Preprint of paper accepted to the JCDL 2007.

A New Generation of Textual Corpora

Mining Corpora from Very Large Collections

Gordon Stewart
Harvard University
Language Resource Center
Cambridge, MA 02138
stewart5@fas.harvard.edu

Gregory Crane
Tufts University
Perseus Project
Medford, MA 02155
gregory.crane@tufts.edu

Alison Babeu
Tufts University
Perseus Project
Medford, MA 02155
alison.jones@tufts.edu

ABSTRACT

While digital libraries based on page images and automatically generated text have made possible massive projects such as the Million Book Library, Open Content Alliance, Google, and others, humanists still depend upon textual corpora expensively produced with labor-intensive methods such as double-keyboarding and manual correction. This paper reports the results from an analysis of OCR-generated text for classical Greek source texts. Classicists have depended upon specialized manual keyboarding that costs two or more times as much as keyboarding of English both for accuracy and because classical Greek OCR produced no usable results. We found that we could produce texts by OCR that, in some cases, approached the 99.95% professional data entry accuracy rate. In most cases, OCR-generated text yielded results that, by including the variant readings that digital corpora traditionally have left out, provide better recall and, we argue, can better serve many scholarly needs than the expensive corpora upon which classicists have relied for a generation. As digital collections expand, we will be able to collate multiple editions against each other, identify quotations of primary sources, and provide a new generation of services.

Categories and Subject Descriptors: H.3.7 Information Systems: Information Storage and Retrieval [digital libraries]

General Terms: Measurement, Documentation

Keywords: OCR evaluation, Ancient Greek, text alignment

1. INTRODUCTION

Image front collections with industrially scanned page images and automatically generated text such as JSTOR and the Making of America spawned large scale digital libraries based on digitized print such as the Million Book Library, the Open Content Alliance, Google Book Search and Microsoft Book Search. Humanists still, however, largely depend upon primary sources that have been manually keyboarded and corrected. These digital corpora have been expensive to produce, usually leave out scholarly information sources such as textual notes about variant readings, and contain only a single edition of each work. We report here on the results of OCR for classical Greek – a field in which no one has, to our knowledge, produced usable corpora based on OCR output. We found to our surprise that we could not only produce useful output but that we could support searching that exceeded the recall of the manually produced corpora on which classicists have depended for a generation. Our results consider three base cases. First, even when we work with OCR-generated text from difficult source material (e.g., a mid-nineteenth century edition of Aristotle in a non-standard Greek font), searching OCR-generated text provides superior recall because the OCR-generated text includes many variant readings. Errorful OCR output of text and variants provides better searching than perfect transcription of the reconstructed text alone.[1] Second, we could correct 50% of errors in individual texts, bringing accuracy of character transcription up from 99.72% to 99.87% in a text with a modern Greek font. Third, when we compared the output of two OCR engines on two editions of the same text, we were able to reach an accuracy of 99.93%, thus approaching the 99.95% accuracy standard in professional data entry.

These results point to a new generation of digital primary sources. These new textual corpora will provide accuracy comparable to hand-crafted corpora on which humanists have relied for a generation but will also build upon the industrial methods and scale of the very large digital collections now taking shape. While these corpora will lack much of the hand-crafted markup of manually produced corpora, they will include variant readings (which almost all large curated corpora leave out) and multiple editions (which almost no large curated collections contain), and thus will better serve many needs of scholarship than their smaller, more expensive predecessors.

In previous publications, we have described work on the automatic classification (e.g., Washington as a person vs. a place) and identification (e.g., Washington, PA, vs. Washington DC) of various named entities (people, places, organizations, dates, times, and other numerical expressions)[13, 38]. This paper takes its departure from [17], which used

[1]By reconstructed text we mean the particular version which an editor constructs from multiple variants. Figure 1 shows part of the reconstructed text and all of the variants on a typical page.

JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.

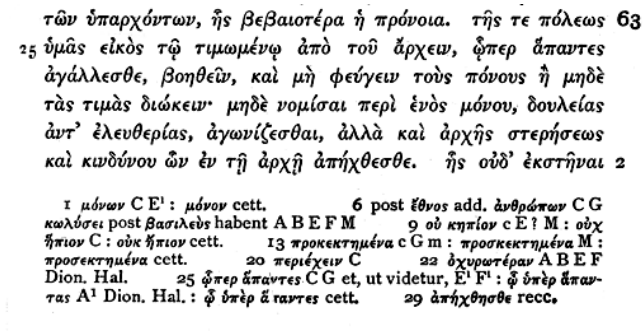
generic texts from Project Gutenberg to evaluate the results of OCR on different editions of the same work. This paper demonstrated that even when editions differed in various points, they could be automatically aligned with each other and their differences used to identify errors.

Since many important texts have appeared in multiple editions – and, indeed, the more heavily a text has been studied, the more numerous its editions – we speculated whether we could use multiple editions of a single text, online within larger digital collections, as sources with which to correct OCR errors that appeared in one text and not in another. Beyond simple error correction we wondered how effectively we could automatically collate multiple editions against one another. Students of literature would benefit from being able to see precisely how different versions of a work changed over time, visualizing differences, calculating their significance as they compare the relatedness of multiple versions of a work.

In the typical case, general approaches are not enough. We need a layer between general alignment/text comparison algorithms and the data. In texts that predate spelling conventions we need to be able to distinguish, with reasonable accuracy, among orthographic variants on the same word ("sonne"/"son"), significant variants between editions (e.g., "son" vs. "sun") and errors (e.g., "scn"). When texts are in historical languages, we need to manage not only orthographic conventions and genuine variants, but special errors that occur when modern OCR assumes it is processing English: the common Latin word *tum* ("then"), for example, often becomes *turn*, as automatic error correction attempts to generate plausible English out of the Latin source. A mature cyberinfrastructure, should, we believe, contain a library of knowledge bases which it can match automatically to whatever text it is processing. The digital library should be able to distinguish early modern Latin poetry from articles on machine translation and bring to bear its resources (e.g., lexica, gazetteers, term lists) to analyze the content of each.

We chose classical Greek as the initial focus of our work. For the first generation of digital work in classics, all Greek text had to be entered by hand, with accented Greek being two to four times more expensive than English (in our experience generally $1,000-2,000 per megabyte for 99.95% accuracy). Classicists have for decades had access to multiple databases of classical Greek texts: the Perseus Digital Library contains the most widely read classical Greek authors while the Thesaurus Linguae Graecae (TLG) has a comprehensive collection of Greek texts that now extends past the classical period. These collections, however, offered only partial solutions to the general problem of scholarship. At the very least, three challenges remain.

1. Large corpora only include the reconstructed text (defined above) and do not include the variant readings. This was a major compromise made for a variety of reasons. First, while many scholars do not feel that the reconstructed text of Sophocles or Vergil was subject to copyright, many felt that the textual notes assembled by scholars were copyrightable. Thus, entering the reconstructed text seemed acceptable from a moral perspective. Second, encoding variants is a laborious task: the scholarly annotation is often ambiguous and requires considerable background knowledge. Very accurate markup of large bodies of textual notes seemed impractical. Third, most text searching systems have no infrastructure for variants. Since variants account for between 14% and 3% of the text on a page (see below), perfect recall on the reconstructed text yields results between 86% and 97%.


τῶν ὑπαρχόντων, ἧς βεβαιοτέρα ἡ πρόνοια. τῆς τε πόλεως 63
ὑμᾶς εἰκὸς τῷ τιμωμένῳ ἀπὸ τοῦ ἄρχειν, ᾧπερ ἅπαντες
ἀγάλλεσθε, βοηθεῖν, καὶ μὴ φεύγειν τοὺς πόνους ἢ μηδὲ
τὰς τιμὰς διώκειν· μηδὲ νομίσαι περὶ ἑνὸς μόνου, δουλείας
ἀντ' ἐλευθερίας, ἀγωνίζεσθαι, ἀλλὰ καὶ ἀρχῆς στερήσεως
καὶ κινδύνου ὧν ἐν τῇ ἀρχῇ ἀπήχθεσθε. ἧς οὐδ' ἐκστῆναι 2

1 μόνων C E¹ : μόνον cett. 6 post ἔθνος add. ἀνθρώπων C G κωλύσει post βασιλεὺς habent A B E F M 9 οὐ κηπίον c E ? M : οὐχ ἥπιον C : οὐκ ἥπιον cett. 13 προκεκτημένα c G m : προσκεκτημένα M : προσεκτημένα cett. 20 περιέχειν C 22 ὀχυρωτέραν A B E F Dion. Hal. 25 ᾧπερ ἅπαντες C G et, ut videtur, E¹ F¹ : ᾧ ὑπὲρ ἅπαντας A¹ Dion. Hal. : ᾧ ὑπὲρ ἅ παντες cett. 29 ἀπήχθησθε recc.

Figure 1: Bottom section of a page from the Greek historian Thucydides. The reconstructed text contains 255 words. The textual notes contain 26 Greek words, of which 18 are variants distinct from the reconstructed text. The text of Thucydides is fairly well-established but even on a randomly selected text of this author the reconstructed text contains only 93.5% of the relevant data. In a preliminary survey we found that in scholarly editions only 86% of the words on a page were in the reconstructed text. In the Loeb Classical Library, which traditionally minimized the variants which it cited (assuming that serious scholars would consult more elaborate editions), we found that the reconstructed text contained only 97% of the words on a page.

2. Multiple editions are not available and no infrastructure exists with which to manage them if they did exist. The TLG – the flagship digital library for classical texts and, after thirty-five years of operation, one of the most established textual corpora in any discipline – systematically replaces older editions with new ones. The TLG's *Canon of Greek Literature* contains only the single edition which the TLG chose to include in its collection. Not only has James Diggle's editon of Euripides replaced that of Gilbert Murray, which was part of the early shipments of TLG texts on magnetic tape (before CD ROMs had appeared), but the TLG's bibliography does not even list the older Murray edition. The only indication that we can find of the earlier edition is in an on-line list of "suppressed" texts with the notation "substituted by new editions."[2] An earlier print edition of the TLG *Canon of Greek Literature* confirms this.

   Preliminary comparison suggests that c. 1-10% of the words in differing editions of the same work vary. In authors where the textual tradition is well-established and we are fairly confident about the reconstructed text (e.g., Plato, Isocrates), editions will vary little from one another. Where the manuscript tradition is garbled or contains many variants, consensus is much

[2]http://www.tlg.uci.edu/CDEworks.html: accessed January 17, 2007.

lower. The reconstructed texts of some widely studied Greek authors (e.g., Aeschylus) vary substantially.

Some (and probably most) scholars are even more interested in the variance between published editions than variants that an editor may have considered worth noting. Put another way, comparison of reconstructed texts effectively filters the potential variants and corrections, allowing readers to see which variants particular editors chose to include in their reconstructed texts. Automatic comparison of editions would allow us to calculate and visualize the relationship between editions of an author as they evolved over the centuries, enabling us to see which editions exerted the greatest influence, which corrections suggested by a modern scholar were adopted by others and other patterns.

3. While scholarly editions are the foundation for textually based work, we also want to know where other works quote particular passages from our source texts. Thus, if we are puzzling over a particular idiom in a particular passage, for example, we would like to know if a dictionary discusses this same passage – the Liddell Scott Jones Greek-English and Lewis and Short Latin-English lexica already in the Perseus Digital Library alone contain 250,000 and 225,000 quotations each. At the same time, many primary and secondary sources quote earlier literature. Literary texts and scholars, ancient and modern, have been quoting earlier texts for thousands of years. Identifying who quotes which passages of Vergil allows us to trace the influence of this Latin poet in subsequent literature and to identify relevant passages of modern scholarship. In each case, analyzing the context of the quotations opens up possibilities for text mining to identify trends in what the quoting authors say about these passages.

2. BACKGROUND

2.1 Generations of Digital Corpora

The history of electronic texts begins in postwar Italy, where the Jesuit Roberto Busa collaborated with IBM to digitize and index the works of Saint Thomas Aquinas.[3] The first major project that set out to create a digital library of works from a range of authors was Project Gutenberg, founded in 1971 and still on-going.[4] Project Gutenberg is a community driven project, with minimal external funding. At least two features characterize Project Gutenberg and efforts like it (such as, in classics, theLatinLibrary.com). First, the goals are pragmatic. Project Gutenberg sets out to create readable texts and chose not to concentrate on the factors on which scholars would focus but that make little difference to the majority of readers. Thus, footnotes and introduction may be ignored and an electronic text may not even cite the print edition on which it is based. Most readers of nineteenth century fiction, for example, don't care where their Moby Dick comes from, so long as it accurately transcribes some version of what Melville produced. Second, the work is highly decentralized, with individuals around the world contributing entire works. More recently, Project Gutenberg developed a pioneering approach in which individuals could sign up to proofread individual pages. Thus, Project Gutenberg was able to tap into the energy of the larger group willing to contribute smaller units of labor than entire books.

Classics was arguably the first academic discipline to create a systematic collection of digital texts to support scholarly analysis. The TLG began work on a corpus of classical Greek in 1972, before the Winchester drive introduced modern disk storage, a decade before personal computers became widespread and two decades before the internet emerged as a major force in the intellectual life of society. Fifteen years later the Packard Humanities Institute (PHI) created a digital library for classical Latin, producing on a CD ROM a Latin counterpart to the TLG. Both are available on CD ROM and both share the same encoding scheme, BetaCode, which is a page layout language for Greek, Latin and other ancient languages. While other Latin databases have emerged, the TLG and PHI collections have remained fundamental tools in classical studies. At least two features distinguish these projects from community driven efforts such as Project Gutenberg and theLatinLibrary.com. First, the TLG and PHI invested heavily in professional data entry, where comparison of independent keyboarding enables firms to guarantee accuracy of 99.95% or higher. Second, classicists checked all texts in the TLG and PHI, not only correcting transcriptional errors but adding a consistent markup scheme. Classicists devised BetaCode to capture not only ancient languages such as Greek and Coptic but to capture the basic page layouts and citation schemes of scholarly editions.[5] The citation schemes were probably more important than the formatting information, since they allowed scholars to search and browse with the various citation schemes by which they mapped their texts (e.g., line number, book-chapter-section).

In the 1980s, a third class of corpora began to emerge. Like their predecessors, these corpora depended upon highly accurate, manual data entry processes (e.g., double keying, cleanup of OCR output or of single keying). Unlike first generation corpora in classics, these corpora adopted SGML/XML as the syntax for their markup and based their markup language on the Text Encoding Initiative Guidelines. This allowed corpus developers to begin encoding semantic information (e.g., marking as text a Latin quotation that happens to be in italics rather than as simply a chunk of italics). These efforts include the Perseus Digital Library of Greek and Latin source texts, the SGML collections in American Memory, the DocSouth collections of Southern Culture at UNC, the Early English Books On-line Text Creation Partnership, the recently released Perseus American collection and many others. Adoption of standardized markup and public DTDs made long term preservation of individual documents and collections far more practical.[11] The ability to encode semantic information (e.g., this string is Latin) rather than page layout (e.g., this string is in italics) began to make possible more powerful queries (e.g., extract all quoted Latin).

In the 1990s, however, the decreasing cost of disk storage and the rise of the internet made possible a fourth strategy, based on storing whole page images and searching text au-

[3]http://www.corpusthomisticum.org;
http://en.wikipedia.org/wiki/Roberto_Busa.
[4]http://www.gutenberg.org/wiki/Main_Page

[5]http://www.tlg.uci.edu/quickbeta.pdf

tomatically generated by OCR. Image-front systems present the user with the page image in front, with the searchable text in the background, where it may be accessible to the reader or hidden. Image-front collections could not, in the general case, approach the accuracy of carefully edited collections and they generally make no attempt at adding semantic markup within the text, with JSTOR providing a notable exception.[6] Image front collections stress quantity over quality and they have been immensely successful because they provide good enough results to serve most users most of the time. For expository prose in modern English print (almost anything printed in the past two hundred years without the long 's' which resembles an 'f') commercial OCR systems provide output that is good enough to provide satisfactory searching. The Cornell-Michigan Making of America and the non-profit JSTOR scholarly journal archive popularized this approach in the 1990s. This image-front strategy has made possible the massive digital libraries being developed by the European Union, Open Content Alliance, Microsoft and Google.

This paper focuses on a fifth class of collection, now becoming feasible, which attempts to synthesize the strengths of the other four classes with an extensible workflow. A number of features characterize these fifth generation corpora:

- They are decentralized, accepting contributions, large and small, from contributors around the world. They thus provide mechanisms whereby communities may fully exploit, augment and drive the strategic goals of the collection. In this they resemble first generation corpora and avoid the centralization which has tended to ossify collections when major funding ends or the product is "good enough" to generate subscription revenue. In the Text Creation Partnerships of Michigan, a centralized production team creates initial texts and then passes these on to a broader scholarly community.[7] The Christian Classics Ethereal Library allows its users to correct errors in and add markup to individual pages via a web interface.[8] In the Perseus Digital Library, users can contribute corrections to automatically generated morphological analyses [10, 9]. The Distributed Proofreaders of Project Gutenberg continues to produce clean copy for increasingly complex texts [28].

- They relentlessly pursue automated methods to generate scalable, semantic markup. Part-of-speech taggers, morphological analysis and named entity identification are three well-established methods of adding markup to large corpora. Citeseer has for years automatically identified bibliographic references, author names, titles and other core elements of structural markup. In March 2006 Perseus published a 55 million word corpus of American English in which we have automatically tagged, among 12 million automatically generated annotations, 1.5 million personal names, 1 million places, 600,000 dates, and 500,000 organizations [13, 24, 12]. As these tools grow more sophisticated, as markup becomes more expressive and extensible (e.g., TEI P5) and as user expectations rise, "annotation factories" are beginning to emerge.[9]

- They synthesize the scholarly demands of capital intensive, manually constructed collections like the classical corpora in Perseus, the TLG and the PHI Latin CD ROM with the industrial scale of very large "million book" libraries now emerging. This paper focuses on this challenge. No project with which we are aware has ever made any effective use of automatically generated OCR text of classical Greek. We discovered that we could not only produce text that was good enough for general information retrieval but that we could create, at scale, digital versions of printed editions that rivaled the transcriptional accuracy of professional data entry and potentially provide better support for the study of Greek and Latin than the manually produced corpora on which scholars depend.

3. RELATED WORK

The work reported in this paper is similar to research in a number of related research areas: OCR and document recognition for historical documents, the alignment and collation of textual variants and multiple editions, and general work in parallel text alignment. The Perseus Project has previously explored the issue of supporting encoding for textual variation and multiple editions [37], but our work here draws most directly on the work of [17]. Other work has also focused on the use of ancient texts such as the Bible as a testground for OCR techniques [20].

The problems of generating usable text and knowledge from images of ancient manuscripts particularly in challenging languages such as Greek and Arabic has been reported previously by a number of researchers, including some recent overviews of the issues by [7], [2] and [22]. Rawat, et. al. have developed an interactive system that continuously improves the results of an OCR system developed for large document image collections in Indian languages [32]. Similarly, [6] developed a model for ancient document recognition that combined several OCRs with a specialized intelligent character recognition based on neural networks, which improved the recognition of rejected characters by almost 5%. Experiments in OCR and text alignment have also been reported by [4] who used specialized OCR and alignment of texts to assist them in the automatic indexing and reformulation of historical French dictionaries. Other research conducted by [21] has explored methods of indexation such as word spotting and computer assisted transcription for when OCR does not produce usable results. Some specific work has also been reported on the development of OCR techniques for handwritten manuscripts in Old Greek [18, 29], but our work has focused on the use of OCR with typeset editions from the late nineteenth and early twentieth centuries.

A number of digital humanities projects have also collected and aligned multiple editions of works, although they have typically been focused on the works of one author. The Canterbury Tales Project[10] has conducted a variety of important research in this area, including exploring the issues of presenting a large number of variant editions and manuscripts [34]. Other important projects that have explored

[6]http://www.jstor.org/about/recent-issues.html
[7]http://www.lib.umich.edu/tcp/
[8]http://www.ccel.org/
[9]http://gate.ac.uk/sale/talks/salzburg06.html
[10]http://www.canterburytalesproject.org/

these issues include the Blake Archive,[11] the Decameron Web,[12] and the Cervantes Project, [13] to name only a few. The Cervantes project in particular has spent a number of years exploring the issues involved in the alignment of different versions of texts by Cervantes and how to visualize these results within a digital library [1, 43]. This work has included the development of an Electronic Variorum Edition of *Don Quixote* and both a multi-variant editor for documents and interactive timeline viewer to visualize the variants [26, 25].

A variety of other researchers have also developed tools or interfaces to support the visualization of variant texts. Schreibman, et. al. have developed the Versioning Machine, a software tool that allows users to compare different versions of a text and to view different textual witnesses side by side [36]. The NINES Project has created an open source tool called JUXTA which allows users to compare multiple editions of a text [16]. Similar research has also been reported by Schmidt and Wyeld [35] who have created an interface to visualize multiple editions of documents. Additionally, the Active Reading Project is working to create an electronic scholarly edition of King Lear that will allow users to visualize the textual variants between different editions of the work [41].

Our research has also drawn on general knowledge from various experiments reported on parallel text alignment, including the alignment of multiple editions or of editions with their multiple translations. Parallel text alignment is frequently used for machine translation systems [15, 23, 42]. Text alignment can also be used for monolingual corpora to support various tasks such as summarization [27, 3]. [30] have reported on various experiments with the automatic alignment of multiple texts including various English language translations of classical texts by authors such as Homer. Ghorbel et. al. explored using a variety of heuristics including lexical, morphological, syntactic and semantic similarities to align prose and verse versions of medieval texts [19]. A variety of research has also focused on developing encoding schemes and algorithms that support more sophisticated collation of variant editions and texts [40, 5, 39], often with the purpose of creating a digital scholarly edition. Riva and Zafrin have also explored the importance of creating more sophisticated digital editions that represent the variants works and texts of an author [33]. Finally, important research into how to create image based editions that link manuscript images with transcripts, translations, and other files has been reported by [14].

4. METHODOLOGY

Classical Greek has several accent marks. These have bedeviled students for thousands of years, but they are later additions to the text and were not present in what Sophocles and Plato wrote – these accents are, in fact, a form of text markup added by later scholars. In our twenty years of development, we have always assumed that we were searching cleanly entered, full accented Greek. But because one of our two Greek OCR engines was designed for modern Greek and could not process classical accents, we examined the impact of ignoring the accents in searching. In an initial sample of 79 unique words randomly selected from a Greek text, ignoring accents meant that our morphological analyzer generated 1.25 possible dictionary entries for each inflected form with accents and 1.3 dictionary entries without accents. The accentuation was much more significant in reducing the ambiguity of morphology – in our sample, we generated 1.30 morphological analyses per inflected forms with accents but 2.35 analyses when we ignored accents – an increase of 80%. If we are supporting image-front searches of Greek, ignoring accents has a marginal impact on precision and no impact on recall.

4.1 Simple Retrieval of Uncorrected Text

We employed a multi-tiered approach to optical character recognition (OCR) of Ancient Greek text that applied two major post-processing techniques to the output of two commercial OCR engines: ABBYY FineReader 8.0[14] and Anagnostis 4.1[15]. A small number of major series published almost all major editions of Greek texts in the 19th and 20th centuries. Table 1 provides results for samples drawn from the Loeb Classical Library, Oxford Classical Texts, the German Teubner and the French Budé series. The final line describes results for the mid-nineteenth century Bekker edition of Aristotle, which uses a very different Greek font. We chose the Bekker Aristotle as a hard case for much of our work.

The figures in Table 1 reflect a lower accuracy rate than the 99.95% standard in professional data entry contracts and lower still than professional data entry that has circulated for years and where many remaining errors have been fixed.

As mentioned above, however, carefully produced collections of classical Greek texts available to scholars have traditionally encoded only the text as reconstructed by the chosen edition. Manual copying of ancient texts over hundreds of years has added considerable noise and our manuscripts encode a wide range of readings. Editors laboriously reconstruct what they feel the ancient author wrote but all serious editors include what they consider to be the most important variants as notes. Serious students of the text are supposed to check those notes as well as the reconstructed text before basing conclusions on the textual evidence.

We conducted an exploratory survey to estimate the number of words listed as variants in most editions. We randomly selected three pages from ten Greek editions and compared the number of words in the reconstructed text with the number of Greek words greater than one alphabetic character in length. For this set we discovered that on the average page 86% of Greek words were in the reconstructed text while 14% of Greek words were in the textual notes. We had thought that Greek poetry might have offered more textual difficulties than prose but were surprised to find that the figure was identical (86% vs. 14%) for both the five prose and five poetry editions. The ten texts included five editions from Oxford, three from Teubner, one early 19th century edition of Aristotle and a Loeb Classical Text of Plutarch's *Moralia*.

The Loeb Classical Library contains English translations as well as source texts and was traditionally designed as reading aids to open up Greek and Latin to a broader audience. Until the past generation, Loebs generally based their Greek editions on the major scholarly editions of the time

[11] http://www.blakearchive.org/blake/
[12] http://www.brown.edu/Departments/Italian_Studies/dweb/
[13] http://www.csdl.tamu.edu/cervantes/english/index.html
[14] http://www.abbyy.com/
[15] http://www.ideatech-online.com/

Table 1: Baseline OCR Accuracy Rates

OCR Engine	Text	Edition	Char. Level Accuracy
ABBYY 8.0	Plutarch *Life of Solon*	Loeb	99.72%
ABBYY 8.0	Plutarch *Life of Solon*	Teubner	98.00%
ABBYY 8.0	Aristotle *Mechanics*	Bekker (ed.)	99.10%
ABBYY 8.0	Aristotle *Nichomachean Ethics*	OCT	99.20%
ABBYY 8.0	Plato *Phaedo*	Budé	99.84%
		Subtotal	99.17%
Anagnostis 4.1	Plutarch *Life of Solon*	Loeb	99.27%
Anagnostis 4.1	Plutarch *Life of Solon*	Teubner	96.30%
Anagnostis 4.1	Aristotle *Mechanics*	Bekker (ed.)	97.10%
		Subtotal	97.56%
		Total	98.57%

and minimized the number of variants which they cited. We thus measured a sample three pages of ten Loeb editions to estimate the minimum number of variants that a reader should see. The proportion of variants was, not surprisingly, much lower – ten of the thirty pages sampled contained no variants at all. In the Loebs 96% of the Greek words were in the reconstructed text and 4% were in the variants. The five prose texts had slightly fewer variants (97% vs. 3%) than the poetry (96% vs. 4%).

The textual notes appear, however, in a smaller font and generate more errors in OCR. We thus chose as a sample text a work from our hardest case, the Bekker Aristotle.

Table 2: Recall from searching OCR-generated text of Bekker's edition of Aristotle's Mechanics. While the OCR of this text is unusually noisy, it still provides 4.8% more of the overall text and variants than a perfect transcription of the reconstructed text alone.

	total words	errors	accuracy
Text	8,649	266	97%
Variants	810	89	89%
Total	9459	355	96.2
Perf. Transcr. w/o variants	9459	810	91.4%
Gain from OCR		-455	+4.8%

We are measuring our ability to find Greek words on a page and we thus define variants as Greek words in the textual notes and do not count, for this purpose, the page line numbers, manuscript identifiers, etc. Since we do not have a clean text of the textual notes, we identified as an error any word that (1) did not appear in a list of valid Greek words and (2) did not generate a valid analysis from our Greek morphological analyzer. While some incorrect words will pass this test, more often we will not have seen a particular inflected form or have the correct stem in our morphological database. We therefore believe the above figure slightly overestimates the error rate.

More work needs to be done. We need to extend our survey of sample pages and control as well for instances where a textual note repeats the word in the reconstructed text before listing variants. We should also consider weighting variants less than the words which editors chose in the reconstructed text. Nevertheless the results so far surprised us. Even without the specialized error correction described below, a few days of work training individual OCR on the fonts in major series of Greek editions could produce texts which provide search results that are at least comparable, and for some purposes superior, to corpora on which we have lavished millions of dollars and decades of work.

4.2 Automatic Correction of Single Texts

So far we have considered a minimal scenario where no language resources are available and someone knowledgeable in Greek spends a modest amount of time optimizing one or more OCR engines: given access to page images of major Greek editions, an advanced student of Greek with no programming skills could produce an open ended number of searchable Greek editions. In this section, we consider the question of how well we can identify and correct errors given a reasonable digital infrastructure for classical Greek. Our work relies upon two major resources: a list of one million inflected Greek words from manually created editions and a morphological analyzer for classical Greek [8].

In the evaluations reported here, we checked our results against a smaller collection of ground-truth and OCR text aligned at the word level. In the general case, we do not have corrected base texts against which to compare OCR output. We thus identify potential errors by applying two tests on each word generated by OCR: an OCR word is considered an error if it (1) does not appear in the million forms from our existing collection or (2) our morphological analyzer is unable to provide any morphological analyses for the word.

We have assumed in our work that measures of error correction at the word level reflect similar correction rates at the character level. For this to be the case, there should be an even distribution of incorrect characters, as identified by comparison with a ground-truth text, over words identified as errors by our error detection method. In an informal survey of 50 error words in Plutarch's *Life of Solon*, we found that the percentage of uncorrected characters matched very closely the percentage of uncorrected words. Of 50 error words, 10% were left uncorrected by our single and parallel text correction methods, compared to 9.6% of a total 52 incorrect characters. We thus have found it acceptable, in certain cases, to extrapolate character level correction rates from word level ones.

Error detection is, of course, a special case of information retrieval. Precision here measures the number of times the error detection routine incorrectly labels correct forms as incorrect. This figure varies widely depending on how sim-

ilar a new document is to the list of existing Greek words and the stems in the morphological analyzer. In the text of Josephus' *Antiquities of the Jews* we produce analyses for 302,451 of 313,121 words (96.5%), but, if we exclude proper names, the rate rises to 99.5%. Since Josephus stresses different people and places from those in many Greek authors but his language is standard Greek, this figure suggests how well the system works for standard Greek with a very different set of proper names. The poems of Theocritus, by contrast, are written in a specialized dialect with stems and endings that differ from standard Greek. For Theocritus, we produce analyses for 20,186 of 21,671 words (93.1%). In this case, excluding proper names brings the figure up to 97.3%.

The recall rate of error detection is arguably more important. It is easier to process words for which we have no morphological analysis than to find errors that were missed in the rest of the text. To estimate recall, we randomly varied one character in each of 8,441 words from a clean text of Aristotle's *Mechanics*. Of these 8,441 damaged words only 12 words were not identified as errors (i.e., the random changes had transformed these words into other valid, but in context incorrect, Greek words), producing a recall rate of 99.85%. In 1,000,000 words of noisy OCR output where 5% of the words contained errors, a recall rate of 99.85% would identify all but 71 of 50,000 errors. While this figure requires further study, all 12 errors that we missed in our sample were short words of five characters or less and thus represent more common terms that vary much less widely than the rare words (which tend to be longer) and proper names that affect precision.

After stripping error words from the text, we generate a list of potential ground-truth terms for each incorrect form. Potential ground-truth terms were generated on the basis of two sets of statistical data: transition probabilities from one character to another observed in our curated texts and a confusion matrix based on how often the OCR engine confused one character with another. We used an implementation of the Viterbi algorithm to rank the probabilities of each possible correction. On the average we generated 10 possible corrections for each error. Where we were able to generate potential corrections, 66.36% of the top ranked corrections matched the corresponding word in ground-truth. Since we are at present trying to generate text for image-front searching, another figure more accurately describes the impact of single text correction: in 75% of the cases, the correct reading was somewhere on the list of suggested corrections or the original word was already correct but did not show up in our word list or generate a valid morphological analysis.

4.3 Using Two or More Texts to Correct One Another

In a true digital research library we will find multiple editions of the same canonical works of literature. Results reported in [17] suggested to us that it might be feasible to use OCR output from different editions – even when these editions differed from one another – to correct errors in parallel texts automatically. In effect, we would be using the multiple editions in a library in a manner analogous to the double-keyboarding technique used by professional data entry operators. While errors are not randomly distributed (some characters and character configurations are harder for OCR than others), our error rates are low enough and errors are sufficiently random that two texts will probably not have the same error in the same place. As noted above, we only consider error correction in this study and do not consider the problem of collating variants – i.e., places where two editions intentionally differ from each other.

First, we align different editions with each other. Following [17] we automatically aligned editions by finding unique strings in each. Each string that occurs once and only once in both editions defines the start of a new chunk. There are 1,166 words that appear (1) once and only once and (2) appear in the same sequence in two editions of Aristotle's *Mechanics*. Since both editions contain c. 9,000 words, we can automatically align the two texts into 8 word chunks.

Figure 2 shows two chunks from one edition of Plutarch's *Life of Solon* that has been automatically aligned with a second edition. The strings "PROSEDECANTO" ("they received") and "NEWTEROI" ("more recent people") both appear exactly once in both aligned editions and they thus serve as unique milestones with which to segment the two texts into parallel chunks. We recognized "XRSTOIS" as a probable error in the simple text correction phase but we failed to generate any plausible corrections for it.

Second, given an error word in one text, we perform a fuzzy search (i.e., we match strings that differ by one or two characters from our query) on the parallel text to locate the potential correct form. Fuzzy searching works well in this context. Table 4 shows the recall rates of exact and fuzzy searches for 250 random ground-truth terms in a corrected copy of Aristotle's *Mechanics*. We conduct a fuzzy search of the text segment with id "NEWTEROI" in the second text for legal Greek strings that resemble "XRSTOIS" and discover "XRHSTOIS", the correct term.

Third, once error words in a base text have been matched against their potential ground-truth counterparts in the parallel texts, we use rules generated by a freely available decision tree program C4.5[31] to determine which, of a number of possible criteria, are most relevant to the classification of variants as either correct or incorrect. In order to generate this ruleset, we created a control set that indicated, for each variant, whether it was correct or incorrect. Of the criteria we recorded, the number of witnesses in parallel texts was the most important. Other less important criteria were the probability of the variant, as determined by our implementation of the Viterbi algorithm, and whether the variant was duplicated in the single and parallel text correction stages. By applying the decision tree ruleset to the generated variants, we were able to produce variant classifications with an average margin of error of under 10.0%.

Table 3 shows the correction rates associated with single and parallel text correction of four parallel "editions" of Plutarch's *Life of Solon*, along with baseline accuracy rates and the corresponding increases in character level accuracy. Editions here refer to combinations of textual editions (e.g., Loeb, Teubner, OCT, etc.) with two different OCR engines. Although baseline accuracy varies widely among the 4 texts (from 96.3% to 99.72%), the parallel text correction rate is consistently high, and in every case this rate is greater than the single text correction rate by at least 5%, and on average by nearly 16%. Because parallel correction harnesses data from multiple texts and subsumes the variants suggested in the initial single text correction stage, we expected that the parallel correction rate would be higher than the single correction rate. However, we were surprised by how effective this technique actually was.

Table 3: Baseline Accuracy, Single and Parallel Text Correction Rates for Four "Editions" of Plutarch *Life of Solon*

Edition	OCR Engine	Baseline Accuracy	Single Text Correction	Parallel Text Correction
Loeb	ABBYY 8.0	99.72%	99.87% (+54%)	99.93% (+75%)
Loeb	Anagnostis 4.1	99.27%	99.73% (+63%)	99.77% (+68%)
Teubner	ABBYY 8.0	98.00%	98.74% (+37%)	99.29% (+65%)
Teubner	Anagnostis 4.1	96.30%	98.66% (+64%)	98.98% (+72%)
Total		**98.32%**	**99.25% (+55%)**	**99.49% (+70%)**

Table 4: Fuzzy Search of Aristotle's *Mechanics*

Search Method	Recall Rate
Exact Search	98.8%
Fuzzy Search	99.2%

```
<anchor id='PROSEDECANTO'>PROSEDECANTO TOUS ARISTOUS A D
OUN OI</anchor>

<anchor id='NEWTEROI'>NEWTEROI TOUS AQHNAIOUS LEGOUSI TAS
TWN PRAGMATWN DUSXEREIAS ONOMASI
  <app>XRSTOIS
    <rdg margin_err='0.104' class='True'>XRHSTOIS</rdg>
  </app>
KAI FILANQRWPOIS</anchor>
```

Figure 2: Automatically aligned OCR-generated Greek text of Plutarch's *Life of Solon*.

5. FUTURE WORK

First, while we can ignore Greek accents in searches where lexical ambiguity is our focus, we need to determine how accurately we can recognize the accents that appear in modern editions. In addition, we need to examine how well we can capture punctuation as well as basic formatting.

Second, we need to evaluate the impact of differing work flows on image quality and subsequent OCR accuracy. The work reported here was conducted on books scanned on flatbed scanners at a constant 600 DPI. Workflows such as those developed by Kirtas and the Open Content Alliance use digital cameras with a fixed resolution: the same CCD is available no matter how large or small the page may be. Thus, the actual resolution can vary. In the Oxford Classical Text series, for example, print occupies c. 6 x 4 inches (24 sq inches). In the 19th century Bekker edition of Aristotle print occupies 9 x 7 inches (63 square inches). Thus, we were not surprised to find that results for OCA scans were not as good as those generated from our own scans that recorded a constant 600DPI. Other factors need to be analyzed, however: the OCA scanned a different print copy which seemed noisier than that which we scanned.

Third, we need to study errors that occur when we compare two different editions of the same text against each other. In some cases, we will suggest a correction from text-2 that would change, rather than correct, text-1. In this study, we simply measure how well we can, with automated methods, create a transcript of the words in text-1.

Fourth, we need to explore methods with which to use multiple editions to improve automatic structural markup. Thus, if we have two OCRd editions with a different set of textual notes at the bottom of each page, can we identify notes, line/chapter/section/etc. numbers, headers etc. by aligning the shared (though not completely identical) reconstructed texts in each edition? Furthermore, many texts have been carefully entered and are available in XML markup. How well can we use one well marked-up text to correct and mark up many other texts for which we have only uncorrected OCR?

Fifth, we need to see how well we can identify smaller excerpts and quotations of source texts embedded in reference materials and secondary sources. To what extent can we link particular quotations with particular passages (e.g., matching a phrase to its occurrence in a particular section of Plato)? How well can we use a growing library of source texts to augment our ability to correct quotations? How well can we identify the editions on which a particular quotation depends?

Sixth, we need better usage models to prioritize the value of any added-value services that would require capital investment. Where are accurate image-front searches good enough? Where do we want more detailed structural markup? How adequate are automatic methods for extracting structural markup for various tasks?

6. CONCLUSION

For a generation students of classical Greek have depended entirely upon manual keyboarding for digital text. First, we have shown that OCR-generated Greek text can support searches that, by including textual notes as well as the reconstructed text, provide recall that is substantially superior to perfect transcriptions of the reconstructed text alone. Second, we have shown that simple error correction techniques based on an existing word list and a morphological analyzer were able to improve the results generated by two OCR engines on the same text. Third, we have shown that, by comparing the results of two OCR engines on two different editions of the same work, we were able to bring the character level transcriptional accuracy to a point just below the accuracy stipulated in standard data entry contracts (99.93% vs. 99.95%). Large, industrially produced digital libraries with multiple editions of canonical works thus provide resources with which automated systems can scalably correct one text against another.

The potential impact of these methods on the communities who work with classical Greek is immense. Not only will be able to search variants for the first time, but we can begin to imagine comprehensive digital libraries with many editions of the same author automatically collated against one

another. Even with simple error correction measures and text alignment, scholars of the language will be able to see, far more precisely than has been feasible before, where and how editions differ from one another. By scanning secondary sources and reference works for Greek text we will be able to create links between particular passages and the documents which cite them. (In this case, Greek presents an easier problem than Latin or English quotations: running Greek OCR over English texts produces garbage output on English and only captures recognizable words when it processes Greek.) In disciplines where scholarship draws upon and richly cites textual sources, this will open up major new possibilities for mining and visualizing trends within scholarship. We can begin to develop virtual variorum editions, which provide customized summaries of scholarly opinion on any arbitrary passage.

Classical Greek texts are only one moderately challenging example of a broader challenge: creating libraries of collated source editions from large digital libraries. Our own work on early modern and 19th century English suggests that Shakespeare and Dickens may be easier, while classical Arabic, Sanskrit and classical Chinese will presumably be more challenging. Our textual corpora have been immense anthologies, which extracted a single edition from many books and left out all the textual notes and scholarly apparatus. By situating corpus production within a digital library (i.e., a collection of authenticated digital objects with basic cataloging data), exploiting the strengths of large collections (e.g., multiple editions), and judicious use of practical automated methods, we can start to build new corpora on top of our digital libraries that are not only larger but, in many ways, more useful than their manually constructed predecessors. From these digital libraries we can begin to mine core services for those who work with the written heritage of humanity.

7. ACKNOWLEDGMENTS

The NSF ITR Program (NSF IIS-0205466) was primarily responsible for the work reported in this paper. The Digital Library Initiative Phase 2 (NSF IIS-9817484), with particular support from the National Endowment for the Humanities, allowed us to create the foundations on which this work was executed. We particularly wish to thank Sayeed Choudhury, Tim DiLauro and their colleagues in the Gamera project for making this work possible. A grant from the Mellon Foundation to study what we can do for million book libraries help frame the work that we report here.

8. REFERENCES

[1] N. Audenaert, R. Furuta, E. Urbina, J. Deng, C. Monroy, R. Sáenz, and D. Careaga. Integrating diverse research in a digital library focused on a single author. In *ECDL 05: Proceedings of the Ninth European Conference on Research and Advanced Technology for Digital Libraries*, volume 3652 of *Lecture Notes in Computer Science*, pages 151–161. Springer, 2005.

[2] H. S. Baird, V. Govindaraju, and D. P. Lopresti. Document analysis systems for digital libraries: challenges and opportunities. In *Document Analysis Systems VI, 6th International Workshop, DAS 2004*, volume 3163 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2004.

[3] R. Barzilay and N. Elhadad. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 25–32, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[4] A. Belaid, I. Turcan, J. M. Pierrel, Y. Belaid, Y. Hadjamar, and H. Hadjamar. Automatic indexing and reformulation of ancient dictionaries. In *DIAL '04: Proceedings of the First International Workshop on Document Image Analysis for Libraries*, pages 342–54, Washington, DC, USA, 2004. IEEE Computer Society.

[5] J. Carlquist. Medieval manuscripts, hypertext and reading, visions of digital editions. *Literary and Linguistic Computing*, 19(1):105–118, 2004.

[6] H. Cecotti and A. Belayd. Hybrid OCR combination approach complemented by a specialized ICR applied on ancient documents. In *ICDAR '05: Proceedings of the Eighth International Conference on Document Analysis and Recognition*, pages 1045–1049, Washington, DC, USA, 2005. IEEE Computer Society.

[7] G. S. Choudhury, T. DiLauro, R. Ferguson, M. Droetthboom, and I. Fuginaga. Document recognition for a million books. *D-Lib Magazine*, 12(3), 2006.

[8] G. Crane. Generating and parsing classical Greek. *Literary and Linguistic Computing*, 6(4):243–245, 1991.

[9] G. Crane, D. Bamman, and A. Babeu. *ePhilology: When the Books Talk to Their Readers.* Blackwell Companion to Digital Literary Studies, edited by Ray Siemens and susan Scheibman. Basil Blackwell, 2007. Forthcoming.

[10] G. Crane, D. Bamman, L. Cerrato, A. Jones, D. Mimno, A. Packel, D. Sculley, and G. Weaver. Beyond digital incunabula: Modeling the next generation of digital libraries. In *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, volume 4172 of *Lecture Notes in Computer Science*. Springer, 2006.

[11] G. Crane, R. F. Chavez, A. Mahoney, T. L. Milbank, J. A. Rydberg-Cox, D. A. Smith, and C. E. Wulfman. Drudgery and deep thought. *Commun. ACM*, 44(5):34–40, 2001.

[12] G. Crane and A. Jones. The Perseus American Collection 1.0. Technical report, Tufts University-Perseus Project, 2005.

[13] G. Crane and A. Jones. The challenge of Virginia Banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 31–40, New York, NY, USA, 2006. ACM Press.

[14] A. Dekhtyar, I. E. Iacob, J. W. Jaromczyk, K. Kiernan, N. Moore, and D. C. Porter. Support for XML markup of image-based electronic editions. *Int. J. on Digital Libraries*, 6(1):55–69, 2006.

[15] Y. Deng, S. Kumar, and W. Byrne. Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 12(4):1–26, 2006.

[16] D. F. Felluga. Addressed to the NINES: The Victorian Archive and the disappearance of the book. *Victorian Studies*, pages 306–319, Winter 2006.

[17] S. Feng and R. Manmatha. A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 109–118, New York, NY, USA, 2006. ACM Press.

[18] B. Gatos, K. Ntzios, I. Pratikakis, S. Petridis, T. Konidaris, and S. J. Perantonis. An efficient segmentation-free approach to assist Old Greek handwritten manuscript OCR. *Pattern Anal. Appl.*, 8(4):305–320, 2006.

[19] H. Ghorbel, G. Coray, and A. Linden. SAM: System for multi-criteria alignment. In *Proceedings of LREC 2002*, 2002.

[20] T. Kanungo, P. Resnik, S. Mao, D. W. Kim, and Q. Zheng. The Bible and multilingual optical character recognition. *Commun. ACM*, 48(6):124–130, 2005.

[21] Y. Leydier, F. LeBourgeois, and H. Emptoz. Textual indexation of ancient documents. In *DocEng '05: Proceedings of the 2005 ACM symposium on Document engineering*, pages 111–117, New York, NY, USA, 2005. ACM Press.

[22] S. Marinai, E. Marino, F. Cesarini, and G. Soda. A general system for the retrieval of document images from digital libraries. In *DIAL '04: Proceedings of the First International Workshop on Document Image Analysis for Libraries*, pages 150–73, Washington, DC, USA, 2004. IEEE Computer Society.

[23] R. Mihalcea and M. Simard. Parallel texts. *Natural Language Engineering*, 11(3):239–46, 2005.

[24] D. Mimno, A. Jones, and G. Crane. Finding a catalog: generating analytical catalog records from well-structured digital texts. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 271–280, New York, NY, USA, 2005. ACM Press.

[25] C. Monroy, R. Kochumman, R. Furuta, and E. Urbina. Interactive Timeline Viewer (itlv): A tool to visualize variants among documents. In *Visual Interfaces to Digital Libraries [JCDL 2002 Workshop]*, pages 39–49, London, UK, 2002. Springer-Verlag.

[26] C. Monroy, R. Kochumman, R. Furuta, E. Urbina, E. Melgoza, and A. Goenka. Visualization of variants in textual collations to analyze the evolution of literary works in the Cervantes project. In *ECDL 02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 638–53, 2002.

[27] R. Nelken and S. M. Shieber. Towards robust context-sensitive sentence alignment for monolingual corpora. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[28] G. B. Newby and C. Franks. Distributed proofreading. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 361–363, Washington, DC, USA, 2003. IEEE Computer Society.

[29] K. Ntzios, B. Gatos, I. Pratikakis, T. Konidaris, and S. J. Perantonis. An Old Greek handwritten OCR system. In *Proceedings of ICDAR '05*, pages 64–69, Washington, DC, USA, 2005. IEEE Computer Society.

[30] C. B. Owen, J. Ford, F. Makedon, T. Steinberg, and C. Metaxaki-Kossionides. Parallel text alignment. *Int. Journal of Dig. Libraries*, 3(1):100–114, July 2000.

[31] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[32] S. Rawat, K. S. S. Kumar, M. Meshesha, I. D. Sikdar, A. Balasubramanian, and C. V. Jawahar. A semi-automatic adaptive OCR for digital libraries. In *Document Analysis Systems VII, 7th International Workshop,*, volume 3872 of *Lecture Notes in Computer Science*, pages 13–24, 2006.

[33] M. Riva and V. Zafrin. Extending the text: digital editions and the hypertextual paradigm. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 205–207, New York, NY, USA, 2005. ACM Press.

[34] P. Robinson. Where we are with electronic scholarly editions, and where we want to be. *Jahrbuch fr Computerphilogie - online*, 5:123–43, 2004.

[35] D. Schmidt and T. Wyeld. A novel user interface for online literary documents. In *Proceedings of OZCHI '05*, pages 1–4, Narrabundah, Australia, 2005. Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

[36] S. Schreibman, A. Kumar, and J. McDonald. The versioning machine. *Literary and Linguistic Computing*, 18(1):101–7, 2003.

[37] D. A. Smith. Textual variation and version control in the TEI. *Computers and the Humanities*, 33(1-2):103–112, April 1999.

[38] D. A. Smith and G. Crane. Disambiguating geographic names in a historical digital library. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 127–136, London, UK, 2001. Springer-Verlag.

[39] M. Spencer, B. Bordalejo, L. Wang, A. Barbrook, L. Mooney, P. Robinson, T. Warnow, and C. Howe. Analyzing the order of items in manuscripts of the Canterbury Tales. *Computers and the Humanities*, 37(1):97–109, February 2003.

[40] M. Spencer and C. Howe. Collating texts using progressive multiple alignment. *Computers and the Humanities*, 38(3):253–70, August 2004.

[41] S. F. Thomas. Finalizing the multiple-text electronic King Lear for use in the classroom. *Proceedings of ACH/ALLC 2005, Victoria, 15 - 18 Jun 2005*, 2005.

[42] J. Veronis. *Parallel text processing: Alignment and use of translation corpora*. Kluwer Academic Publishers, 2000.

[43] A. B. Zaslavsky, A. Bia, and K. Monostori. Using copy-detection and text comparison algorithms for cross-referencing multiple editions of literary works. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 103–114, London, UK, 2001. Springer-Verlag.