more practical and with a particular suggestion for action. This paper emerges from a July 2002 conference about lexica and few would be willing to argue that classicists could not improve upon the foundations left to us by Liddell, Scott, Jones, and the others who labored on our shared Greek-English Lexicon. But simply because we could improve upon what we already have does not tell us what we should build now. We do not need a better nineteenth-century lexicon. We do not even need the best possible twentieth-century lexicon. We need to develop lexicographic resources to serve non-lexicographic readers of the twenty-first century.

Our perspective flows from the following observations:

First, texts derive their value from being read – an idea that can be traced in academic discourse to Plato's *Phaedrus*. One could argue against simple quantitative measures – a few hours of passionate, engaged reading might outweigh thousands of hours spent by press gangs of students slogging heavily through their assigned rounds. Nevertheless, the heavy feet of dutiful student reading has traditionally broken the sod in which more lasting analysis has subsequently flourished. Few, if any, students in the early twenty-first century are forced to learn classical Greek. The young who seek direct advancement visit fields far removed from Homer in the original Greek.

Those of us who are Hellenists have, as our first and overriding goal, to pass on the practice of reading classical Greek to subsequent generations. We need to strain every muscle to attract more into the difficult, even harsh practice of learning and reading Greek. Where college level Latinists can rely upon a small army of high school Latin teachers to provide new students, Hellenists must, for all practical purposes, train their students from scratch.[1]

Those Hellenists who believe that we currently fulfill our obligations in an acceptable fashion need not read further. Our students are with us for all too brief a period. Even if we get to them in their first year, they rarely acquire the skills to immerse themselves fully in classical Greek. Having painfully absorbed some Greek and read a small set of texts, they move on to their careers, their hard-won knowledge of Greek slipping away in the hubbub of busy lives.

Second, the center of serious inquiry has shifted irrevocably to a networked, electronic world. Whatever resources we design must be designed to flourish within such a world and to serve an audience that has little patience for the friction of moving within a print world.[2] Simple online page images such as PDF are problematic and definitely inadequate for reference works such as lexica. Electronic lexica seamlessly adapt their formats to the particular background and immediate needs of particular readers. Thus the same system would adapt itself to the varying needs not only of different users (e.g., the professor doing research vs. the student learning to read the language)[3] but of the same users at different points (e.g., intermediate

[1] A relatively recent study indicated that while the number of college students studying classical Greek was relatively stable from 1990 to 1998, between 1974 and 1990 there was a 50% drop in the numbers of students (Lafleur 2000).

[2] For more on the impact of technology and the digital world on classics, please see Crane (2004), McManus and Rubino (2003), Latousek (2001), and Hardwick (2000).

[3] A variety of research has been conducted into how systems can best automatically adapt themselves to the needs of different readers. For an overview, please see Russell (2003) and for a specific application of a personal reader see Dolog et al. (2004).

students of Greek working through a particular passage vs. studying a particular key term as part of a research paper).[4]

Third, the primary research problem that now confronts us in humanities and in classics above all lies not in the production of new knowledge but in learning how to communicate the knowledge that we now possess to the widest possible audience.[5] We have to a large extent allowed our colleagues in the sciences and professional schools to define our attitudes towards research. We know an immense amount about our subject – indeed, our focus upon Greco-Roman culture gives us a key advantage as we compete to engage new generations of students. Our greatest challenge for the immediate future lies in converting as much of this existing knowledge from inert print, comprehensible only to a small set of highly trained human readers, and into a form that can make itself useful to the widest possible audience. The print lexica, commentaries, editions, encyclopedias, monographs and similar instruments provide us with a starting point that our colleagues in many fields can only envy, but each genre of publication needs to be rethought from the ground up.

Fourth, we cannot flourish as a discipline if we aim at the audiences of specialist scholars and full-time students on whom we have traditionally concentrated. The networked world has spurred the creation of new audiences and new forms for collaborative contribution.[6] The heart of modern culture beats vigorously but, for the most part, far beyond the channels of scholarly communication. Consider one example from popular culture, a series defiantly named *Buffy the Vampire Slayer*. Two English professors publish *Slayage: the Online International Journal of Buffy Studies*,[7] while other web sites aggregate and manage contributions from academics, the general public and the commercial world. When this series was in production, individuals would intercept the satellite feeds to individual network affiliates and post the digitized episodes online a day or so before they aired. Others would, in turn, create scene by scene descriptions and commentaries of the new 50 minute productions, making their work available before the show aired the next day. Serious students of the series would, by the time of broadcast, have already studied a community generated edition of the broadcast episode. The broadcast would then become a period for interactive comment and exchange via personal messaging and chat rooms.

It would be easy to criticize and even dismiss such activity. While students of vanished worlds would be sensible not to accuse enthusiasts for popular culture of frivolity, our standards of producing, vetting, disseminating and preserving the apparatus to decode complex cultural artifacts is highly developed, if archaic. We have much to offer and much to gain by drawing upon, enhancing and working with communities from beyond the classroom. Our best course as a field is to stimulate interest from such user communities, creating a new infrastructure in which specialist and novice alike contribute.[8]

The assumptions listed above aim at situations such as the following:

As a reader confronts a passage of classical Greek, the reading support system considers the following:

- Is the reader working her way through a continuous text, looking at the results of a word search or checking a citation from a secondary source? Each of these three cases has implications for the kinds of information that this particular reader might find useful. If the reader has been systematically working through the *Iliad* and is calling up the next unread section, then the system may marshal information relating this passage to the previous. If the reader is conducting a word

[4] The field of adaptive hypermedia (AH), particularly, has explored how to model the needs of individual learners and to track the changes in the knowledge state of users as they interact with a system. For an overview of AH please see Brusilovsky (2002), and for specific applications see Kavcic (2004) and Weibalzahl and Weber (2002).
[5] Some interesting work in making research materials more accessible to a wider audience can be found at the Public Knowledge Portal (Willinsky 2003).
[6] As examples such as Wikipedia illustrate, there are an immense number of users who want to make contributions to both scholarly and popular culture ventures. Different models exist for encouraging contributions from user communities, including Project Gutenberg (Newby and Franks 2003) and Planet Math (Krowne and Halbert 2003).
[7] http://www.slayage.tv/
[8] One interesting project that has sought to harness the collective power of both expert and user contributions to create an encyclopedia is described in Kolbitsch and Maurer (2005).

search, then the system may want to present initial thumbnail information about date, genre and general context. If the reader is working through a study on marriage, the system should be able to use that information to provide context to help understand the contribution of a particular text to this particular topic.[9]

- What is the state of the reader's knowledge? Reading support systems should help their users keep track of what they have studied.[10] Thus, when the introductory student in later first year Greek confronts a passage of Lysias, the system should be able to keep track of not only what vocabulary that particular student has learned but which senses of a word that student has encountered: thus, the system should identify unfamiliar definitions for common words that the reader does not yet know.[11] The system should likewise track syntactic phenomena as well, identifying constructions with which the reader should be aware from those which may be new. Mature systems should have well-developed models of memory and usage. The system should be able to identify linguistic phenomena of particular interest to a particular reader, whether that interest lies in its difficulty or in the fact that the reader is interested in a particular usage.[12]

  Mature systems should be able to analyze question patterns and diagnose problems that their readers are encountering.[13] They should be able to quiz readers, testing their knowledge and allowing them to evaluate for themselves the degree to which they grasp the texts before them.[14] We need models that allow for the automatic generation of such diagnostic questions so that these systems can be scalable and support those reading texts beyond heavily worked canons.

  Mature systems also need to preserve not only core data but individual experiences of long periods of time.[15] Thus, many potential readers wish to return to their Greek later in life, when the demands of family and career have begun to abate and they are, in many ways, better prepared to appreciate the literature that they first encountered in their youth. Reading support systems should be able to remember what these readers learned decades before, their previous learning styles and their current cognitive practices.

- What does the reader wish to accomplish at this particular time? Reading support systems should not only understand the particular backgrounds of their users but also their immediate purposes.[16] Thus, readers may be skimming the results of a word search, trying to ferret out the basic meanings from many different passages, reading quickly through a particular text, or studying a particular passage exhaustively. Systems should be able to combine background knowledge and

---

[9] Learning how to support the contextual needs of readers as they move through a system or across systems is a growing area of research in both personalization and e-learning. For examples, please see Henze and Nejdl (2002) and Niederee et al. (2004).

[10] A variety of work has been done on reading support systems that help track what the user has learned and provide recommendations for review and further study. For examples, please see Rouane et al. (2003) and Wang and Chen (2004).

[11] For similar work in vocabulary tracking (albeit with elementary school students), please see Brown and Eskenazi (2004).

[12] Advanced work in reading support systems has been reported by Terras (2005) who created a "papyrologists assistant" or a computer program to aid experts in reading ancient documents.

[13] The analysis of question patterns and personalization is a growing research area. See, for example, Koutrika and Ioannidis (2004).

[14] Some interesting work in having a system generate specific questions at various points in a user's interaction to test their current understanding has been proposed by Fleming (2004).

[15] The idea of a permanent personal digital archive or storehouse of lifetime memories and knowledge has been well articulated by the creators of MyLifeBits (Gemmell et al. 2006). Neil Beagrie has also explored this concept (Beagrie 2005).

[16] A variety of research has explored how to create user models and personalization systems that can dynamically reflect the changing needs or interests of its users. For examples, please see Kim and Fox (2004) and Teevan et al. (2005).

immediate purpose as they customize displays of information.[17]

- What language does the reader wish to use? The audience for classical Greek extends, of course, beyond the English speaking world and should extend beyond the cluster of languages supported by the EU. A mature system should provide access not only to Classical Greek literature but Sanskrit, Classical Chinese, Arabic and other textual corpora from the West and beyond. The reading environment should encourage cross-cultural reading and research.

  Some features of a reading support system will inevitably be language dependent (e.g., the English distinction between a river bank and a financial bank) but many terms (e.g., nominals of concrete objects, terms for particular cultural practices such as *xenia*, and virtually all morphological and syntactic information) can be represented in a language independent fashion.

  Even bad machine translation coupled with morphological and syntactic data can provide substantial help to human readers working with a new language.[18] Technologies such as cross-language retrieval (in which users pose a question in one language, e.g., English, and retrieve results in multiple other languages, e.g., Chinese, Arabic and Spanish) have made substantial strides.[19] We must design our resources for classical Greek to enhance the tools available

- What decisions have other readers made as they read the same or similar texts? Classicists are, of course, familiar with commentaries, and reading support systems such as those in the Perseus Digital Library already link multiple commentaries to individual texts. Such formal commentaries are, however, labor intensive, require expensive editorial intervention, are available online behind subscription firewalls (if at all) and often do not cover the mundane questions that bedevil most readers. Readers should be able to share their judgments as to how a particular sentence is constructed, what the meaning of a particular term might be, and more interpretive responses.[20] Digital commerce has already produced models for representing trustworthiness that would allow us to distinguish and make productive use of the annotations from a range of users.[21] While users may be able to pick and choose between many accomplished annotators of canonical texts such as Sophocles' *Oedipus Rex*, any substantive annotations in the vast stretches of Galen or other texts beyond the traditional canon will be welcome to many.

- What would the current reader be willing to share with others? Formal comments and questions are obvious categories of annotation, but systems can mine useful information from actions as well (e.g., in a given lexically ambiguous term in a particular text, which dictionary entry do most users choose to examine?). Data mining has applications for the support of reading Greek just as it does for commerce.[22]

The functions outlined above can be implemented effectively with existing technologies. The reading environments that projects such as Perseus have provided for classical, and that various commercial efforts

---

[17] Some recent interesting work has explored how to best visualize users' current purposes as they navigate a system (Fisher and Everson 2003) as well as how to provide customized recommendations based on both the user's current context and the documents that have been viewed before (Billsus et al. 2005).

[18] Church and Hovy (1993).

[19] For an overview of recent work in cross-language retrieval, please see Gey et al. (2005).

[20] There has been a great deal of recent interest in exploring different ways of supporting both scholarly and general user annotations within online communities. For example, the museum community is exploring allowing users to add online annotations to exhibits (Kateli and Neville 2005), and by encouraging users to add keywords to images to increase retrievability (Bearman and Trant 2005). A number of research projects such as COLLATE (Thiel et al. 2005) and CYCLADES (Renda and Straccia 2005) allow scholars to both annotate resources and share these annotations with other.

[21] For example, the commercial site Slashdot allows users to achieve online reputation or karma points through posting reviews or moderating comments (Lampe and Resnick 2004).

[22] Data mining has been used to determine groups of user communities in digital libraries (Papatheodorou et al. 2003) and to explore previous patterns of resource use to support more appropriate linkages between resources (Elango et al. 2004).

have provided for New Testament Greek, constitute only an early generation. More powerful environments will come. Now classicists must wrestle with the questions of how they wish to contribute to those environments, for digital reading is not the future, but already the established practice for many, if not most, of those who have learned to read classical Greek in the past ten years.[23]

We need good research to help us rethink the form in which we shape our ideas about the classical Greek world. The print based methods at our disposal are little more effective than the academic Latin of nineteenth-century dissertations, for they assume access to paper documents and, worse, an immense amount of useless logistical effort. Secondary literature is secondary in importance and should be subordinated to the practice of reading the primary literature. While our goal may be to encourage thoughtful reading, human reading now depends upon prior conversations between multiple electronic systems – even the production and distribution of a printed book requires collaboration between a range of systems from word processors through library catalogues. The more effectively the electronic systems can communicate, the more effectively we can support the task of human reading.[24]

Even if the broad features outlined above do prove to characterize conventional reading environments in the future, we need to understand our users, present and potential, much more thoroughly if we are to plan effectively. Topics for research include the following:

- Who are the current readers? Our experiences in supporting the Perseus Digital Library suggest that this set is broader than we had thought and the weblogs provide a useful starting point. Emails suggest that readers come from lunch time breaks in financial institutions, isolated kitchens in the Appalachian mountains, naval vessels, and other locations far beyond traditional dorm rooms and academic libraries.

- Who are our potential readers? If we want to expand the set of readers, who might these new readers be? Potential areas for growth include not only traditional students but also previous classics majors and adult learners from a global population.

- What questions do various readers of classical Greek actually confront as they struggle through a text? What problems do they really face? My students rarely find the questions addressed in many commentaries "aimed primarily at students at university and in the upper forms of schools" to be useful, probably because such commentaries are really written to impress committees for tenure and promotion.

- How might the problems and expectations of our readers shift? What new kinds of questions do we need to be able to support? Current readers may find dictionary lookups useful, but readers may soon expect more sophisticated linguistic help.[25] Certainly, readers will expect links between semantic and encyclopedic data: thus, readers should expect that any reference to an archaeological site such as Delphi should naturally lead to a self-organizing database on the history and topography of Delphi.[26] Moreover, we need to anticipate much broader cross-cultural reading and research, in which our readers expect to explore, for example, Confucius and Plato in ways that were simply not feasible in the past.

While human readers are the center of our field, we also need to scrutinize emerging language technologies.[27] Information technology – whether print or digital – constrains the set of questions that we can ask. Emerging technologies thus not only affect our ability to ask the same questions but potentially break ground for new sets of questions as well. While we must not let the technology drive our work, we must also understand what technologies can do at present and where they may proceed in

[23] For an analysis of the recent shift to online reading among students, please see Marshall (2005).
[24] For more on this issue, please see Crane (2006) and Crane (2005).
[25] Related work on developing more sophisticated linguistic tools to assist readers can be found in Carlquist (2004) and Sinclair (2003).
[26] For more on the need for semantic links to historical reference works, please see Crane and Jones (2006).
[27] For a more in-depth review of some of these issues, please see Crane et al. (2005).

the future. There are at least three reasons to track particular technologies, First, we may be able to influence the ways in which technologies evolve. Second, we may be able to identify tools of interest. Third, we need to understand what tools exist today and may emerge in the near future so that we can consider how, if at all, we may make our work interact more powerfully in increasingly sophisticated reading environments.

Consider a simple technology in which classicists have invested substantial effort and that has been immensely successful: standardized citations such as Thuc. 1.35 (= Thucydides' *History of the Peloponnesian War*, Book 1, Chapter 35) constitute a sophisticated reference language by which classicists have for generations been able to identify fine grained chunks of text. The system is robust, covering multiple editions of the same work, each of which may differ in small ways. While different works may instantiate the system in somewhat different ways (e.g. "Thuc.," "Th.," "T." may all serve as abbreviations for Thucydides), the basic coordinates (e.g., 1.35) are remarkably consistent. The stability of this citation language has allowed us to convert citations into links: the Perseus Digital Library contains hundreds of thousands of such links derived from more than a century of print scholarship. Shakespearean scholarship, by contrast, conventionally renumbers the plays in their canon with each new edition, making it much harder to map a particular reference (e.g. Hamlet 3.1.44) to a particular segment of text. While this citation language emerges from print, it nevertheless constitutes a powerful information technology in which classicists can take pride. By using the same methods to identify segments of canonical texts, classicists were able to promote interoperability in print and electronic form alike.

Consider now a second technology less well served by print conventions. The Perseus Digital Library provides morphological analysis for classical Greek and the resulting database has proven useful for teaching and research alike.[28] While a great deal of software maps inflected forms to their canonical dictionary entries, the most important component to the system consists of a database of stems and endings. Thus, the system knows that πεμπ- is the stem for the first principle part of a conventional Greek verb and that it can take endings such as -ω, -ει, -ετε and augments such as ἐ-. Thus, when it sees the inflected form ἐπέμπετε, it can analyze the augment, stem and ending, check the accent and identify this as the 2nd person plural imperfect indicative active of the verb πέμπω.

Building such a morphological database for classical Greek is far harder than building the morphological analyzer. When we set out to digitize the Intermediate Liddell Scott Greek-English Lexicon and then, five years later, LSJ 9, we were primarily interested in extracting the morphological data. We were able to mine the online versions of these lexica for tens of thousands of stems but the process was messy and imperfect. While certain patterns did frequently recur, different dictionary entries would represent morphological information in slightly different ways. Human readers had little difficulty managing such variations because they bring an immense amount of knowledge with them. Automated information extraction routines rely upon regularities of expression that human authors and editors rarely follow.[29] Thus, the information extraction routines missed a small but substantial amount of the morphological information that the human lexicographers encoded in the lexicon. The automatic morphological analyzer is, however, probably the most important single "reader" for the morphological data in the lexica: its ability to analyze morphological data allows it to provide the linking that thousands of users daily use to search and read classical Greek.

The form that we give to our ideas constrains the questions that we can subsequently ask. Our field established print conventions that have proven highly effective for citation linking but much less so for morphological analysis. While language technologies are rapidly evolving and will not stabilize in the near future, we as philologists need to understand both what we can do today and plan as best we can for what will emerge.[30] The following technologies already exist and should be available for use with classical Greek:

---

[28] Please see Crane (1991) and Crane (1996).

[29] For an overview of information extraction technology, please see McCallum (2005).

[30] Some early work in humanities computing has explored the potential intersection of philology and new technologies – see Bozzi and Calabretto (1997), and for more recent work Spencer and Howe (2004) and Bøe et al. (2004).

- Clustering: Given a set of words, phrases, passages or whole documents, identify statistically significant patterns.[31] The Perseus Digital Library already includes several types of clustering. We associate statistical correlations with Greek dictionary entries: 1) words with similar definitions in Greek and Latin; 2) words that are used together. A third type of clustering is available for searches: search for a term and the system looks for words or phrases associated with it. Such automatic clustering provides a scalable way to identify patterns for further analysis. The richer the underlying databases, the more sophisticated the clusters can be: we automatically analyze place names in full text, mapping these to their geographic coordinates (e.g., distinguishing Salamis of Athens from Salamis in Cyprus).[32] We could thus use this information to cluster texts by geographic focus and allow readers to ask for texts relevant to a particular place. If we combined this with dates (which we also extract)[33], we could support geo-temporal queries: e.g., show me information relevant to Marathon in the latter half of the fourth century.[34]

- Machine translation (MT): Translations into modern European languages exist for most of the canonical texts of classical Greek, but digital libraries such as the TLG now make available many texts for which translations are not available online or even in print. MT already has a long history but systems that can rival human translators are a long way off.[35] Nevertheless, even bad machine translation can provide a helpful tool to readers struggling with a text, especially when they have other reading aids (such as morphological and dictionary lookups).[36] Moreover, MT can be used to generate translations not only into Western languages but into Chinese, Arabic, Hindi or other important languages less well served with manual translations.

- Word sense disambiguation: Far more modest than MT (and, indeed, a component within many MT systems), word sense disambiguation maps a given instance of a word to its closest dictionary meaning, identifying, e.g., the use of the word "bank" either as financial institution or as the area next to a river.[37] Evaluating the performance of WSD systems is difficult because success rates depend upon the complexity of the dictionary entries and the depth to which the automated system needs to correctly achieve a match.[38] Nevertheless, existing technologies should allow us to present the best dictionary meaning to a reader 80% of the time.

- Cross-language information retrieval (CLIR): Pose a query in one language and retrieve results in multiple languages (e.g., Greek and Latin, as well as Sanskrit, Classical Chinese etc.).[39] While one can, of course, apply MT to a query (e.g., convert the English query into Arabic, then search the Arabic), there turn out to be more effective approaches. Both US and EU evaluation forums have tested various systems against a range of languages.[40] The results have been surprisingly good – the best CLIR systems have occasionally outperformed the monolingual control systems: e.g., posing a query in English provided better results from an Arabic corpus than posing the query in Arabic.[41] CLIR opens the possibility for new avenues of cross-cultural research that few, if

[31] For a general overview of document clustering algorithms, please see Steinbach et al. (2000). The use of clustering technologies has only recently begun to make its way into the humanities. For one interesting example of how clustering was used to create automatic topic categories see Krowne (2005).
[32] For an overview of place name extraction in the Perseus Digital Library, please see Smith and Crane (2001).
[33] See Smith (2002).
[34] Interesting work in this area has been done by Pouliquen et al. (2004).
[35] For two useful overviews of machine translation, please see Ney (2005) and Marcu and Knight (2005).
[36] For an overview of machine translation with particular applications in the humanities, see Smith (2006).
[37] For an overview of WSD, see Ide and Véronis (1998).
[38] The SENSEVAL workshop and evaluation exercise (Kilgarriff 1998) addresses this difficulty by providing a common task on which WSD systems can test their performance; the most recent of these (SENSEVAL-3) took place in 2004, with SENSEVAL-4 currently underway.
[39] For an overview of recent technical advances in CLIR, please see Kishida (2005).
[40] For a review of the evaluation forums and recent results for CLIR, please see Peters et al. (2004).
[41] Ibid, Gey et al. (2005).

any, of us could pursue in a print environment.[42]

- Automatic summarization: At its simplest, this technology produces summaries of modern documents, assumes that such documents have unambiguous messages that can be reduced to summaries, and competes with human generated abstracts.[43] More recently, however, summarization has grown more sophisticated, recognizing that the same document may contain different messages for different readers at different times. While automatic summarization may be best suited to secondary literature, it can also be used to extract information about particular themes from larger corpora: e.g., prepare a report on opinions about slavery in later Greek prose. [44]

- Question answering: Question answering systems look for the answers to discrete questions (e.g., "how much steel did Poland produce in 1993?").[45] QA systems are best suited to communities that pose concrete questions of corpora that contain many easily quantified propositions, but their applications may prove substantial even when applied to pre-modern primary sources.[46]

The above list concentrates on well-defined technologies where multiple systems have been subjected to systematic evaluation. Such technologies can be combined into more complex systems and only constitute an initial sample.

We in classics are surprisingly well positioned to exploit such technologies. While the software required can be quite complex, software requires structured data to do its work. The following are resources that are already available in electronic form and that will allow classicists to draw upon a variety of technologies.

1) Textual corpora: The Thesaurus Linguae Graecae[47] provides classicists with access to a comprehensive digital library with more than 75 million words of classical Greek. Perseus[48] contains a much smaller textual corpus (7.5 million words) and the editions are often older than those in the TLG, but the corpus is based on public domain materials.

2) Morphological analysis: systems need to be able (1) to extract the morphological information implicit in inflected terms and (2) to map inflected terms to possible dictionary entries. We have tools that can accomplish both of these tasks already in place.[49] The morphological analyzer in Perseus already supports our clustering algorithms and can enhance the performance of many systems addressing all of the technologies listed above.

3) Aligned corpora of translations and source texts. Statistical methods identify possible translations for words and phrases by building language models from large sets of equivalent bilingual texts.[50] Perseus contains substantial aligned corpora for Greek and Latin. The finer grained the citation scheme, the more effective the statistical methods, but preliminary analysis suggests that even the

[42] For the use of cross language searching with classical texts, please see Rydberg-Cox et al. (2004). Similar work with modern texts was reported by Cheng et al. (2004).
[43] State of the art summarization technologies and techniques are evaluated annually by the Document Understanding Conference (DUC). For the most recent overview of DUC, please see Dang (2005). For a good overview of summarization technologies, see also Hahn and Mani (2000).
[44] For some interesting uses of summarization technologies based on extraction of relevant information from multiple documents, please see Sengupta et al. (2004) and Zhou et al. (2004).
[45] Advances in question answering technologies are explored at conferences such as DUC and TREC (Text Retrieval Conference). For an overview of recent work, see Maybury (2004).
[46] For an intriguing look at the possibilities of question answering for the humanities see Cohen and Rosenzweig (2005) and Cohen (2006).
[47] http://www.tlg.uci.edu/
[48] http://www.perseus.tufts.edu/
[49] See Crane (1991) and Crane (1998).
[50] This is generally associated with the work of Brown et al. (1990), who used the Hansard texts, a bilingual English-French corpus of Canadian Parliamentary proceedings, for English-French MT. For recent work in this area, please see Lambert and Castell (2004), Tiedemann (2005) and Mihalcea and Simard (2005).

coarser citation schemes (e.g., page to page) will provide useful results.

4) Machine readable dictionaries: These provide a wealth of information. We already use the morphological data and make some use of the tagged definitions, but computational linguists can do far more with such resources. MRDs are predominantly used in word-sense disambiguation (as described above) but are also the core of a multitude of other tasks as diverse as ontology induction,[51] cross-language information retrieval[52] and document clustering.[53]

5) Encyclopedias of people, places, institutions etc.: Where traditional classical lexica concentrate on semantic data, we also have elaborate encyclopedias on a range of topics, and some of the older ones are already in Perseus. These can provide authority lists – stable lists of proper names – while the associated articles help us analyze full text (e.g., we can use the content of articles on multiple Alexandrias or Cleopatras to determine which Alexandria or Cleopatra is meant in a particular passage).[54]

All electronic resources are subject to refinement: we can add more markup to capture more phenomena, align our parallel texts and translations more precisely, improve the morphological analysis, etc. The major challenges that classicists face are, however, more organizational and logistical. We need to make resources that now exist more accessible, providing both flexible rights and smooth electronic access.

While research on the needs of our field is the most important next step, one major project can already be identified. While we can certainly improve on our existing Greek lexica, we should first build a database of syntactically parsed Greek sentences – a database of parse trees or "treebank." Treebanks exist for a variety of languages and are core resources for corpus linguistics, of which the study of classical Greek is a small subset.[55] Such treebanks can encode various amounts of data: the more comprehensive the syntactic information, the more expensive they are to produce. At its simplest, a treebank for classical Greek would encode the dependencies within individual sentences, including many of the questions that pedagogues have asked students for centuries: this adjective depends on this noun which serves as the subject of a given verb. Even the simplest such treebank would include the structural outlines for individual sentences.

The resulting treebank would serve a variety of purposes. On the one hand, it provides a foundation for a new lexicon – one could even argue that a treebank constitutes the prerequisite for any serious new lexicon and that no new lexicon should be attempted until such a treebank has been established. The treebank encodes basic lexicographic information, such as which words serve as subjects/objects of particular verbs, which adjectives modify which nouns. Combining such data with semantic databases allows us to find broader patterns: e.g., a given verb for eating takes animals as its subject a given percentage of time in a given corpus. Such usage statistics provide a much firmer foundation than we have had before for systematic lexicographic research. Even more important, such a treebank would be expandable and would allow for a range of queries suited to particular research agendas.[56]

This kind of lexical power can perhaps be illustrated best with an example that arose in our planning for the construction of a Latin treebank. An interesting social question is bound up with the evolution of the Latin verb *libero* ("to free," "liberate"). If we want to find out how this verb changes in meaning across time, we can ask a simple question: what do Latin authors want to be "liberated" from, and how does this change? We can imagine that an orator of the republic has little need to speak of liberation from "eternal death," while an apostolic father is just as unlikely to speak of being freed from another's monetary debt. A

[51] Nichols et al. (2005).
[52] Aljlayl and Frieder (2001).
[53] Hotho et al. (2003).
[54] Please see Crane (2005b).
[55] For an overview of the use and development and treebanks, see Abeille (2003). Established treebanks exist for English (Marcus et al. 1993) and other fixed word order languages including German (Brants et al. 2002), Spanish (Moreno et al. 2000), French (Abeille et al. 2000), Italian (Montemagni et al. 2000) and Japanese (Kurohashi and Nagao 1998), and also for Czech (Hajic 1999), a relatively free word order language like Latin and Greek.
[56] For a review of treebank querying please see Lai and Bird (2004).

treebank allows us to ask this question in a systematic way and quantify the result. We can translate the question above into a treebank query by turning it into a syntactic one: what are the predominant objects of prepositional phrases modifying *libero* headed by *a* or *de* or with no head at all (i.e., bare ablative nouns)? In even a small test corpus we find an interesting answer: in 100 sentences of Cicero containing *libero*, the most common modifiers are *danger*, *fear, care* and *debt*; in 100 from the Vulgate, the most common are *hand* (e.g., *the hand of the Egyptians*), *death* and *mouth* (e.g., *the mouth of the lion*). Where traditional lexica such as Lewis and Short and the Oxford Latin Dictionary illustrate how a word is "typically" used by means of representative samples from various authors, with a treebank we can quantify that usage exactly.

Cicero

periculis	20%
metu	12%
cura	9%
aere	4%
scelere/sceleribus	4%
sollicitudine	4%
suspicione	4%
bello	3%
culpa	3%
molestia	3%
rege	3%
regno	3%

Jerome

manu/manibus	44%
morte	6%
ore	6%
latronibus	4%
inimico/inimicis	4%
bello	4%

Figure 1: Objects "liberated from" in Cicero and Jerome that occur with frequency > 1 in a corpus of 100 sentences containing any inflected form of the verb *libero*.

A database of syntactically parsed sentences, of course, also gives us a means to describe the state of syntax in Latin at any given time and also chart its transformation through different authors. By far the predominant method for Cicero to express the entity "liberated from" is with the use of a bare ablative noun (e.g., *res publica maximis periculis sit liberata*). In the Vulgate, however, the predominant method is a prepositional phrase headed by *de* (e.g., *liberet me de omni angustia*). In our 100-sentence sample, in fact, the bare ablative is not used by Jerome even once to denote the entity liberated from. With these two data points, we can pinpoint a clear syntactic change.

Cicero

--	97%
a/ab	3%
de	0%

Jerome

de	72%
a/ab	28%
--	0%

Figure 2: The head of the syntactic phrase containing the object "liberated from" in Cicero and Jerome.

A treebank is clearly crucial for supplementing lexica with this kind of information and for opening up an entirely new form of philological inquiry into classical texts. If our main goal is, however, to support those reading classical Greek, a treebank is probably the most important new tool that we could develop. For all their weaknesses, the lexicographic tools at our disposal serve the needs of our readers fairly well. New lexica can make substantial improvements but these improvements operate at the margin. A treebank

would, for the first time, provide comprehensive syntactic support. While systematic research of our readers' actual needs remains to be done, a substantial amount of reading time surely rests in puzzling over problematic syntactic structures. The ability to determine which words depend upon which would solve many problems of those struggling to work their way through classical Greek text.

We can, of course, argue that too much reading support is a bad thing and enervates our potential readers. Such an approach works well if the study of Greek constitutes the prerequisite for some compelling, material good – e.g., "to be a gentleman, one need not know Greek, but one must have forgotten it." But if, on the other hand, we welcome those who devote themselves to this language and cherish their time and respect their intentions, then we can hardly provide too much help. If we fret that the rising generation of students will never really learn their Greek, then we should devote ourselves to developing systems that will help new readers track the questions that they ask and help them understand what knowledge they have made their own.

There are various ways to construct such a treebank and substantial analysis would need to precede any serious attempt.[57] The number of words that can be analyzed in an hour varies depending upon the language, the availability of trained readers, the existence of an automated syntactic parser to do some of the preliminary work, and the amount of syntactic information that the treebank sets out to encode. One English treebank was able to process 750 words per hour,[58] while a Chinese treebank, proceeding without a syntactic analyzer, was able to cover 240 words per hour.[59] The Chinese project increased this throughput to almost 480 words/hour by adding a simple syntactic parser that could capture easily identified phenomena and such an approach should work well also for Greek.

The most heavily read texts of Greek comprise a corpus of c. 5 million words. If we can process at least 250 words per hour of labor, then we would require 20,000 hours or roughly 10 years of labor to create an initial treebank for the most heavily read texts. This is a major project by the standards of a small field such as classics but the total cost would be under one million dollars – an immense sum, but something that either U.S. or European funding agencies could pay for over time. If the project used simple pre-processing of the Greek texts, provided a framework for collaborators to mark up their particular authors and an editorial board to test the results, then the work could be done with substantially less monetary support.

Once we had treebanks for the main classical texts, we could use this database to train automated parsers that would automatically process the remaining 70+ million words. Research in Latin syntactic parsing at Perseus suggests that the accuracy of an untrained parser – one used without being trained on a treebank – hits a performance ceiling of about 70%. The accuracy of trained parsing would still be less than that of the manually curated corpus but it would, however, probably capture 80-90% of the syntactic relations accurately and provide a major new instrument for the study of Greek.

We are only now beginning an extended period of change for academia as a whole and for classics in particular. Decisions that we make now will constrain the extent to which we participate in broader currents of thought and exploit immense investments in computational tools for the study of language and culture. We cannot simply compare the present to the past and strive to do better what we have done before. While our core mission – the reading of Greek texts and the passionate engagement with the culture that produced them – may remain the same, the ways that we approach this problem and, more importantly, the audiences to which this field is accessible will – and arguably should – evolve along new paths. While we cannot see far, we can see the beginnings and we can, as a field, determine what steps we should now take.

---

[57] For some different examples, please see Rosen et al. (2005) and Tanaka et al. (2005).
[58] Marcus et al. (1993).
[59] Chiou et al. (2001).

References

Abeille, A., L. Clement, A. Kinyon, and F. Toussenel (2000). "Building a Treebank for French." *Proceedings of the Second Conference on Language Resources and Evaluation*, pp. 87-94.

Abeille, A. (ed.) (2003). *Treebanks: Building and Using Parsed Corpora*. Series: *Text, Speech and Language Technology*, Volume 20 (Dordrecht: Kluwer Academic Publishers).

Aljlayl, M. and Frieder, O. (2001). "Effective Arabic-English Cross-Language Information Retrieval via Machine Readable Dictionaries and Machine Translation." *ACM Tenth Conference on Information and Knowledge Management*.

Beagrie, Neil (2005). "Plenty of Room at the Bottom? Personal Digital Libraries and Collections." *D-Lib Magazine* 11.6, http://dlib.anu.edu.au/dlib/june05/beagrie/06beagrie.html

Bearman, David and Jennifer Trant (2005). "Social Terminology Enhancement through Vernacular Engagement: Exploring Collaborative Annotation to Encourage Interaction with Museum Collections." *D-Lib Magazine* 11.9, http://www.dlib.org/dlib/september05/bearman/09bearman.html.

Billsus, Daniel, David H. Hilbert and Dan Maynes-Aminzade (2005). "Improving Proactive Information Systems." *IUI '05: Proceedings of the 10th International Conference on Intelligent User Interfaces*, pp. 159-66.

Bøe, Hilde, Jon Gunnar Jørgensen and Stine Brenna Taugbø (2004). "Philology Meets Text Encoding in the New Scholarly Edition of Henrik Ibsen's Writings." *Literary and Linguistic Computing* 19.1, pp. 55-71.

Bozzi, Andrea and Sylvie Calabretto (1997). "The Digital Library and Computational Philology: The BAMBI Project." *Lecture Notes in Computer Science 1324, Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pp. 269-85.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith (2002). "The TIGER Treebank." *Proceedings of the Workshop on Treebanks and Linguistic Theories* (Sozopol, Bulgaria).

Brown, Jonathan and Maxine Eskenazi (2004). "Retrieval of Authentic Documents for Reader-Specific Lexical Practice." *Proceedings of InSTIL/ICALL Symposium 2004*.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (1990). "A Statistical Approach to Machine Translation." *Computational Linguistics* 16.2, pp. 79-85.

Brusilovsky, Peter (2002). "From Adaptive Hypermedia To The Adaptive Web," *Communications of the ACM* 45.5, pp. 30-3.

Carlquist, Jonas (2004). "Medieval Manuscripts, Hypertext and Reading. Visions of Digital Editions." *Literary and Linguistic Computing* 19.1, pp. 105-118.

Cheng, Pu-Jen, Jei-Wen Tang, Ruei-Cheng Chan, Jenq-Haur Wang, Wen-Hsiang Lu and Lee-Feng Chien (2004). "Translating Unknown Cross-Lingual Queries in Digital Libraries Using a Web-Based Approach." *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 146-53.

Chiou, Fu-Dong, David Chiang, and Martha Palmer (2001). "Facilitating Treebank Annotation Using a Statistical Parser." *Proceedings of the First International Conference on Human Language Technology Research HLT '01,* pp. 1-4.

Church, Kenneth W. and Eduard H. Hovy (1993). "Good Applications for Crummy Machine Translation," *Machine Translation* 8.4, pp. 239-58.

Cohen, Daniel and Roy Rosenzweig (2005). "Web of Lies? Historical Knowledge and the Internet?" *First Monday* 10.12, http://firstmonday.org/issues/issue10_12/cohen/index.html.

Cohen, Daniel (2006). "From Babel to Knowledge: Data Mining Large Digital Collections." *D-Lib Magazine* 12.3, http://www.dlib.org/dlib/march06/cohen/03cohen.html.

Crane, Gregory (1996). "Building A Digital Library: the Perseus Project as a Case Study in the Humanities." *Proceedings of the First ACM International Conference on Digital Libraries*, pp. 3-10.

Crane, Gregory (2004). "Classics and the Computer: An End of the History." Susan Schreibman, Ray Siemens and John Unsworth (eds.), *A Companion to the Digital Humanities* (Oxford: Blackwell Publishing).

Crane, Gregory (1991). "Generating and Parsing Classical Greek." *Literary and Linguistic Computing* 6.4, pp. 243-5.

Crane, Gregory (1998). "New Technologies for Reading: the Lexicon and the Digital Library." *Classical World* 92, pp. 471-501.

Crane, Gregory (2005b). "No Book is an Island: Designing Electronic Primary Sources And Reference Works for the Humanities." H. van Oostendorp, Leen Breure, Andrew Dillon (eds.), *Creation, Use, and Deployment of Digital Information* (Erlbaum 2005), pp. 11-26.

Crane, Gregory (2005). "Reading in the Age of Google: Contemplating the Future with Books That Talk to one Another," *Humanities* 26.5, http://www.neh.gov/news/humanities/2005-09/readingintheage.html.

Crane, Gregory (2006). "What Do You with a Million Books." *D-Lib Magazine* 12.3, http://dlib.ejournal.ascc.net/dlib/march06/crane/03crane.html.

Crane, Gregory and Alison Jones (2006). "Text, Information, Knowledge and the Evolving Record of Humanity." *D-Lib Magazine* 12.3, http://purl.pt/302/1/dlib/march06/jones/03jones.html.

Crane, Gregory, Kalina Bontcheva, Jeffrey A. Rydberg-Cox, and Clifford Wulfman (2005). "Emerging Language Technologies and the Rediscovery of the Past: a Research Agenda." *International Journal on Digital Libraries* 5.4, pp. 309-316.

Dang, Trang Hoa (2005). "Overview of DUC 2005." http://duc.nist.gov/pubs/2005papers/OVERVIEW05.pdf.

Dolog, Peter, et. al. (2004), "The Personal Reader: Personalizing and Enriching Learning Resources Using Semantic Web Technologies." *Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pp. 85-94.

Elango, Aravaind, Michael Nelson and Johan Bollen (2004). "Dynamic Linking of Smart Digital Objects Based on User Navigation Patterns." *The Computing Research Repository*, http://arxiv.org/abs/cs.DL/0401029

Fisher, Michelle and Richard Everson (2003). "Representing Interests As A Hyperlinked Document Collection." *CIKM '03: Proceedings of the Twelfth International Conference on Information And Knowledge Management*, pp. 378-385.

Fleming, Michael (2004). "The Use of Increasingly Specific User Models in the Design of Mixed-Initiative Systems." *Proceedings of Advances in Artificial Intelligence: 17th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI*, pp. 434-8

Gemmell, Jim, Gordon Bell, and Roger Lueder (2006). "MyLifeBits: A Personal Database for Everything." *Communications of the ACM* 49.1, pp. 88-95.

Gey, Fredric C., Noriko Kando and Carol Peters (2005). "Cross Language Information Retrieval: the Way Ahead." *Information Processing & Management* 41.3, pp. 415-31.

Hajic, Jan (1999). "Building a Syntactically Annotated Corpus: The Prague Dependency Treebank." E. Hajicová (ed.), *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevova* (Prague: Charles University Press).

Hahn, Udo and Inderjeet Mani (2000). "The Challenges of Automatic Summarization." *Computer* 33.11, pp. 29-36.

Hardwick, Lorna (2000). "Electrifying the Canon: The Impact of Computing on Classical Studies." *Computers and the Humanities* 34, pp. 279-95.

Henze, Nicola and Wolfgang Nejdl (2002). "Knowledge Modeling for Open Adaptive Hypermedia." *Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pp. 174-183.

Hotho, A., S. Staab, and G. Stumme (2003). "WordNet Improves Text Document Clustering." *Proceedings of the SIGIR 2003 Semantic Web Workshop*.

Ide, Nancy, and Jean Véronis (1998). "Word Sense Disambiguation: The State of the Art." *Computational Linguistics* 24.1, pp. 1-40.

Kateli, Behzad and Liddy Neville (2005). "Interpretation and Personalisation: Enriching Individual Experience By Annotating On-Line Materials." Museums and the Web 2005, http://www.archimuse.com/mw2005/papers/kateli/kateli.html.

Kavcic, Alenka (2004). "Fuzzy User Modeling for Adaptation in Educational Hypermedia." *IEEE Transactions on Systems, Man and Cybernetics*, *Part C* 34.4, pp. 439-449.

Kilgarriff, A. (1998), "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs." *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pp. 581-585.

Kim, Seonho and Edward A. Fox (2004). "Interest-Based User Grouping Model for Collaborative Filtering in Digital Libraries." *7th International Conference of Asian Digital Libraries, 13-17 Dec 2004, Shanghai, China,* pp. 533–542

Kishida, Kazuaki (2005). "Technical Issues of Cross Language Information Retrieval: A Review." *Information Processing & Management* 41.3, pp. 433-455.

Kolbitsch, Josef and Hermann Maurer (2005). "Community Building Around Encyclopaedic Knowledge." *Journal of Computing and Information Technology*, to appear. Online at: http://www.iicm.edu/iicm_papers/community_building_around_encyclopaedic_knowledge.pdf.

Koutrika, Georgia and Yannis E. Ioannidis (2004). "Rule-based Query Personalization in Digital Libraries," *International Journal of Digital Libraries* 4.1, pp. 60-3.

Krowne, Aaron (2003). "Building a Digital Library the Commons Based Peer Production Way." *D-Lib Magazine*, 9.10, http://www.dlib.org/dlib/october03/krowne/10krowne.html.

Krowne, Aaron and Martin Halbert (2005). "An Initial Evaluation of Automated Organization for Digital Library Browsing." *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 246-255.

Kurohashi, S., and M. Nagao (1998). "Building a Japanese Parsed Corpus while Improving the Parsing System." *Proceedings of the First International Conference on Language Resources and Evaluation*.

Lafleur, Richard A. (2000). "Latin and Greek Enrollments in America's Schools and Colleges." *ADFL Bulletin* 31.3, pp. 53-8.

Lai, Catherine and Steven Bird (2004). "Querying and Updating Treebanks: A Critical Survey and Requirements Analysis." *Proceedings of the Australasian Language Technology Workshop*.

Lambert, Patrik and Núria Castell (2004). "Alignment of Parallel Corpora Exploiting Asymmetrically Aligned Phrases." *Proceedings of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*.

Lampe, Cliff and Paul Resnick (2004). "Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 443-550.

Latousek, Rob (2001). "Fifty Years of Classical Computing: A Progress Report." *CALICO Journal*, 18.2, pp. 211-22

Marcu, Daniel and Kevin Knight (2005). "Machine Translation in the Year 2004," in *Proceedings of Acoustics, Speech and Signal Proceedings (ICASSP 2005*), Volume 5, pp. 965-8.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). "Building a Large Annotated Corpus of English: the Penn Treebank." *Computational Linguistics* 19.2, pp. 313-330.

Marshall, Catherine C. (2005). "Reading and Interactivity in the Digital Library: Creating an Experience that Transcends Paper." Deanna Marcum and Gerald George (eds.), *Digital Library Development: The View from Kanazawa* (Westport, Connecticut: Libraries Unlimited), pp. 127-145

Maybury, Mark, ed. (2004). *New Directions in Question Answering* (Cambridge: MIT Press).

McCallum, Andrew (2005). "Information Extraction: Distilling Structured Data from Unstructured Text." *ACM Queue* 3.9, pp. 48-57.

McManus, Barbara F. and Carl A. Rubino (2003). "Classics and Internet Technology." *American Journal of Philology* 124.4, pp. 601-8

Mihalcea, Rada and Michel Simard (2005). "Parallel Texts." *Natural Language Engineering* 11.3, pp. 239-46.

Montemagni, S., F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, R. Delmonte (2000). "The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation." *Proceedings of the COLING Workshop on "Linguistically Interpreted Corpora (LINC-2000)."*

Moreno, A., R. Grishman, S. López, F. Sánchez and S. Sekine (2000). "A Treebank of Spanish and its Application to Parsing." *Proceedings of the Second Conference on Language Resources and Evaluation.*

Newby, G. B. and C. Franks (2003). "Distributed Proofreading." *Proceedings of the Joint Conference on Digital Libraries,* pp. 361–363.

Ney, Hermann (2005). "One Decade of Statistical Machine Translation: 1996-2005." *Proceedings of the MT Summit X*, pp. i-12 - i-17

Nichols, Eric, Francis Bond and Daniel Flickinger (2005). "Robust Ontology Acquisition from Machine-Readable Dictionaries," *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1111-1116.

Niederée, Claudia, et al. (2004). "A Multi-Dimensional, Unified User Model For Cross-system Personalization." *Proceedings of Advanced Visual Interfaces International Working Conference (AVI 2004) -Workshop on Environments for Personalized Information Access*, pp. 34-54.

Papatheodorou, Christos, Sarantos Kapidakis, Michalis Sfakakis and Alexandra Vassiliou (2003). "Mining User Communities in Digital Libraries." *Information Technology and Libraries* 22.4, pp. 152–157.

Peters, Carol, Martin Braschler, Khalid Choukri, Julio Gonzalo, and Michael Kluck (2004). "The Future of Evaluation for Cross-Language Information Retrieval Systems." *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pp. 841-844.

Pouliquen, Bruno, Ralf Steinberger, Camelia Ignat, and Tom De Groeve (2004). "Geographical Information Recognition and Visualisation in Texts Written in Various Languages." *Proceedings of the 2004 ACM Symposium on Applied Computing*, pp. 1051-8.

Renda, M. Elena and Umberto Straccia (2005). "A Personalized Collaborative Digital Library Environment: A Model and an Application." *Information Processing and Management* 41.1, pp. 5–21.

Rosen, Victoria, Koenraad De Smedt, Helge Dyvik and Paul Meurer (2005). "TREPIL: Developing Methods and Tools for Multilevel Treebank Construction." *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories, TLT 2005*, pp. 161-72.

Rouane, Khalid and Claude Frasson and Marc Kaltenbach (2003). "Reading for Understanding: A Framework for Advanced Reading Support." *Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies*, pp. 394-5.

Russell, John (2003). "Making It Personal: Information That Adapts To The Reader." *SIGDOC '03: Proceedings of the 21st Annual International Conference on Documentation*, pp. 160–166.

Rydberg-Cox, Jeffrey A., Lara Vetter, Stefan Rüger, and Daniel Heesch (2004). "Cross-Lingual Searching and Visualization for Greek and Latin and Old Norse Texts." *Proceedings of The 4th ACM/IEEE-CS Joint Conference On Digital Libraries*, p. 383.

Sengupta, Arjit, Mehmet Dalkilic, and James Costello (2004). "Semantic Thumbnails: A Novel Method for Summarizing Document Collections." *Proceedings of the 22nd Annual International Conference on Design of Communication: The Engineering of Quality Documentation*, pp. 45-51.

Sinclair, Stéfan (2003). "Computer Assisted Reading: Reconceiving Text Analysis." *Literary & Linguistic Computing* 18.2, pp. 175-84.

Smith, David A. and Gregory Crane (2001). "Disambiguating Geographic Names in a Historical Digital Library." *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'01*), *Lecture Notes in Computer Science*, pp. 127–136.

Smith, David A. (2006). "Debabelizing Libraries: Machine Translation by and For Digital Collections." *D-Lib Magazine* 12.3, http://www.dlib.org/dlib/march06/smith/03smith.html.

Smith, David A. (2002). "Detecting Events with Date and Place Information in Unstructured Text." *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 191-196.

Spencer, Matthew and Christopher Howe (2004). "Collating Texts Using Progressive Multiple Alignment," *Computers and the Humanities* 38.3, pp. 253-70.

Steinbach, Michael, George Karypis, and Vipin Kumar (2000). "A Comparison of Document Clustering Techniques." *KDD Workshop on Text Mining*. http://www.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf.

Tanaka, Takaaki, Francis Bond, Stephan Ooepen, and Sanae Fujita (2005). "High Precision Treebanking—Blazing Useful Trees Using POS Information." *Proceedings of the 43rd Annual Meeting of the ACL,* pp. 330–337.

Teevan, Jaime, Susan T. Dumais and Eric Horvitz (2005). "Personalizing Search via Automated Analysis Of Interests And Activities." *Proceedings of the 28th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval*, pp. 449-456

Terras, Melissa (2005). "Reading the Readers: Modelling Complex Humanities Processes to Build Cognitive Systems." *Literary and Linguistic Computing*, 20.1, pp. 41-59.

Thiel, Ulrich, Holger Brocks, Andrea Dirsch-Weigand, André Everts, Ingo Frommholz, and Adelheit Stein (2005). "Queries in Context: Access to Digitized Historic Documents in a Collaboratory for the Humanities." *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments*, *Lecture Notes in Computer Science*, Volume 3379, pp. 117-27.

Tiedemann, Jörg (2005). "Optimization of Word Alignment Clues*." Natural Language Engineering* 11.3, pp. 279-93.

Wang, Chin Yeh and Gwo-Dong Chen (2004). "Extending E-Books with Annotation, Online Support and Assessment Mechanisms to Increase Efficiency of Learning." *SIGCSE Bulletin,* 36.3, pp. 132-136.

Weibalzahl, Stephan and Gerhard Weber (2002). "Adapting to Prior Knowledge of Learners." *Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, (AH2002*), pp. 448-451.

Willinsky, John (2003). "Opening Access: Reading (Research) in the Age of Information." C. M. Fairbanks, J. Worthy, B. Maloch, J. V. Hoffman, & D. L. Schallert (eds.), *51st National Reading Conference Yearbook* (Oak Creek, WI: National Reading Conference), pp. 32-46.

Zhou, Liang, Miruna Ticrea and Eduard Hovy (2004). "Multi-Document Biography Summarization." *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 434-41.