URN: urn:nbn:de:kobv:b4-opus-24327

GREGORY CRANE & ALISON BABEU,
Global Editions and the Dialogue among Civilizations,

in:
Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011, herausgegeben von Ingelore Hafemann, Berlin 2013, S. 11-80.

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae 4

Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Thesaurus Linguae Aegyptiae

4

BERLIN 2013

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN

Perspektiven einer corpusbasierten historischen Linguistik und Philologie

Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12. – 13. Dezember 2011

herausgegeben von Ingelore Hafemann

BERLIN

2013

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Dieser Band wurde durch die gemeinsame Wissenschaftskonferenz im Akademienprogramm mit Mitteln des Bundes (Bundesministerium für Bildung und Forschung) und des Landes Berlin (Senatsverwaltung für Wirtschaft, Technologie und Forschung) gefördert

URN: urn:nbn:de:kobv:b4-opus-24310

INHALTSVERZEICHNIS

VORWORT

Die internationale Tagung „Perspektiven einer corpusbasierten historischen Linguistik und Philologie“ vom 12. – 13. Dezember 2011 am Akademienvorhaben „Altägyptisches Wörterbuch“ der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) war dem Thema des Aufbaus und der Nutzungsperspektiven elektronischer Textcorpora und Wörterbücher in den historischen Sprachen gewidmet. Die Teilnehmer, Vertreter der Ägyptologie, der Hethitologie, Indogermanistik sowie Referenten aus der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen diskutierten vor allem über die Veränderungen, die mit dem Einsatz elektronischer Erfassungs- und Verarbeitungsprozeduren einhergehen. Vertreter der Computerlinguistik vom „Zentrum Sprache“ der BBAW wurden in die Diskussionen einbezogen. Dort beschäftigt man sich seit Jahren mit dem Aufbau großer elektronischer Textcorpora (DWDS), darunter auch solcher, die historische Texte (DTA) für die elektronische Nutzung ermöglichen.

Die größte Herausforderung dieser neuen elektronischen Corpora und Wörterbücher ist es, sowohl den Methoden und damit den wissenschaftlichen Ansprüchen der traditionellen Philologie und Lexikographie unbedingt verpflichtet zu bleiben als auch neue Gebiete wie die Corpus- und Computerlinguistik für die historischen Sprachen zu öffnen. Die Teilnehmer haben gemeinsam und disziplinenübergreifend die Möglichkeiten und Grenzen der Datenerfassung, ihrer Präsentation und den Nutzen neuer Auswertungsprozeduren diskutiert.

Unter dem ersten Thema „Historische Corpusprojekte – synchron und diachron“ wurden elektronische Corpora vorgestellt und ein intensiver Austausch darüber geführt, welche Datenstrukturen die linguistischen Inhalte in adäquater Weise abbilden. Wichtig war die Frage, auf welche Resonanz diese elektronischen Corpora bei den Nutzern gestoßen sind und welche Erwartungen und Anforderungen aus den verschiedenen Fachdisziplinen an die Projekte herangetragen werden. Der Austausch über Nutzungsperspektiven elektronischer Corpora schloss auch die Diskussion über die Erarbeitung projektübergreifend einsetzbarer Standards der Codierung und Strukturierung historischer Textdaten mit ein. Hinsichtlich einer mittel- und langfristigen Nutzbarkeit sowie einer langfristigen Datensicherheit stehen solche Fragen zunehmend im Focus und einige aktuelle Initiativen dazu wurden vorgestellt. Spezielle technische Aspekte

elektronischer Datenerfassung und automatischer Analyse- und Speicherungsverfahren elektronischer Textdaten konnten am letzten Tag als ein Themenschwerpunkt mit den Programmierern diskutiert werden.

Ein zweiter Schwerpunkt waren konkrete Fragstellungen aus der historischen Lexikographie und diachronen Textanalyse. Für das Ägyptische ist der diachrone Ansatz auf Grund der über viertausendjährigen Textüberlieferung von großer Relevanz. Themen wie historischer und/oder textgattungsspezifischer Wortgebrauch, die Erarbeitung diachroner Wortlisten und Aspekte des kontaktindizierten Sprachwandels konnten disziplinübergreifend zwischen den Ägyptologen und den Kollegen der historischen Lexikographie des Mittel- und Frühneuhochdeutschen und des Altfranzösischen behandelt werden.

Mit dem Abendreferenten Gregory Crane, dem Begründer der „Perseus Digital Library“, wurde ein breites Publikum angesprochen. In seinem Vortrag hat er noch einmal die hohe Relevanz und die neuen Möglichkeiten der Einbeziehung zahlreicher Wissenschaftler und einer interessierten Öffentlichkeit in die Projektarbeit demonstriert, die das Internet auf völlig neue Weise eröffnet hat. Die Herausgeberin ist sehr froh, seinen programmatischen Beitrag zu diesem Thema, dessen schriftliche Form er gemeinsam mit Alison Babeu erarbeitet hat, ebenfalls in diesem Band präsentieren zu können.

Wir danken der Berlin-Brandenburgischen Akademie der Wissenschaften für die umfassende Unterstützung unserer Projektarbeit und ganz speziell der Vorbereitung dieser Konferenz sowie der Möglichkeit, die Akten auf dem E-Doc-Server der Akademie veröffentlichen zu können.

Der Hermann und Elise geborene Heckmann Wentzel-Stiftung sei hiermit ausdrücklich für die unbürokratische und großzügige finanzielle Unterstützung dieser erfolgreichen Tagung gedankt.

Das Akademienvorhaben „Altägyptisches Wörterbuch“ konnte sich als aktives Mitglied des Weiteren auf das „Zentrum Grundlagenforschung Alte Welt“ stützen, dem alle altertumswissenschaftlichen Vorhaben der BBAW angehören. Dem Zentrum ist es zu danken, dass der Abendvortrag von Gregory Crane einem breiteren Publikum dargeboten werden konnte.

Allen Autoren dankt die Herausgeberin für ihre anregenden Diskussionen und die qualitätvollen Beiträge in diesem Band.

Auf eine Gesamtbibliographie wurde verzichtet und die Abkürzungen der in den ägyptologischen Beiträgen erwähnten Zeitschriften und Reihen folgen dem Lexikon der Ägyptologie, herausgegeben von Wolfgang Helck und Wolfhart Westendorf, Band VII: Nachträge, Korrekturen, Indices, Wiesbaden 1992, XIV-XIX.

Ganz besonders sei schließlich Frau Angela Böhme für die gewissenhafte redaktionelle Bearbeitung der Manuskripte gedankt sowie Dr. Simon Schweitzer für seine Hilfe beim Erstellen des Layouts.

Berlin, Mai 2013 Ingelore Hafemann

GLOBAL EDITIONS AND THE DIALOGUE AMONG CIVILIZATIONS

GREGORY CRANE & ALISON BABEU

> "If we want to identify one idea which through the whole of history is visible in ever broader effect, if any [idea] proves the often contested, but even more often misunderstood perfection of all mankind, it is the idea of Humanity, the struggle to remove the hostile boundaries which prejudices and biased perspectives have placed between human beings and to treat all of humanity without regard to religion, nationality, or color, as one great, closely related family, as a single whole for the achievement of a single goal, the free development of individual power. This is the final, external goal of sociability at the same time the inborn inclination of human beings to the unconstrained expansion of their destiny." – "On the duties of the historian," Wilhelm von Humboldt (1821)[1]

> "By selecting these two specimens of German scholarship we should indeed adduce the most favourable instances which could be found, but should not exemplify the general character of the German philologer. For, in their activity of mind and body, Hermann and Lachmann came nearer to Englishmen than 99 out of 100 Germans." – John William Donaldson (1856)[2]

This paper is about the reinvention of editing source texts from the human record. Editing may be largely a technical, frequently a tedious, and almost always an underappreciated task, but editing can have profound effects upon the world. We have an opportunity, one could argue an urgent necessity, to establish a dialogue among civilizations. When information flows back and forth across the world in real time, the alternative to dialogue is conflict. The quotations above illustrate two fundamental forces that strain against

---

[1] VON HUMBOLDT, W., 1821: *Über die Aufgabe des Geschichtsschreibers*, Berlin: „Wenn wir eine Idee bezeichnen wollen, die durch die ganze Geschichte hindurch in immer mehr erweiterter Geltung sichtbar ist; wenn irgendeine die vielfach bestrittene, aber noch vielfacher missverstandene Vervollkommnung des ganzen Geschlechtes beweist: so ist es die Idee der Menschheit, das Bestreben, die Grenzen, welche Vorurteile und einseitige Ansichten aller Art feindselig zwischen die Menschen gestellt, aufzuheben; und die gesamte Menschheit ohne Rücksicht auf Religion, Nation und Farbe als einen großen, nahe verbrüderten Stamm, als ein zur Erreichung eines Zweckes, der freien Entwicklung innerer Kraft, bestehendes Ganzes zu behandeln. Es ist dies das letzte, äußere Ziel der Geselligkeit und zugleich die durch seine Natur selbst in ihn gelegte Richtung des Menschen auf unbestimmte Erweiterung seines Daseins."

[2] DONALDSON, J. W., 1856: *Classical scholarship and classical learning considered with especial reference to competitive tests and University teaching*, Cambridge, 157, http://books.google.com/books?id=riACAAAAQAAJ.

one another whenever anyone reflects upon the past. Wilhelm von Humboldt, a Prussian aristocrat and product of the Berlin Enlightenment, sees in the study of history an opportunity to lower the barriers that separate humanity. John Donaldson reduces the study of Greek and Latin to a proxy for the superiority not only of European culture within the world but also of the British upper classes within Europe.

If we follow a path such as Humboldt described, our goal is to increase understanding across humanity. The goal is not to eradicate difference but to promote a dialogue among civilizations – a dialogue that European and North American voices do not impose upon the rest of the world. In 1998, the then Iranian President Mohammed Khatami called for a dialogue among civilizations as an alternative to the "Clash of Civilizations" which thinkers such as a Samuel Huntington had seen as a successor the Cold War.[3] President Khatami's call did not fall upon deaf ears and the United Nations (UN) declared a year of Dialogue among Civilizations. "I see," Secretary General Kofi Annan asserted, "dialogue as a chance for people of different cultures and traditions to get to know each other better, whether they live on opposite sides of the world or on the same street."[4] The official UN English website introduced the topic: "What does a dialogue among civilizations mean? One could argue that in the world there are two groups of civilizations – one that perceives diversity as a threat and the other which sees it as an opportunity and an integral component for growth. The Year of Dialogue Among Civilizations was established to redefine diversity and to improve dialogue between these two groups. Hence, the goal of the Year of Dialogue Among Civilizations is to nurture a dialogue which is both preventive of conflicts – when possible – and inclusive in nature."[5]

It would not be difficult to find similarly contrasting statements in every major language – narrow exclusivity is inherent in our Hobbesian, primate natures, but the cosmopolitan aspirations that we find in Humboldt appear – and will always reappear. Every nation with the opportunity to do so has fallen far short of Humboldt's ideas in the two centuries since they were composed but these failures only emphasize the need to reassert a shared humanity

---

3 HUNTINGTON, S., 1996: *The Clash of Civilizations*, New York.

4 http://www.un.org/dialogue/.

5 http://www.un.org/dialogue/background.html.

and to view in the complexity and diversity of human cultures an opportunity for each of us to learn and to grow. Nor does such a dialogue of civilizations reflect a European or North American attempt to reduce cultures to their own categories. The then president of Iran, Mohammed Khatami, called for such a dialogue and the United Nations responded by declaring a year for the Dialogue among Civilizations. That year was 2001 and the events of 9/11 set in motion a new chain of violence that smothered dialogue but the need for that dialogue remains and is only the greater. When the bombs fall or the door is kicked in before dawn, dialogue may seem a futile, even a laughable instrument. But dialogue, born not only of solemn respect but also of curiosity and delight, provides an essential instrument against violence and for civilization, if that word is to have any meaning.

Greek, Latin, and the Dialogue among Civilizations

As a practical initial goal, we should build a space whereby those who can work with any one of several modern languages can work directly with a range of historical languages.

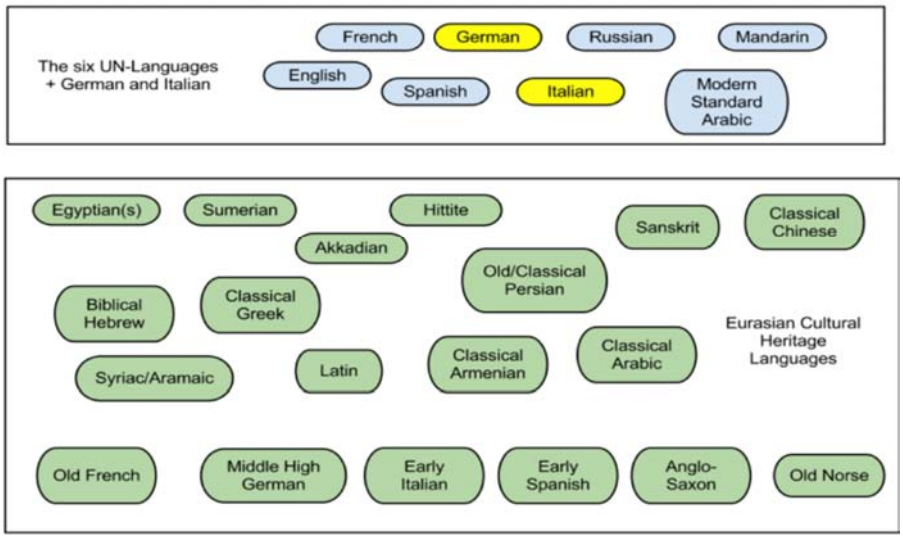


Figure 1: A Euro-centric view of major languages (the six UN languages including German and Italian because of their historical importance in the study of Greek and Latin).

The figure above lists eight modern languages; those in blue boxes are the six official languages of the United Nations. A European contribution to the dialogue among civilizations would probably need to consider including support as well for German and Italian because a great deal of information about the Greco-Roman world is available in these languages. A speaker of Chinese or Russian should, for example, be able to work with information about the Greco-Roman world that is available in French or German. Here the task is to optimize very large systems already emerging to help individuals work across multiple modern languages.[6] Students of historical languages should shrewdly track, exploit and, where appropriate, contribute to new multilingual services such as improving machine translation, information extraction, and cross-language information retrieval.[7] Different communities could extend the coverage to meet their own needs – the European Union might, for example, well want to provide coverage for more European languages, while India might consider support for Hindi, Bengali, Telugu and other major languages.

The lower part of the above figure illustrates a selective and Eurocentric subset of nineteen historical language types. Some of the languages, such as Persian and Egyptian, refer generally to languages that have evolved over thousands of years, from records in cuneiform and hieroglyphics through classical sources in Arabic script. Some of these languages (e.g., Latin, Classical Chinese) remained languages of publication for thousands of years. If we are to support a substantive dialogue among civilizations, we might begin by developing an environment to enable anyone who can understand one of the modern languages above to work directly with materials in any of the other supported modern languages and with any from a subset of historical languages such as those listed below. Thus, a Chinese speaker interested in Alexander the Great should be able to work directly with the lives of Alexander that survive by Plutarch and Quintus Curtius Rufus in Greek and Latin respectively, as well as any

---

[6] For a detailed overview of the use of multilingual technologies to provide cross language access to digital libraries, see DIEKEMA (2012) and for the growing need for such tools in technology enhanced language learning, see ANTONIADIS *et al.* (2009).

[7] A useful overview of the potential of these and other natural language processing technologies for cultural heritage texts and historical languages has been provided by PIOTROWSKI (2012) and SPORLEDER (2010) while a particular focus on the use of these tools for manuscripts has been presented by VERTAN (2010).

supporting scholarship in English, French, German, Italian, Spanish and Russian.

Europe and the Americas can contribute to, but could not, even if they wished, control, a dialogue among civilizations. The people of Europe and the Americas must depend upon their fellows elsewhere to support languages such as classical forms of Chinese, Sanskrit, Persian, Arabic other historical languages. Analysis of the most recent statistics from the Modern Language Association (MLA) indicates that, in the United States at least, the early modern big three of Classical Greek, Latin, and Biblical Hebrew account for more than 95% of all enrollments in historical languages (66,668 of 68,877). Greek and Latin alone accounted for more than three quarters of the total (53,246). Personal experience and conversations with colleagues suggest that the situation in Europe is not much different.

		2006	2009
Latin	Classical and Medieval	31,400	31,369
Greek	Ancient, Koine, Biblical, Old Testament	22,788	21,877
Hebrew	Biblical	14,098	13,422
Aramaic		2,556	562
Sanskrit		607	483
Arabic	Classical	4	285
Chinese	Classical	113	202
Akkadian		129	195
Egyptian[8]		56	110
Slavic	Old Church	133	73
German	Middle High	9	55
Others		223	244
Totals		72,116	68,877
Greek + Latin		54,188	53,246

[8] The MLA statistics do not define what "Egyptian" means in this context. The figure above probably counts those studying the dialect of Arabic currently spoken in Egypt but the figure is included because Egyptian could cover earlier forms of the language (e.g., Coptic, Demotic, Hieroglyphic).

percentage	**75.1%**	**77.3%**
Greek + Latin + Hebrew	**68,286**	**66,668**
percentage	**94.7%**	**96.8%**

Table 1: Enrollments in historical languages based upon figures from the Modern Language Association.[9]

The goal is not to reduce the number of students studying Greek, Latin and Hebrew but to increase the number of those engaged with every historical language – the aggregate 2006 and 2009 enrollments of 72,000 and 68,000 are far too low. Each student of a historical language serves also as a proxy both for broader interest, and access to classes, in a given language. The vanishingly small numbers listed for Classical Sanskrit, Arabic and Chinese reflect the economics of brick and mortar universities and colleges, where each class must draw a minimum number of students to be taught. As distance learning evolves, we will be able to draw upon much larger populations of students and staff courses on more languages – it is easier to find 15 students for a language in a population of 500,000 students (such as represented by the US http://www.cic.net/) than in a liberal arts college of 2000.

In the short run, if we in Europe and the Americas wish to advance a global dialogue among civilizations and to advance a digital infrastructure to support that dialogue, we need to begin by focusing upon Greek and Latin for both diplomatic and practical reasons. First, Greek and Latin are the two major cultural heritage languages to which no region outside of Europe or the Americas can assert a proprietary claim and feel usurped by a Western hegemony. And, second, because there are not enough students of languages other than Greek, Latin, and Hebrew in Europe and the Americas to do the work that is needed – for, as this paper will suggest, our automated systems have now created immense needs and opportunities for intellectual activity of every kind.

Digital Editions

The methods by which we disseminate Greek and Latin are based upon the limitations and possibilities of print technology. They are

[9] http://www.mla.org/2009_enrollmentsurvey; http://www.mla.org/2006_flenrollmentsurvey.

obsolete – indeed, our editions are cultural fossils, retaining archaic forms that now assume and perpetuate a dwindling specialist audience. These forms were, however, originally designed to reach beyond barriers of language, religion and nation. Our task is to re-imagine how to address that ancient goal with the methods available in a digital space.[10]

Non-specialists, interested in the Greco-Roman world, may shake their heads curiously if they happen to pick up the new print editions that specialists still create for one another. The introductions are still, for the most part, exercises in Latin prose composition. The textual notes consist of telegraphic abbreviations that can only partially represent the sources upon which they are based. And the most sophisticated editions still all too often lack an accompanying translation. Editors, of course, have very definite, often distinct, ways of understanding texts in which they have scrutinized every word but the editorial conventions of major editions still assume specialist audiences who can read the Greek or Latin source text on their own. The Greek and Latin editions of the twentieth century were monuments of a closed intellectual culture.

Greek and Latin editions played a different role in early modern culture. When the first editors of printed editions wrote their introductions and notes, even their translations from Greek, in Latin, they were asserting membership in a cosmopolitan European culture that transcended the petty duchies and kingdoms in which they lived. To write in Latin was to advance a transnational republic of letters and to assert a broader identity. The rise of vernaculars – much heralded as a triumph of mass culture – replaced a single language of publication to which no one ethnic group could lay special claim with a handful of culturally dominant dialects. As languages such as French, German, Italian and English emerged as literary media, speakers of these languages could dispense with Latin. Speakers of Croatian and Danish simply had to learn another foreign language – and to accept, in some measure, cultural, if not political domination, of more numerous contemporaries.

The editors of the twenty-first century can now pursue again – and indeed far more effectively – the cosmopolitan goals of their intellectual ancestors. We now have the tools at hand by which to

[10] A series of articles dedicated to this very topic were published in a special issue of *Digital Humanities Quarterly* in 2009, entitled, "Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure," http://www.digitalhumanities.org/dhq/vol/3/1/.

begin developing a new generation of editions, ones designed to serve not merely a European but also a global audience. The grand challenge for editors is not simply to represent a text in a general format but to do so in a format that allows the speaker of Chinese or Arabic to work directly with sources in Greek, Latin, and other European cultural heritage languages.

Adding a translation in a modern language with extensive computational support provides an initial first step: machine translations from English to Mandarin or from French to Arabic may be problematic but they exist and are steadily improving. A great deal more can be done – and the next generation of scholars can congratulate itself on its good fortune in reaching maturity just as our understanding of Greek, Latin, and every cultural heritage language is being reborn. The past is not simply a foreign country but a truly new world, ready to be discovered. Some prototypes exist but we are still in the incunabular stage of invention. No true digital editions exist for any authors.[11] After a generation of experimentation, however, the outlines of new editorial practices are beginning to appear.

The outlines may shift and the subject is in flux – an editor today could put their bets on the wrong services and find their work obsolete even as it is published. We do not know the precise nature of the future – but it hard to believe that the conventions of print will be those of the digital world. Conservative practice is the most promising path to obsolescence and, at best, a sighing sympathy from future readers. The safe bet – producing another edition on the print model – is the safest bet for failure. As students of Greek and Latin, we participate in a conversation that extends centuries and millennia into the past. Our print editions have been mature since Karl Lachmann in the nineteenth century if not before. We have an equal obligation to write, as best we can, for the future and to think in terms of decades and generations to come, rather than the practices that we have inherited.

Digital editions[12] must have the following characteristics:

---

[11] Paolo Monella has also commented on this phenomenon in a recent article, "Why are there no digital scholarly editions of "classical" texts?" http://folk.uib.no/hnooh/filologiadigitale/abstracts/Monella.pdf

[12] The topic of digital editions and how best to design them is a topic of intense discussion within the digital humanities community, and providing support for digital editions is frequently cited as an important task by large humanities cyberinfrastructure research projects, see for example NEDIMAH (Network for

1. Not texts, but multi-texts. Editions must be multi-texts, capable of representing the relationships between any number of versions that the text has assumed.[13] Print conventions present single reconstructions of an original source (a critical edition) or diplomatic representations of particular versions of that text (a diplomatic edition of a manuscript).[14] They represent a finite number of textual differences as manually constructed abbreviated formulas in the notes. These textual notes are often not machine actionable – we cannot dynamically reconstruct from these notes what different versions looked like or see immediately how different versions resembled one another. And different versions should include not only manuscripts and critical editions but also quotations and paraphrases. A digital edition should, as much as possible, trace the entire history of a text.

Within this framework, editors may argue for particular readings or suggest new corrections. They can also create complete networks of suggested readings but these readings constitute – as they have always constituted – a network of annotations that produces one particular version of the text while alluding to many other possible reconstructions. In a truly digital edition, the annotations are immediately separable, whether these constitute the original decisions in an editio princeps or a new anthology of earlier readings.[15]

In some, if not many cases, the earlier states of a text are more important than any new edition, however improved. The works of Galen in Greek, as well as in translations into Arabic and then from Arabic into Latin, served as medical textbooks for more than a thousand years. A new edition of a work by Galen, however much better it captures the original text, should never again inform medical practice. Literary, historical and philosophical works may

---

Digital Methods in the Arts and Humanities) recently announced expert meeting on scholarly editions (http://www.esf.org/index.php?id=8752).

[13] The literature regarding the utility of the digital environment for representing not only different versions of classical or historical texts but also their textual evolution is quite extensive; two recently published books have a number of chapters discussing this topic, see McCarty (2010) and Peursen (2010). For other important work in this area, see also Schmidt & Colomb (2009) and Monella (2008).

[14] For a discussion of "diplomatic editions" in the digital age, see Pierazzo (2011).

[15] For some interesting work in digitally mapping conjectures and variants to textual decisions within editio princeps, see Boschetti (2007) and Cisne *et al.* (2010).

continue to be important in their own right but Machiavelli's text of Livy or the editions behind Gibbon's Decline of the Roman Empire were not those that we use today. If we wish to understand the significance of historical sources in any language, we need editions that help us trace the history of those sources as fully as possible.

2. At least one aligned translation into a modern language. Digital editions must contain at least one major modern language, ideally with a translation that is aligned to the original source text. The modern language translation not only provides basic intellectual access to those who understand that language but also links the original text indirectly to the multi-lingual services available to the modern language (e.g., English) but either not available or not as fully developed for the source language (e.g., Classical Greek). Automatic systems can identify the relationship between most of the words in a Greek or Latin source text and the corresponding words in a modern language translation[16]. Editors can refine these automatic alignments and even optimize their translations to make the alignments more precise. Such optimization can affect the structure and vocabulary. Different translators will, as they always have, pursue different philosophies about how closely the translation should follow the original.

3. Machine actionable annotations as the foundation. Third, digital editions must more fully capture the linguistic interpretations of their editors. Print editions have for centuries added annotations not present in the manuscripts, inscriptions, or other original sources. These include punctuation, capitalization, paragraph breaks, indentation, and indices of people and places. Digital editions should include annotations that represent the editor's understanding[17] and that traditional print markup cannot represent nearly as well if at all.[18] Annotations should include, at a minimum, one or more interpretations of the morphological and syntactic structure of every

---

[16] Work in parallel text alignment is particularly applicable to this task (for a fairly recent overview of the state-of-the-art, see MIHALCEA & SIMARD (2005), and for some interesting work using parallel text alignment and markup projection, see BAMMAN *et al.* (2010)).

[17] O'DONNELL (2009) expands upon this idea of how digital editions can both build upon and improve the traditional practice of print critical editions in representing various textual witness and expert editorial opinions.

[18] For example, the EpiDoc schema (http://epidoc.sourceforge.net/), created for encoding inscriptions can be used to provide for far more sophisticated markup as well as multiple interpretations than is possible with the Leiden conventions, see CAYLESS *et al.* (2009).

word, identifications of every person, place, and similar named entity, metrical analyses, as well as alignments to at least one modern language translation.[19] Since editors traditionally invest a great deal of time pondering the function of every word in a text, the added labor of creating such annotations should be marginal. In practice, annotation should not be a final stage but should constitute a key element of digital editing, with editors using the discipline of linguistic annotation to make sure that they have considered every single word. Digital editions must also contain major alternative annotations.

4. Adequate expository argument to explain the decisions behind the machine-actionable annotations. Digital editions must contain sufficient explanations to justify the choices that their editors make. Even as digital editions exploit machine actionable annotations, expository narrative should justify the substantive decisions that these annotations reflect. There is no reason to have a volume of textual notes separate from the main edition or to create a distinct editio minor without most of the editorial data. The arguments traditionally printed in introductions, commentaries, and accompanying volumes are thus, if anything, more tightly integrated into the edition.

5. Open architectures. Digital editions must have open architectures[20] and can be dynamically constructed from many different elements, each of which has clearly identified provenance. Provenance[21] in turn includes the date at which a conjecture was first published or the number of editors who have endorsed a particular

---

[19] The importance of not only supporting different types of annotations within digital editing and textual scholarship but also the need for shared annotation models to provide interoperability between digital projects is quite vast. For an overview of the nature of digital annotations, see AGOSTI & FERRO (2007), and for recent work combining two of the most prominent annotation models, the Open Annotation Collaboration (http://www.openannotation.org/spec/core/) and the Annotation Ontology, see HUNTER & GERBER (2012).

[20] A number of recent projects have sought to develop open architectures (e.g. shared data models, services, tools and infrastructure) for the creation of digital scholarly editions including Interedition (http://www.interedition.eu), the Virtual Manuscript Room (http://vmr.bham.ac.uk), and TextGrid (http://www.textgrid.de/en/ueber-textgrid.html). For a detailed examination of the importance of developing critical editions as open access texts (including both the marked up text and any code used to generate the edition), see BODARD & GARCÉS (2009). Peter Robinson has also explored the importance of open architectures for the creation of digital editions, see ROBINSON (2010a, 2010b).

[21] For a recent look at designing workflows that support the unique needs of data provenance for philological research, see KÜSTER *et al.* (2011).

variant from one or more manuscripts. Provenance allows readers to reconstruct and to compare particular versions, the contributions that particular sources have made over time and who has endorsed those contributions. The open architecture allows readers to view a new edition in isolation or in conjunction with earlier editions and subsequent reviews. The open architecture also allows readers to link new proposed annotations immediately to the relevant passages in particular texts. The open architecture also allows members of the community to create new translations in a wide range of languages.

6. Dynamic knowledge bases rather than static visualizations. Printed editions – and their PDF imitations – are static visualizations. Digital editions are dynamic entities that evolve over time. Editors may still create comprehensive editions, in which they produce new translations and re-examine many old questions, publishing their own selection of earlier annotations and of their own conjectures. But with digital editions readers can integrate new materials as they appear. Students of the text will add notes on particular passages, studies of particular phenomena, and surveys of the reception of a text.[22] Readers have the freedom to define the texts according to parameters that they choose.

The situation in 2012

According to the criteria listed above, no digital editions yet exist – and no digital editions will soon fully satisfy all six criteria for any textually complex work. But the services, collections and even communities are now in place that can begin to build the textual sources needed to enable broader dialogue and deeper understanding of the human record than has ever before been possible. Computational linguistics, broadly construed, allows us to extract machine actionable text from analogue representations such as images and sound files and then to detect meaningful patterns across vast bodies

[22] There is growing recognition of the need to design digital editions as dynamic sources that lend themselves to both student contributions and collaborative editing between scholars, teachers and students. For example, the Textus Project (http://textusproject.org/), from the Open Knowledge Project, is an "open source platform for working with collections of texts" that "enables students, researchers and teachers to share and collaborate around texts using a simple and intuitive interface." Similarly, the INKE (Implementing New Knowledge Environments) project is examining how best to design tools and interfaces to support an intersection of social media and the creation of "online scholarly editions" (SIEMENS *et al.* 2012). For some other related perspectives, please see BEAULIEU & ALMAS (2012) and GIBBS (2011).

of texts composed in hundreds, if not thousands, of languages.[23] Where computational linguistics focuses largely upon automated processes that can be applied to open ended collections, corpus linguistics develops well-defined, ever more richly annotated corpora to study linguistic phenomena.[24] In the traditional terminology of information retrieval, computational linguists excel at recall (they can detect far more phenomena than human annotators could ever manually examine) while corpus linguists emphasize precision (they focus on annotations of high accuracy in scientifically designed corpora).

As this document is composed in late 2012, many on-going efforts in Europe and the Americas are laying tangible foundations for new digital editions of historical languages such as Greek and Latin. These efforts include at least five different threads, each of which contributes to an emergent fabric of intellectual life; (1) mass digitization, (2) scalable, highly granular collections, (3) customized Optical Character Recognition (OCR), (4) transcription and structural markup, (5) text-reuse detection, (6) machine actionable annotations such as named entity identification and morpho-syntactic analysis, and (7) more decentralized structures for intellectual activity, integrating the contributions of student researchers and citizen scholars.

1. Mass digitization. Gallica[25], Google Books[26], and the Internet Archive[27] are only the most prominent efforts that have made digital images of millions of documents openly accessible to a net public that has, by recent estimates,[28] reached 2.3 billion – one third of humanity. These digital images represent not only books but also manuscripts, papyri, inscriptions and virtually every text-bearing

---

[23] The use of computational linguistics, particularly text mining and data mining, to find patterns across digitized historical corpora, has an ever growing body of literature. One of the best known papers that made us of n-gram detection within Google Books introduced the term "culturomics" to describe this type of work (MICHEL *et al.* 2011). For an overview of the potential of text mining, see UNSWORTH (2011), and for some recent experimental work, see CLEMENT (2012) and ODIJK *et al.* (2012).

[24] For more on the differences as well as the intersection between computational and corpus linguistics, see LÜDELING & ZELDES (2007).

[25] http://gallica.bnf.fr/?lang=EN.

[26] http://books.google.com.

[27] http://www.archive.org.

[28] "The World in 2011: ITC Facts and Figures", International Telecommunications Unions (ITU), Geneva, 2011 (http://www.itu.int/ITU-D/ict/facts/2011/material/ICTFactsFigures2011.pdf).

object. Documents include every major historical language, from Classical Chinese, Sanskrit, Cuneiform languages of the Near East such as Sumerian, Akkadian, Hittite, and Persian, every form of Egyptian from hieroglyphic through Coptic, Classical Arabic, and every language from Europe for which significant written traces survive.[29]

A great deal needs to be done for the coverage of every language. For Greek and Latin, the raw materials are, however, now available. Virtually every major source surviving from antiquity and an immense body of post-classical Latin is available as a scanned image book from some source. Some editions have been poorly scanned or scanned from damaged originals. And even if we have one version of every major source, the multi-text model assumes that we are able to view the textual history of a work as fully as possible – not just one critical edition but every version, including both critical editions and original sources on manuscript, papyrus or stone.

The mass digitization efforts have provided a foundation upon which library professionals can build. Many libraries can now digitize materials from their own holdings and thus many different institutions can add new content and replace problematic scans. The challenge here is to represent the logical contents of, rather than simply the physical form, of the digitized objects. The objects of interest are no longer simply the physical objects that preserve the textual record of the past.

The focus upon books as physical objects rather than upon their contents emerges quickly if one tries to study change over time using digitized books with the default library metadata. This metadata normally records only the date at which a physical book was published rather than including the date as well when the contents of that book were composed. Thus, we find that the vast majority of books catalogued as being in Latin from the Internet Archive list publication dates in the nineteenth century because most of those books were originally printed in that century. Analysis of a subset of 7,000 books that are in fact in Latin and that contain works that can be reasonably assigned single composition dates reveals the actual distribution, with the classical period providing a major, though, interestingly, not dominant, cluster. Interestingly, the nineteenth century remains the major period at which Latin books were

[29] For an overview of the extensive amount of digitized materials available in these various historical languages, see BABEU (2011).

composed – even in the nineteenth century, a great deal of Latin was being produced.[30]

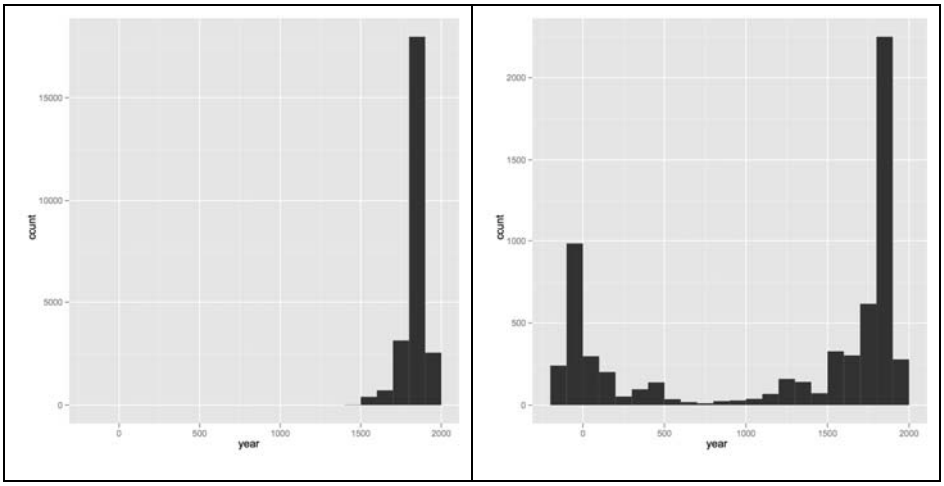


Figure 2: Left, 25,886 books downloaded from the Internet Archive that were catalogued as Latin, charted by publication date; right, analysis of 7,055 Latin books from the Internet Archive, charted by date of composition.

Even the reviewed date figures above provide only preliminary data. The spike of Latin produced in the first century BCE surely reflects not the absolute amount of Latin that survives from that period but the large number of editions for Cicero, Vergil, Horace and other authors from that period. By contrast, the large spike of nineteenth century materials will surely consist much more often of single editions and will thus contain an even larger collection of unique documents than the first century BCE spike. Latin was – and remained through the nineteenth century – a major language of publication within Europe, with many critical scientific, philosophical, and legal as well as literary texts produced in Latin. One could argue that the idea of Europe evolved most purely among those who chose Latin rather than their local language as a means of expression.

In October 2012, the 10,556,524 volumes digitized in the HathiTrust[31] (about ½ the 20 million that Google has already digitized) include 80,069 books identified as being in Latin and

[30] For more on the work that produced this data, see BAMMAN & SMITH (2012).

[31] http://www.hathitrust.org/home.

9,369 as being in Ancient Greek. Such estimates are only rough initial guides – substantial Greek and Latin will appear in books that are not catalogued as being in these languages. Nevertheless, these figures provide a first approximation for a lower bound of Greek and Latin that survive in printed form. An analysis of 9,000 Latin books downloaded from the Internet Archive shows that they include 385 million words. The HathiTrust thus probably contains close to 4 billion words of Greek and Latin. Each of these words is an object of interest that we need to be able to represent and each word can also be the target of an open-ended number of annotations representing an open-ended set of annotation types (e.g., links from a transcribed word to the corresponding section of a page, a link from a name to an encyclopedia entry, a morphological or syntactic analysis of a particular word).[32] By contrast, the *Thesaurus Linguae Graecae* (TLG)[33] contains approximately 100 million words of Ancient and Byzantine Greek. If we focus only upon Classical Greek and Latin (e.g., surviving documents produced through 600 CE – after Justinian and before the Prophet Mohammed), the total is roughly 60 million words of Greek and 40 million words of Latin. The HathiTrust of 2012 already contains about 40 times as many words of Greek and Latin. Of course, many of these books are restricted by copyright law but the counts of Ancient Greek and Latin books in the public domain[34] are 5,587 and 61,659 respectively – about 3 billion words.

catalogued	actual	precision	missed	total	recall
25,886	15,623	60.35%	6,790	22,413	69.71%

Table 2: Book level metadata provides an imperfect tool for locating books in Latin. Out of 1.2 million books downloaded from the Internet Archive, 25,886 were listed as being in Latin. Only 60% of these books were in fact primarily in Latin (many were editions of Greek with Latin introductions) while analysis of the language in 1.2 million book collection revealed 6,790 Latin books that were not catalogued as Latin.[35]

[32] For more on the need to design digital libraries that can deal with analyses at the level of trillions of individual words, see CRANE *et al.* (2012).

[33] http://www.tlg.uci.edu/.

[34] http://www.hathitrust.org/visualizations_languages.

[35] BAMMAN & SMITH (2012).

While it is useful to know that we have 3 billion words of public domain Greek and Latin, such a figure is only a very coarse measurement. Many of these 3 billion words will be represent different versions of the same text – canonical works will have been republished and quoted thousands of times. Each new publication, each excerpt in an anthology, and each quotation represent a decision made at a particular point, with its own context and background. In many instances, readers are not interested in a book but in a logical work such as the Homeric *Iliad* or the *Odes* of Horace. Such logical works often do not correspond to physical books – simple cases such as single volume editions of Dickens' *Oliver Twist* or Shakespeare's *Hamlet* are just one case and even these single volume editions are complex – a text of *Hamlet* will often include not only an introduction but also notes on the bottom of the page below the text.

2. Scalable, highly granular collections. Few researchers actually work with a million, much less ten million, digitized books. Massive collections contain many different potential corpora, each connected to many other corpora but each having its own center of gravity and its own communities. One challenge before us is to create dynamic relationships between smaller, subject-oriented curated collections such as emerged in the first generation of digital scholarship and the massive bodies of data from Gallica, Google and the Internet Archive.

The Perseus Digital Library provides one framework that can be generalized over the 90,000 or so books listed as being in Ancient Greek or Latin and the many citations of Greco-Roman culture scattered throughout millions more books. Perseus serves a number of purposes but its fundamental task is to provide a catalogue of logical documents – it is oriented not around the physical books but around their contents. This approach had evolved already when Perseus began in the 1980s, when CD ROMs had emerged as distribution media and the Internet as it is known today had not yet emerged.[36]

[36] CRANE, G., 2004: Classics and the Computer: An End of the History, in: *Companion to Digital Humanities*, 46-55, Malden Massachusetts. (http://www.digitalhumanities.org/companion).

Figure 3: Visualization of data relevant to chapter 1 of book 1 of Livy's History of Rome in the Perseus Digital Library.

The figure above visualizes results from a query that, in effect, says: "show me everything available about the first chapter of the second book of the *History of Rome* by Livy." The result includes materials of various kinds:

1) Three Latin editions of this particular chapter (with one of these editions the default display for this user). Note that none of the Latin editions contains the whole of Livy's history: two Latin editions come from volumes that contains books 1-10 of Livy, while the third comes from a volume that contains books 1-4. A normal catalogue cannot automatically determine which volumes contain editions of book 2 – or book 32 or 41 – of Livy.

2) Three English Translations of this particular chapter of Livy. Again, each of these translations comes from books that contain varying sections of Livy's work.

3) Three versions of an ancient summary of the first book of Livy's history, two in Latin and one English translation. For most of the works of Livy – and for the works of a number of other authors,

only ancient summaries survive. Summaries are thus an important document type that users need to track.

4) One commentary on this particular chapter. Commentaries are central resources for the study of historical sources. Canonical texts can have not only multiple commentaries composed during centuries of print scholarship but also commentaries preserved in complex formats in earlier manuscripts. These older commentaries are called scholia and a twelfth century CE manuscript can include material produced in Alexandria 1500 years before.[37] Commentaries follow the structure of the work that they explicate, often quoting particular phrases and passages.

5) Livy, like many Greek and Latin authors, has a detailed canonical citation scheme – much as a coordinate system allows people to describe particular regions of the earth, a canonical citation scheme allows scholars to identify particular regions of a text. The existence of these citations allows us to identify passages that mention the first chapter of the second book of Livy's History of Rome. Such references to this chapter of Livy appear (in the figure above) in commentaries on other parts of Livy, in a machine-readable index of Livy, in a reference grammar for Latin, and in an encyclopedia of daily life. Obviously, referenced to Livy will appear in every category of publication.

The structure underlying the figure above is based upon categories that are very old but the visualization depends upon the ability to analyze and manipulate chunks of text dynamically. The volume and page structures of print culture provide a framework out of which the deeper logical structures of logical documents must be extracted and then represented.

Perseus had developed the concept of abstract bibliographic objects (ABO)[38] to represent the distinction between a work, such as Livy's *History of Rome* and the various forms and derivations such as editions, translations, commentaries, and summaries. In the 1990s, the International Federation of Library Associations (IFLA) addressed a similar (though less complex) challenge with its Functional Requirements for Bibliographic Records (FRBR). The FRBR hierarchy provides a framework for organizing dozens--in some cases hundreds

---

[37] For a digital project working with the Scholia of the Homeric Epics, see www.homermultitext.org.

[38] For more on the concept of ABOs, see SMITH *et al.* (2001).

and thousands--of documents associated with canonical works. In the simplest case, FRBR identifies a *work* such as *Hamlet* or *Huckleberry Finn*. Different editions of *Hamlet*, such as those in the Riverside or the Norton Shakespeare, then constitute *expressions* of *Hamlet*. FRBR uses the concept of *manifestations* to distinguish between different physical forms that a particular manifestation can take. The traditional Riverside Shakespeare version of Hamlet, a Braille printing and an audio book constitute three distinct *manifestations* of the same expression. FRBR, in turn, uses the concept of item, to distinguish physical copies of the same manifestation. In traditional libraries, items are central--if the one copy of a book or CD ROM is out on loan or damaged or lost, then no one else can use it. In a digital environment, the *item* still can matter: the FRBR *item* allows us to distinguish the particular copy of a Greek edition of Demosthenes in which John Adams added notes from all other copies of that same edition.

The default FRBR model was originally designed as an entity-relationship model by a study group appointed by IFLA during the period 1991-1997, and was published in 1998.[39] This model was designed to manage print copies of items that frequently had multiple editions. Items become particularly complicated in a digital setting where we can, for example, have multiple scans of the same book, text generated from each scanned page by multiple OCR-engines, then multiple versions of a TEI (Text Encoding Initiative) XML[40] transcription derived from the OCR output (or simply typed in). A more recent effort, FRBRoo,[41] has emerged to provide a metadata standard that mapped the terms of museum documentation and bibliographic description.

For editions of Greek and Latin, Perseus has since 2007 been developing metadata inspired by the FRBR data model.[42] The goal was to develop an extensible bibliography with at least one edition of each Greek and Latin work surviving from antiquity. As an initial focus, the lists of works and editions used by the Lewis and Short Latin-English Lexicon (LS), the Liddell-Scott Jones Greek-English Lexicon (LSJ) and the Oxford Latin Dictionary (OLD) were used to create this initial bibliography. LS dates from the nineteenth century but it covers later Latin, while the more recent OLD focuses upon

---

39 For the full guidelines and model, see IFLA (1998).

40 http://www.tei-c.org.

41 http://www.cidoc-crm.org/frbr_inro.html.

42 For more on this work, see MIMNO *et al.* (2005) and BABEU (2008).

Latin authors through the second century CE. OLD still lists many of the editions that were current when it began work, most of which are now in the public domain. LSJ provides broad coverage for Classical authors, with selective coverage of later sources. Comparison with the TLG Canon – the extensive checklist of editions used by the *Thesaurus Linguae Graecae* – reveals some of the gaps. The largest eleven missing sources are all Christian sources: John Chrysostom (TLG# 2062), Cyril of Alexandria (TLG# 4090), Theodoretus of Cyrrha (TLG# 4089), the series of commentaries on the New Testament known as the Catenae (TLG# 4102), Gregory of Nyssa (TLG# 2017), Didymus the Blind (TLG# 2102), Athanasius (TLG# 2035), Basilius (TLG# 2040), the Ecumenical Councils (TLG# 5000), Epiphanius (TLG# 2021), and Gregory of Nazianzus (TLG# 2022) – a collection that contains more than 13 million words. LSJ documents the great shift of philology away from Christian Greek.

At present, the Perseus FRBR catalogue documents 5,055 Greek and Latin works. Works, at this point, can include not only such well-defined units as Plato's *Republic* or Vergil's *Aeneid*, but also fairly random groups (e.g., the four "epigrams" of Phaedimus that happen to appear in the Byzantine collection known as the *Greek Anthology*) and even phantom works that do not exist in their own right (e.g., the fragmentary quotations and allusions to a lost work or author). The FRBR catalogue represents, however, perhaps the first effort to create a framework by which to track multiple editions of both Greek and Latin authors that may be split among multiple printed volumes or be buried in large, heterogeneous collections such as the *Greek Anthology*.

Out of these 5,055 works, 3,262 have a record describing a particular edition. In 5,935 instances these records include the start and end page of a particular work in a particular printed edition. These records in turn contain 5,195 page level links to image books available in Google Books, the HathiTrust, or the Internet Archive so that users can go directly to a human-readable digitized copy of the books.

	links	works	image books
	0	210	0
	1	962	962
	2	2,037	4,074
	3	53	159
totals		3,262	5,195

Table 3: Image books associated with catalogued works.

This dataset lays the foundation for automatically extracting the sections of books that contain particular works. Ultimately such data will make it possible to feed pages containing particular Greek and Latin works to OCR software and then to use the OCR output to align the new edition with others already online. Rights restrictions still make it impossible often to download the high-resolution versions of the page images needed for best results from OCR software but the underlying data – works, start pages, end pages, and machine actionable links to digital copies – illustrates the necessary architecture for such a system.

Page numbers provide, of course, just a first step towards multitexts. Every word and every character on every surviving object is itself an object of interest. Our metadata must be able to track every word in every surviving version of a work. In addition, students of texts have regularly developed canonical citations schemes as coordinate systems by which to describe very precise chunks of the same text. The surface forms may vary (e.g., Thuc. 4.14 vs. Th. iv, 14) and in cases be ambiguous (e.g., is Th. iv, 14 the fourth *Idyll* of Theocritus or the fourth book of the history of Thucydides) but once properly decoded such citation strings define very precise chunks of text (e.g., chapter 14 of book 4 of Thucydides' *History of the Peloponnesian Wars* or line 14 of the fourth *Idyll* of Theocritus). The contents of these chunks will vary from edition to edition and multitexts need to be able to track those variations, allowing students to recognize, for example, that a particular instance of *fecerit* in one version of a text corresponds to *dixit* in another version. The Canonical Text Services (CTS) protocol[43], which builds upon the FRBR data model, provides a well-defined framework with which to express such relations.

[43] For further explanation of the CTS protocol, see SMITH (2009).

`urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:1.1-1.10`

The uniform resource name (URN) above describes a textual object within the Canonical Text Services Name Space. The basic elements above describe the following features:

greekLit: the work belongs to the category Greek literature.

`tlg0012`: This first field describes a **Text Group**, a category for traditional, convenient groupings of texts such as "authors" for literary works, or corpus collections for epigraphic or papyrological texts (e.g. "Homer," "Aristotle", "inscriptions from a given site"). The string `tlg0012` follows the numerical identifier used by the TLG to designate the Homeric epics.

`tlg001`: Within each TextGroup are **Works**, notional entities, each with a unique identifier within a TextGroup. Each work includes one or more titles (such as titles in different languages). The string `tlg001` follows the numeric identifier used by the TLG to designate the *Iliad*.

`perseus-grc1`: Works, in turn, may appear as **Expressions** which are specific versions of a notional work. Each has a unique identifier within the Work. Within the context of Greek and Latin, expressions are commonly **Editions, Translations, Indices, Commentaries**, author-specific **Lexica** (such as a Lexicon of Homer), and **Summaries**. The string `perseus-grc1` designates a particular Greek edition of the Homeric *Iliad*.

`1.1-1.10`: This designates a range within the canonical citation scheme for the particular work, in this case line 1 of book 1 of the *Iliad* through line 10 of book 1 of the *Iliad*. These URNs can provide the basis for precise and sustainable annotations across documents. Thus, for example, we often need to define the relationship between original source texts and modern language translations. If an English translation of the *Odyssey* begins "Tell me, O Muse, of the man of many devices" and we wish to express the assertion that "of many devices" corresponds to the Greek word *polutropon* in the Greek, we can use the following URNs.

```
urn:cts:greekLit:tlg0012.tlg002.perseus-
eng1:1#of[1]-devices[1]
```

The URN above describes a particular translation of the Odyssey (that of A. T. Murray published in Cambridge, MA, in 1919) and does not assume that this translation contains line numbers. It describes instead a string that begins at the first instance of the word "of" and ends with the first instance of "devices" in book 1 of this translation.

```
urn:cts:greekLit:tlg0012.tlg002:1.1#πολύτροπον[1]
```

The URN above defines the first instance of the Greek word πολύτροπον in line 1 of book 1 of the Odyssey. Like many, if not most, references mined from print sources, this URN does not define a particular edition but instead assumes that the text is sufficiently stable that we can resolve this reference across multiple editions. If the URN above exploits the full expressiveness of the CTS URN syntax, it can easily add a string such as `perseus-eng1` (a critical edition in Perseus) or `hmt-msA` (a particular manuscript of the *Iliad*) to resolve any ambiguities:

```
urn:cts:greekLit:tlg0012.tlg002.hmt-
msA:1.1#πολύτροπον[1]
```

The simplified URN above reflects the reality that most canonical citations are not linked to particular editions. The CTS URN syntax allows for graceful degradation for less precisely specified citations.

The examples above do not address every case in a digital space: we will immediately have multiple OCR-generated transcriptions of different scans of the various physical copies of the same page from a print edition, each of which contains errors. In other cases, different editors will transcribe the same word or abbreviation in a manuscript, papyrus or inscription differently and then occasionally change their minds. We thus need additional specificity, including time-stamps.

Ultimately, accessing the URNs above will yield a digital text, an electronic version of an Edition, Translation, or one of their Exemplars, which will contain one **Online** element. This element contains information about the citation scheme as well as information the server could use to translate the abstract reference into terms needed for local retrieval, such as a filename or database

lookup. Nevertheless, the CTS syntax above provides a precise foundation upon which to build.

In a mature digital space, where we need to align multiple versions of the same work, individual TEI XML transcriptions play a different but important role. In the first generation of digital corpora, researchers depended upon having access to a single, reasonable edition of each work represented in a documented format (ideally, TEI XML). In a multitext space, the transcription becomes a framework around which to cluster and to organize many other editions. Thus, if we can associate a line such as

<l>Arma virumque cano, Troiae qui primus ab oris</l>

with a URN such as `cts:latinLit:phi0690.phi003.perseus-lat1:1.1`,[44] we then can find a very large number of other passages that belong to editions of Vergil's *Aeneid*, or that quote all or part of the above line. Where other versions differ from the base text, we can represent those differences in well-established forms for edit operations (e.g., substitute string X with string Y or insert string Y after string X etc). Once we have one edition of a work encoded with a canonical citation scheme, we can align many others, even when other transcriptions consist of noisy OCR-generated text, and allow users to compare different versions. The TEI XML transcription becomes, in a multitext world, an entry point into a network of different versions. A transcription such as that listed above constitutes both data in its own right and metadata (i.e., data to find related data).

Many Greek and Latin sources exist in digital form but do not support digital scholarship because they are in idiosyncratic formats (such as the page layout description language, developed in the 1970s, in which many Greek and Latin texts are stored), have restrictive front-ends that prevent downloading, and include licensing, enforced with threats of legal action, that prevents the re-use, repurposing and redistribution which are central to digital scholarship. At times, sources are restricted because of all of these reasons.[45]

---

[44] PHI stands for Packard Humanities Institute which published a collection of Classical Latin Texts and assigned identification numbers to authors and works. Here phi0690 designates Vergil and phi003 the *Aeneid*: http://latin.packhum.org/.

[45] CAYLESS (2010) has made a strong case for the role of re-use in long-term digital preservation, whereas a panel at the Digital Humanities in 2009 explored the

Approximately 20 million words of Greek and Latin – roughly 20% of the classical corpus – are either already available, or have been entered and are being formatted, in TEI XML with Creative Commons open licenses.[46]

Greek and Latin editions		
versions	TEI XML transcriptions	total
1	970	970
2	22	44
3	3	9
Subtotal	995	1,023
English Translations		
1	539	539
2	96	192
3	2	6
Subtotal	637	737
Total	1,632	1,760

Table 4: TEI XML transcriptions in the Perseus Digital Library representing original language editions and English translations of Greek and Latin sources.

The Perseus Digital Library currently has 995 distinct Greek and Latin sources in TEI XML, along with English translations for 637 of these works. The collections in Perseus provide breadth but the handful of instances where more than one edition and translation are available have provided an opportunity to develop and demonstrate initial methods by which to manage multiple versions of the same work.

We can represent trillions of relationships between billions of words digitally but we cannot transcribe, much less annotate, 4

difficulties of reusing even open-source objects within digital classics (BODARD 2009).

[46] The major sources for on-line TEI XML transcriptions of Greek and Latin are the Perseus Digital Library (http://www.perseus.tufts.edu/hopper/opensource/download) and http://www.papyri.info/. A Mellon-funded Project centered at Harvard has entered, and is now formatting, several million words of Greek scientific and medical texts.

billion words of Greek and Latin. We must depend upon automated methods if we are to organize even such a modest collection as the surviving body of Greek and Latin (which account for less than 1.5% of the digitized books in the HathiTrust). Many of these 4 billion words will be different versions of the same work – but book level metadata alone would not allow us to determine how many versions of book 4 of Vergil's *Aeneid* or of Sophocles' *Oedipus the King* are within this massive collection: one volume may contain three plays of Sophocles, one play, or all seven remaining plays, while many edited documents are quite short and appear as sections in larger publications. And each version of a document is a historical event in its own right – the school anthology may, for example, draw upon a standard edition but the fact that it drew upon a particular edition and the selections that it drew shed light upon intellectual and educational practices of the time. There is no good way to determine how many unique words of Latin from how many works are within this vast space without analyzing the texts themselves.

If we are to manage the vast body of materials already available to us we need a two-fold transformation of scholarship. We obviously need to draw upon automated methods of every kind relevant to the analysis of textual data in many multiple languages. But automated methods are not enough – there is just too much work to be done and too many instances where human input is necessary. Even if all library professionals and advanced researchers shifted their focus away from book-level metadata creation and specialist publications and towards the myriad tasks by which to make these billions of words ever more intellectually accessible to an ever widening set of humanity, the labor available would still not be enough. Professional students of Greek and Latin must welcome student researchers and citizen scholars as collaborators – in the United States, the 3200 or so members of the American Philological Association (APA)[47] must, in other words, turn not only to the 55,000 students of Greek and Latin in postsecondary education but also to the almost 150,000 secondary

[47] This figure is based upon the statement at http://apaclassics.org/index.php/about the APA/director report/executive director report for 2011/ that 800 represents 27% of the individual members of the American Philological Association. This figure, which includes some who are not professional classicists and others who are not from the United States, serves as a rough estimate for the number of professional Classicists in the United States.

school students studying Latin.[48] Such a shift in the relationships between teacher and student and between learning and research would presumably have an effect upon the students who enroll in Greek and Latin and, inevitably, the number of jobs for those teaching them.

The explosion of digital access to Greek and Latin has transformed the relationship between those languages and society. At the least, a global public could view a range of Greek and Latin sources which were previously only available in research libraries. This physical access challenges students of Greek and Latin to provide the intellectual access needed to understand these sources. That challenge in turn provides the most inward looking specialist with a material reason to look outwards and to engage a wider audience. We cannot pursue our research fully without a new collaborative, laboratory culture. Every aspect of digital editing depends upon not only new automated methods but also new, more broadly based forms of collaboration.

3. OCR for historical languages: Human beings can read images of writing – indeed, high resolution, multispectral and 3D scans of text-bearing objects can make some surfaces more readable than the original objects were to the naked eye.[49] But we cannot transcribe billions of words of Greek and Latin. OCR works well for modern printed Latin texts if the OCR system knows that it is analyzing Latin and if it has access to a Latin dictionary/word list so that it does not try to turn Latin into some other language (e.g., Latin t-u-m, "then," can become English t-u-r-n if the OCR system expects English). But commercial OCR performs much less well for earlier printed books in Latin and indeed in any language. Substantial work remains to be done if we are to extract high quality text from these earlier printed sources.[50]

---

[48] The figure of 150,000 is a rough approximation based upon the 148,000 students who registered for the 2012 National Latin Exam: http://www.nle.org/pdf/ExamResults2012.pdf.

[49] For example, using such technologies has provided unprecedented access to the Archimedes Palimpsest (http://www.archimedespalimpsest.org), see SALERNO (2007).

[50] While still a relatively specialized area, the development of OCR tools (both the modification of commercial tools and the adaptation of open source systems) for historical languages has grown dramatically in the last five years. See for example, the results of the recently concluded Improving Access to Text (IMPACT) project (http://www.impact-project.eu) as well as the newly funded Early Modern OCR Project (http://emop.tamu.edu/). For a review of the state-of-the-art in this area, see PIOTROWSKI (2012).

At the same time, while OCR may need to be optimized for recently published Latin, Classicists have never had access to reasonable OCR-generated text for Ancient Greek. For the forty years since the TLG was founded in 1972, they have had to depend upon manual keyboarding – a labor intensive, inherently expensive process. It has not been possible to image working with thousands of books printed in Ancient Greek. That situation changed when Gordon Stewart published the first paper documenting the effective use of OCR for Classical Greek.[51] He demonstrated that in 2007 a modern Greek OCR system (Anagnostis), trained to ignore the accents in Classical Greek, could generate transcriptions of the alphabetic characters in 19th and twentieth century Greek editions. Because this OCR method also included textual variants and because these variants account for between 8 and 15% of the words on a given page, OCR generated text for editions immediately provides better recall than error-free transcriptions that only include the reconstructed text.

In the subsequent five years, Federico Boschetti and Bruce Robertson carried this work further.[52] Commercial OCR systems had serious limitations: they could not be trained to recognize Classical Greek directly or they could not run on large bodies of text or their licensing systems were not designed to support multi-processor systems. Boschetti and Robertson undertook to train open source OCR systems to recognize Classical Greek and to develop the error checking methods needed to correct the output.

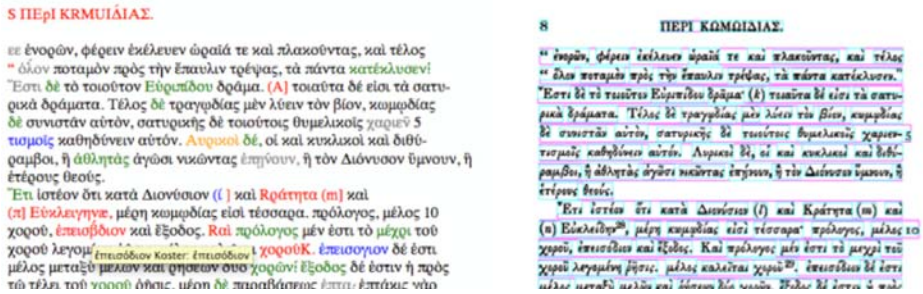


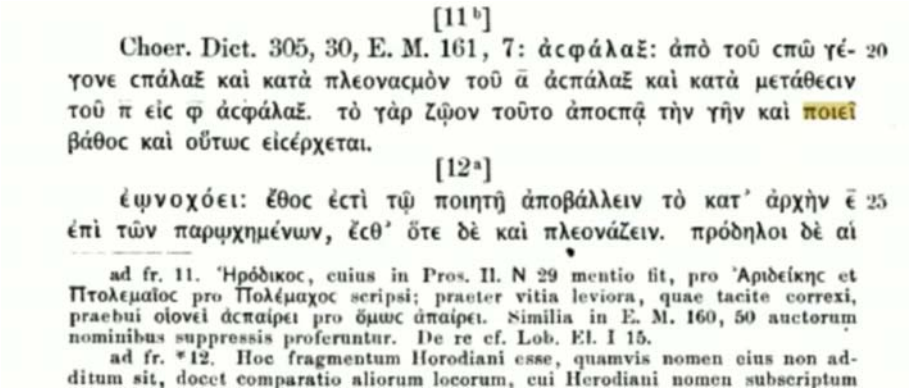
Figure 4: Error identification in Greek OCR developed by Federico Boschetti. Color indicates classes of error. The HOCR format above includes (1) suggestions for corrections based upon standard spell-checking strategies; (2) suggestions based upon words as they appear in another edition on-line (near ground truth).

[51] STEWART *et al.* (2007).

[52] For more on this work, see BOSCHETTI *et al.* (2009) and ALMAS *et al.* (2011).

This multi-text approach to digital editions creates editions that are, in effect, self-correcting as they include OCR-generated text from multiple print editions, even where these individual transcriptions contain substantial error rates. Suppose OCR for two different editions of a text (perhaps one a Teubner and one a Loeb) generates an error in every 10th word. If the errors are randomly distributed, then the probability that one or the other OCR-generated text contains a valid reading rises to 99%. If we add a third edition under the same conditions, the probability that we will have at least one correct transcription rises to 99.9% and so on. Of course, different editions will have different forms up to 5 or 10% of the time but as more editions become available, the probability that the same reading will be correct somewhere will rise. Errors will remain but the nature of the discussion has now shifted from never having variants to doing a better job of capturing a growing body of variants.

Once we align OCR-generated text not only with the page images from which it was derived but also with other editions of the same text, we can create image-front searching long familiar to academics from JSTOR[53]. We search for Greek and Latin and fault tolerant searching locates probable hits and displays the results either as text or as clips from the image of the printed page.


[11 b]

Choer. Dict. 305, 30, E. M. 161, 7: ἀcφάλαξ: ἀπὸ τοῦ cπῶ γέγονε cπάλαξ καὶ κατὰ πλεοναcμὸν τοῦ ᾱ ἀcπάλαξ καὶ κατὰ μετάθεcιν τοῦ π εἰc φ ἀcφάλαξ. τὸ γὰρ ζῷον τοῦτο ἀποcπᾷ τὴν γῆν καὶ ποιεῖ βάθος καὶ οὕτωc εἰcέρχεται.

[12 a]

ἐψνοχόει: ἔθοc ἐcτὶ τῷ ποιητῇ ἀποβάλλειν τὸ κατ' ἀρχὴν ἓ ἐπὶ τῶν παρῳχημένων, ἔcθ' ὅτε δὲ καὶ πλεονάζειν. πρόδηλοι δὲ αἱ

ad fr. 11. Ἡρόδικος, cuius in Pros. Il. N 29 mentio fit, pro Ἀριδείκης et Πτολεμαῖος pro Πολέμαχος scripsi; praeter vitia leviora, quae tacite correxi, praebui οἱονεὶ ἀcπαίρει pro ὅμωc ἀπαίρει. Similia in E. M. 160, 50 auctorum nominibus suppressis proferuntur. De re cf. Lob. El. I 15.

ad fr. *12. Hoc fragmentum Herodiani esse, quamvis nomen eius non additum sit, docet comparatio aliorum locorum, cui Herodiani nomen subscriptum

Figure 5: Image-front, morphologically-aware searching of OCR-generated Greek text. (Bruce Robertson, demo of the Squeegee search prototype, developed as part of a Digging into Data Phase 1 Project).[54]

[53] http://jstor.org.

[54] http://heml.mta.ca/RobertsonGreekOCR/.

The major challenge at this point is to develop the workflow that will feed scanned editions of Greek and Latin to the appropriate OCR software and then allow members of the community to correct the output as they see necessary. This becomes now a question of software development and of the diplomacy needed to make high-resolution scans of public domain Greek and Latin editions available.

4. Transcription and structural markup: More than 25 years ago, the TEI began to develop shared conventions for representing texts in digital form. A major goal of the TEI was to enable semantic markup – rather than labeling a string in italics and then letting the reader determine if the string were in italics because it was the title of a book, because the author wanted to emphasize the text, because the text was in a foreign language, or because of some other reason, the TEI offered conventions to express these deeper purposes. Formatting software could then convert titles and German quotations into italics for printing, while the text preserved these distinctions in a machine-actionable form. The TEI published its fifth edition of Guidelines (TEI P5) in 2007. Off-the-shelf commercial XML editors such as Oxygen[55] exist that support editing TEI XML. Workshops regularly introduce neophytes to the basic (and not, in the end, so terribly challenging) basics of TEI XML.[56] An individual or small working group can now create individual TEI XML transcriptions of texts in Greek, Latin, and many other languages.

The problem now is one of scale. In fall 2012, roughly 35,000 individual users each month work with more than 17 million words of Greek and Latin texts in Perseus. How can we enable any of these users to correct residual data entry errors in, or add additional TEI XML markup, within this corpus as a whole? What happens as the amount of OCR-generated Greek and Latin text ready for editing increases to billions of words and the audience of potential contributors expands beyond the largely English-language users of Perseus?

There are two approaches to this problem. In the simplest case, texts are uploaded to Wikisource[57] and the Wiki community makes corrections as they choose.[58] The Wiki formatting language is not as

---

55 http://oxygenxml.com.

56 See for example the resources offered by the Women Writers Project at Brown University (http://www.wwp.brown.edu/outreach/resources.html).

57 http://wikisource.org/.

58 The potential of collaborative transcription and the creation of TEI-XML documents has been investigated by the Transcribe Bentham project, see CAUSER *et al.* (2012). There are also a number of tools other than WikiSource that have

expressive as TEI XML but it can capture the basic page layout and some fundamental semantic concepts. Texts corrected in a Wikisource space provide an excellent starting point for more elaborate TEI markup. And, with a little work, most corrections to a Wikisource version of a text could, in most cases, be automatically integrated into a parallel TEI XML transcription. In this model, the Wiki infrastructure provides the framework for basic text correction.

Another approach focuses upon the challenge of precisely representing many different changes to a collection, some involving isolated changes to particular documents, others covering thousands of passages. In the Wikipedia model, corrections converge on a single canonical transcription of a master print source. Scholarly editing will, however, produce many different versions of the same text and the editorial workflows quickly diverge as different groups potentially create their own version of the same text. To address this case, papyrologists, funded by the Mellon Foundation, developed a more complex workflow, the *Son of Suda Online*. (SoSOL).[59]

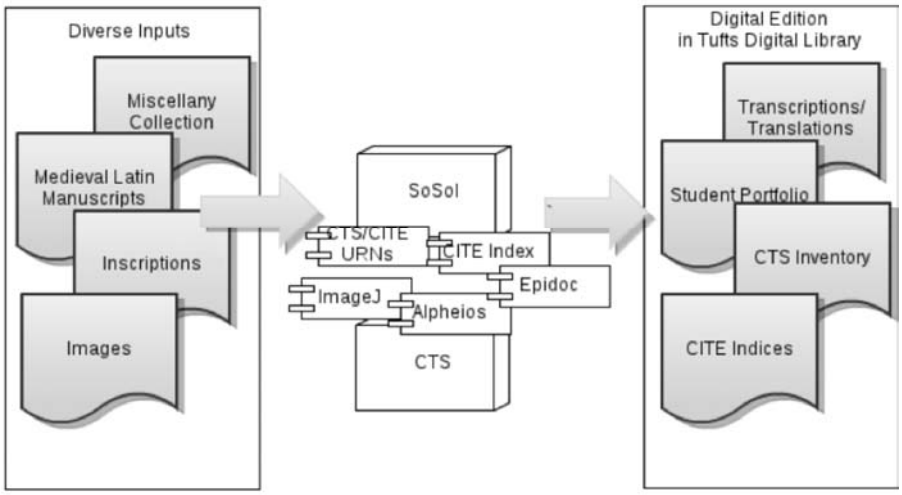


Figure 6: SoSOL as it is being adapted to work with materials in the Perseus Digital Library.

been created to aid in the creation of collaborative manuscript and or text transcriptions, including Scripto (http://scripto.org) and T-Pen (http://t-pen.org/TPEN/).

[59] http://idp.atlantides.org/trac/idp/wiki/. For more details, see SOSIN (2010).

Much work remains, however, to make SoSOL scale up beyond dozens of papyrologists to thousands of contributors working with Greek and Latin in general. Nevertheless, SoSOL can track a large number of very precise editorial events and it constitutes a fundamental step in the direction of scalability.

5. Automatic cataloguing, including language and text reuse detection: Once we have a collection of OCR-generated texts, we can begin to look for instances where one text re-uses another. Book level metadata provides, of course, only a very coarse guide. Books that are primarily in Latin or Ancient Greek can contain distinct documents from different periods (e.g., the Byzantine collection of Greek poetry known as the *Greek Anthology*) and genres (e.g., inscriptions from the same site and covering many genres are customarily published together). Documents also quote each other: Porphyry quotes Plato but Plato also quotes Homer. The self-standing edition and the text that draws upon an earlier text represent two ends of a continuum that we need to track if we are to understand the history of a text.

The Proteus Project,[60] developed with support from the National Science Foundation (NSF)[61] by researchers at the University of Massachusetts, had addressed the problem of identifying duplicate versions of the same work in collections that are large (greater than 1 million books) and that can, in depending upon OCR-generated text contain numerous errors.

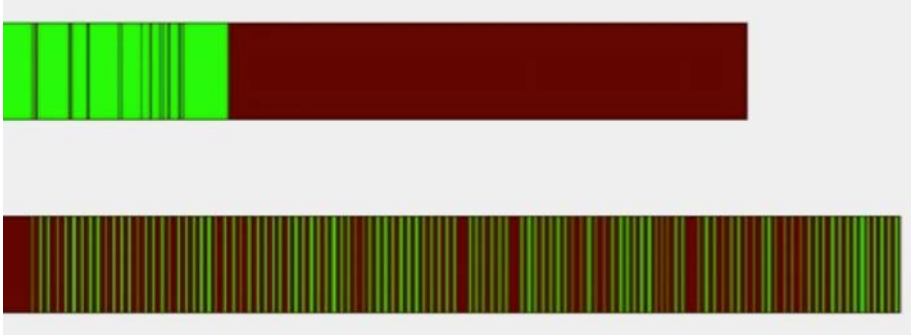


Figure 7: Text alignment is also used for finding groups of texts whose structure corresponds in other ways, such as works published in different languages, or texts and their commentaries. Here, for instance, we see an automatically

[60] http://books.cs.umass.edu/beta-sprint/.

[61] http://www.nsf.gov.

generated alignment between the Latin text of Vergil's Aeneid *and a commentary. The first bar depicts the first eight books of the* Aeneid*. The green in this first bar indicates the aligned portions, from which we can tell that the commentary only deals with the first three books of the* Aeneid*. The second bar depicts the commentary. Its green portions are brief passages from the text of the* Aeneid*, and the intervening red bars are the commentary, which does not align.*[62]

At the other extreme, one text quotes or paraphrases small sections of another (e.g., Plato quoting Homer). In this case, at least three issues complicate the process. First, it is not always clear when one text is directly citing another – we generally need to know the composition dates of various documents so that we can automatically determine which document cites the other. Second, text reuse can include short phrases (e.g., "to be or not to be") and it may not be clear whether the phrase represents an intentional allusion to a particular text (e.g., to *Hamlet*) or has simply become an idiom with no widely recognized single origin. Third, one text may paraphrase, rather than directly quote, another, thus making it hard to detect the textual reuse by searching for repeated strings.

The UMASS Proteus system has also explored methods to detect and to visualize text reuse in large collections. The Proteus visualization of documents that quote *Hamlet* maps one text onto a restricted number of quoting documents. This visualization allows readers to compare a single text with a finite number of documents that quote it.[63]

[62] Text drawn from: http://books.cs.umass.edu/betasprint/Demonstration/Entries/2011/8/3_Aligning_the_Aeneid_and_commentary.html.

[63] For further discussion of this work, see SMITH *et al.* (2011).

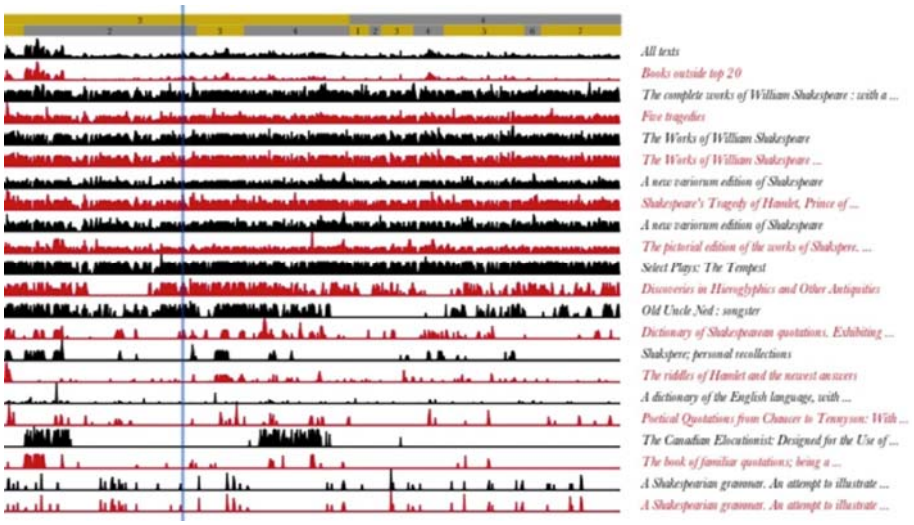


Figure 8: At the top are listed the acts and scenes of the play. Below are histograms showing the amount of textual overlap between each line and various other books. Five Tragedies, for instance, contains the complete text of Hamlet and thus overlaps completely. But we can also see other genres such as a Dictionary of Shakespeare, which uses quotes to illustrate word definitions, or The Canadian Elocutionist, which excerpts speeches for practice by aspiring public speakers, or The riddles of Hamlet *and the newest answers, which is a work of literary criticism*[64]*. The text reuse patterns are represented using the Highbrow visualization tool.*[65]

The eAqua[66] and subsequent eTraces[67] projects, located at the University of Leipzig and funded by the German Federal Ministry of Education[68] also explored the problem of detecting text reuse within a corpus. The first visualization illustrates how subsequent students of Plato used the author's *Timaeus*. The visualization illustrates how this work grew dramatically in importance as Neo-Platonism replaced Middle Platonism. It also shows which passages the Middle and Neo-Platonists most often cited (thus showing a shift in interest within the work). In addition, the visualizations show which authors most often cited this work.

---

64 Text drawn from: http://books.cs.umass.edu/beta-sprint/Demonstration/Entries/2011/8/2_Quotation_detection%3A_Hamlet.html.

65 http://osc.hul.harvard.edu/highbrow/.

66 http://www.eaqua.net/index.php.

67 http://etraces.e-humanities.net/.

68 http://www.bmbf.de/.

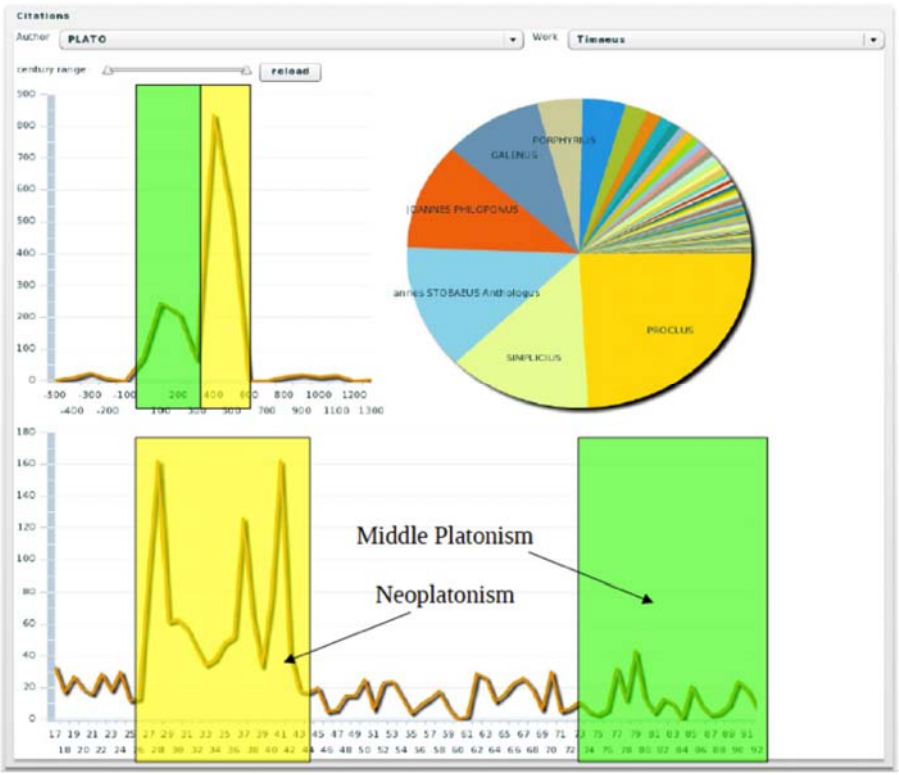


Figure 9: The Leipzig-based eAqua project explored the relationship between different texts. Here we see the frequency with which later authors cite portions of Plato's Timaeus*. In the above left, the green and yellow boxes distinguish quotations by Middle and Neo-platonists, demonstrating the surge of interest in the* Timaeus *among the neo-Platonists. The pie chart on the upper right hand illustrates which authors most frequently cite the* Timaeus*. The graph below shows which sections of the* Timaeus *are most frequently cited. The yellow and green boxes illustrate the sections of greatest interest to the Middle and Neo-Platonists.*

The eAqua and eTraces projects[69] also developed "heat maps" to track which sections of an author's work are most frequently quoted in subsequent Greek literature and thus to see as well which authors are more frequently quoted than others. The heat maps below illustrate the quotation frequency of passages in the surviving works of Xenophon, Plato, Aristotle, and Plutarch. Not surprisingly, Plato

[69] For some related publications regarding the work of both projects, see for eAqua (BÜCHLER *et al.* 2010) and for eTraces (BÜCHLER *et al.* 2012).

and Aristotle are much more heavily quoted than either Xenophon or Plutarch. The heat map for Plato shows a particularly striking pattern of black (i.e., rarely if ever quoted) passages among much more heavily cited passages. The heat map captures a one-to-many relationship (e.g., how often one text is cited in a collection of open ended size). The heat maps above can provide summary views of all subsequent citations while the UMASS visualization shows relationships with specific texts.

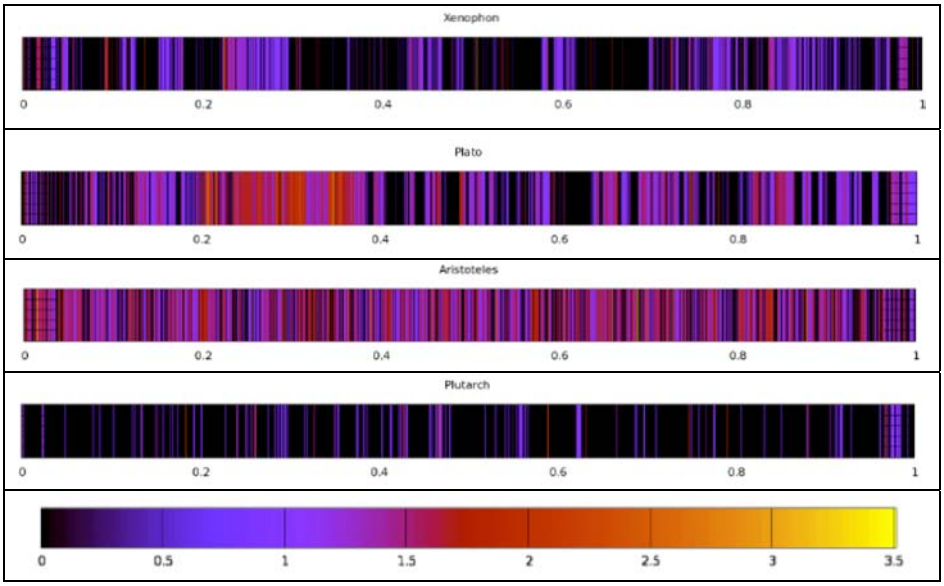


Figure 10: The heat maps above reflect the frequency with which sections of an author's surviving work have been quoted. Sections of the work that have not been quoted appear as black. The more frequently the section has been quoted, the brighter the color, with yellow indicating passages quoted more than three times by other authors.

Translation is a special case of text reuse: a translator takes words in one language and represents them, more or less closely, in another. Automated methods can detect in most cases which words in a source language correspond to their equivalents in a translation – assuming there are enough parallel texts so that the system can learn which words in one language correspond with words in another.[70]

[70] The use of parallel texts for translation alignment has also proven useful as one step in finding translations within massive digitized collections of books (YALNIZ

The Alpheios project[71] has provided tools whereby human editors can refine the results of this machine alignment of source text and translation. The figure below shows a human edited alignment of Greek and English words in the opening of the Homeric *Odyssey*. The textual data is here visualized as a traditional interlinear translation (such as were developed when Greek and Latin were staples of education and many students had to struggle through a few canonical texts).

Text reuse becomes an object of scholarly concern in particular when the quoted source does not itself survive and the quotation is not necessarily verbatim. Thus in the following passage, a speaker in Athenaeus' *Banquet of the Wise Men* quotes an earlier source.

> Ἴστρος δ᾽ ἐν τοῖς Ἀττικοῖς οὐδ᾽ ἐξάγεσθαί φησι τῆς Ἀττικῆς τὰς ἀπ᾽ αὐτῶν γινομένας ἰσχάδας, ἵνα μόνοι ἀπολαύοιεν οἱ κατοικοῦντες: καὶ ἐπεὶ πολλοὶ ἐνεφανίζοντο διακλέπτοντες, οἱ τούτους μηνύοντες τοῖς δικασταῖς ἐκλήθησαν τότε πρῶτον συκοφάνται.
>
> "And Istrus, in his Attics, says that it was forbidden to export out of Attica the figs which grew in that country, in order that the inhabitants might have the exclusive enjoyment of them. And as many people were detected in sending them away surreptitiously, those who laid informations against them before the judges were then first called sycophants." (tr. C. D. Yonge)

Scholars have tried to reconstruct from such fragmentary pieces lost works of Greek and Latin – most of the works of which we know only survive insofar as they are quoted, paraphrased or mentioned.[72] In the passage above, we need to decide what words we believe come from Istros and what words were produced by Athenaeus. We need to mark "says" as the so-called *verbum dicendi* (the word of speaking) so that we can compare it with other similar words (e.g., "asserts", "claims", "reports") and so that we can detect the ways in which one author describes their use of sources. Ultimately we move from automated services that detect textual reuse to close scholarly analysis.

While we may wish to use textual alignment to identify multiple editions and quotations of a work, methods also exist by which to identify translations and then to align many of the words in the

---

& MANMATHA 2012), as well as for markup projection between text in Greek and Latin and modern language translations (BAMMAN *et al.* 2010).

71 http://alpheios.net.

72 For some preliminary work on the encoding of fragmentary works within digital editions and libraries, please see BERTI *et al.* (2009).

original text to their equivalents in the translation. Such parallel texts are fundamental to many, if not most, multilingual services now in use – statistical methods are used to determine automatically which words co-occur. Such parallel texts also enable new lexicographic and semantic tools that grow more and more useful as collections grow larger and purely manual techniques become less feasible.[73]

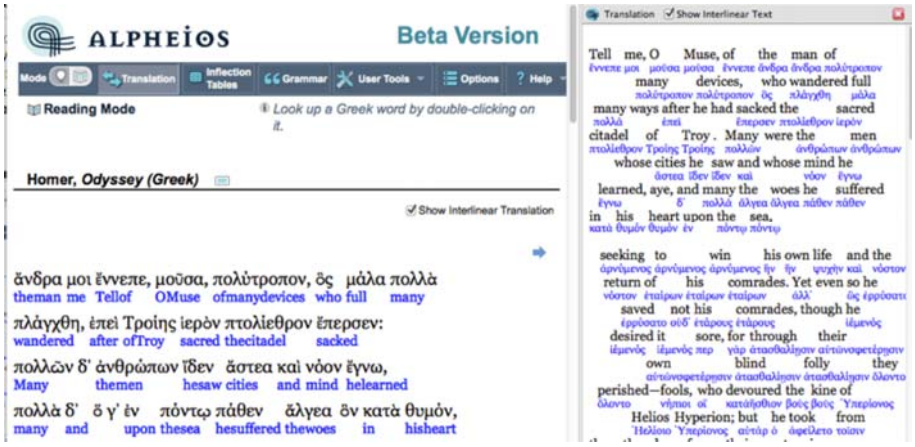


Figure 11: Visualization of Greek and English words aligned to one another in the Alpheios parallel text browser.

The figure above visualizes a Greek text of the Odyssey aligned to an English translation and the corresponding English translation as it is aligned to the Greek. The alignments above were first generated automatically and were then edited.

With the source text/translation alignment, however, we also enter into the world of reading support. The more precisely a source text and its corresponding translation correspond, the more support readers have in picking apart the granular form of a source text in a language that they may have never studied. With aligned source texts and translations we begin to provide a fundamental instrument for global editions that must serve many different linguistic and cultural audiences. The links from Greek to English above connect the Homer text to vast and growing resources being developed to make English (or any other major language) available to a global net audience.

---

[73] For example, see work on the Dynamic Lexicon (BAMMAN & CRANE 2008).

6. Annotation of named entities and morpho-syntactic features: Digital editions should also include machine actionable annotations on various features relevant to their readers. The identification of people, places, ethnic groups and other named entities essentially extends the print practice of adding indices of people and places.[74] Machine actionable annotations for the morpho-syntactic analysis of each word have ancient intellectual roots in pedagogical practice – students have been asked for thousands of years to state which word a given noun or preposition depends upon in a sentence.

Support from the National Endowment for the Humanities (NEH)[75] and the Institute for Museum and Library Services (IMLS)[76] allowed Perseus to develop named entity classification services for Greek and Latin. In the following pasage of Greek text, the names Plato, Menelaos, Homer, Patroklos, and Hector are all classified as being the names of people.

οὐ δεόντως γοῦν <name type="person">Πλάτων</name> τὸν <name type="person">Μενέλεων</name> ἐνόμισεν δειλόν, ὃν ἀρηίφιλον <name type="person">Ὅμηρος</name> λέγει καὶ μόνον ὑπὲρ <name type="person">Πατρόκλου</name> ἀριστεύσαντα καὶ τῷ <name type="person">Ἕκτορι</name> πρὸ πάντων πρόθυμον μονομαχεῖν

Semantic classification by itself is useful, but for many purposes we want to be able to assert that the Plato in a particular passage does indeed describe the famous Greek philosopher rather than the comic playwright of the same name. In some cases, this information can be mined from digitized print indices[77] (although it is not always easy to determine automatically that Alexander-5 in one index is Alexander-3 in the index for another author). In some cases the precise identity of the Antigonus or Alexandria in a given passage is not clear and is the object of scholarly analysis.

---

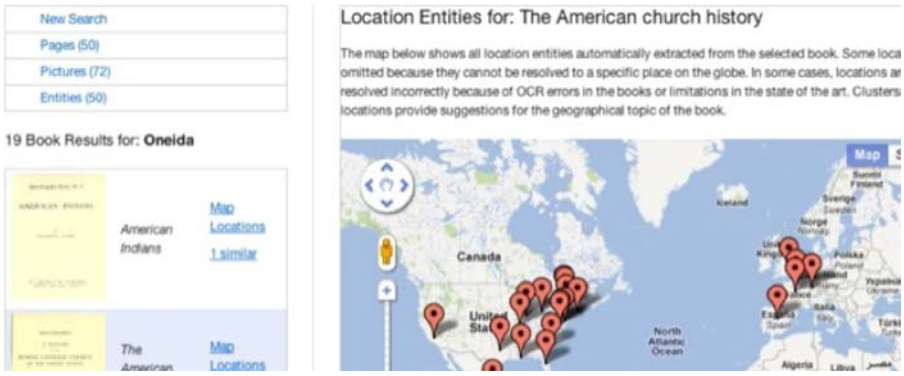
[74] The importance of supporting both the automatic annotation of various named entities within diverse types of historical texts as well as the creation of tools to support users in identifying and annotating such entities has received a great deal of attention in the last few years. For some recent work in these various areas, ZHANG *et al.* (2010), CLOUGH *et al.* (2009), and TOBIN *et al.* (2008).

[75] http://www.neh.gov.

[76] http://www.imls.gov.

[77] For some interesting work on the mining of digitized print indices from historical books for personal and place name identification see PIOTROWSKI (2010) and ROMANELLO *et al.* (2009).

Having the identity of the particular people and places, for example, enables new classes of analysis and visualization. We can, for example, begin to build on machine actionable social network data to trace members of a family or group.[78] A great deal of work has gone into the automatic idenification of places[79] (an inherently easier problem because there are fewer places than people and places do not have children and grand-children nearly so often as do people). The UMASS group has included named entity identification in its architecture. The figure below illustrates frequently mentioned places in a book on church history.



For students of historical languages, richly annotated corpora may be the most important new phenomena from the shift to a digital space. Editors have long included punctuation, capitalization, paragraph breaks and other print annotations based upon their own analysis of the text in order to support contemporary readers. The field of corpus linguistics has developed methods by which to systematically record the linguistic features in a text. An annotated corpus can be queried and its features retrieved and quantified for analysis.

[78] The exact identification of historic individuals is one of the tasks of prosopography and there is growing work in the field of "digital prosopography" with social networks and visualization tools, see for example the project Berkeley Prosopography Services (http://code.google.com/p/berkeley-prosopography-services/) described in SCHMITZ 2009, and also interesting work by GRAHAM & RUFFINI (2007).

[79] Relevant work in place name recognition, particularly in terms of historical language resources and the field of classics, has been reported by the Googling Ancient Places project, see ISAKSEN *et al.* (2012), as well as by the HESTIA project, which has made use of Perseus TEI-XML texts as part of its work, see BARKER *et al.* (2010).

Grammars can then be constructed directly from the full corpus, with explicit statements about the frequency of particular phenomena and links directly back to the

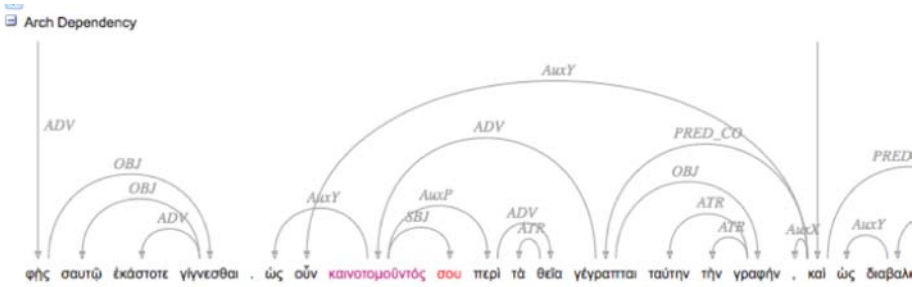
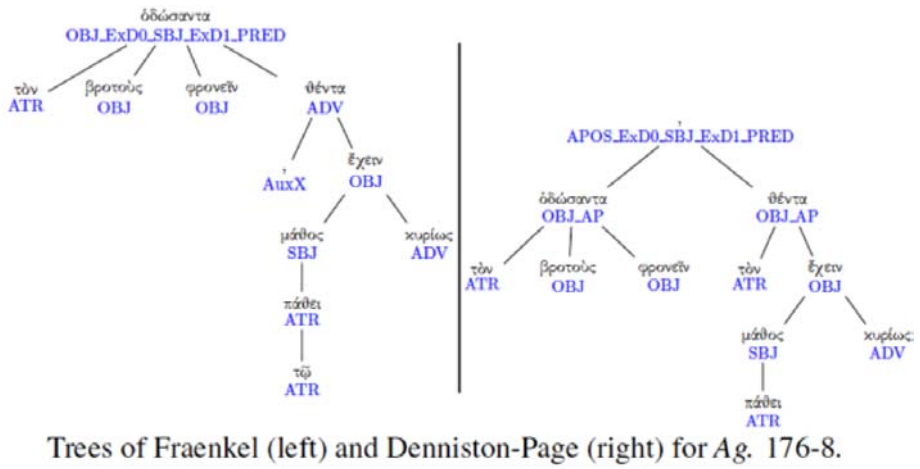


Figure 12: A genitive absolute retrieved from the Euthyprho *of Plato, morpho-syntactically annotated by Giuseppe Celano.*

Richly annotated corpora with systematic morphological and syntactic analyses are often called treebanks because the syntactic structures can be visualized as trees.

Linguists often (in practice) focus upon developing the largest possible corpora because they are looking for typical (and thus repeated) phenomena. More data is, in this case, better data because quantification and statistical significance are fundamental to evidence-driven linguistic research. Philologists focusing intensely on particular texts are often more concerned with exploring multiple ways to construe a particular sentence or phrase. In this case, the goal is not to provide a single plausible interpretation of each sentence but to represent variant interpretations. In the example below, two competing interpretations for one sentence in Aeschylus have been encoded in a dependency grammar. The two hypothetical readings can then be compared to the other sentences in Aeschylus, Greek tragedy or larger corpora as these become available.



Trees of Fraenkel (left) and Denniston-Page (right) for *Ag.* 176-8.

*Figure 13: Interpretations of the same sentence in Aeschylus as proposed by two twentieth-century editors and represented in machine actionable form by Francesco Mambrini (*BAMMAN *et al. 2009).*

Morpho-syntactic analyses are, however, fundamental to global editions because they reveal the underlying structure of a sentence in a general format. Readers with the morpho-syntactic analysis of a sentence and an aligned translation into a language with which they are familiar have the tools with which to pull apart every word in a source of interest to them. The 350,000 morpho-syntactically analyzed Greek and Latin words available in the Perseus Greek and Latin Treebanks provide support for readers regardless of whether their primary language is English, German, Arabic or Chinese.[80] Those who understand English can combine the treebanks with aligned English translations and can begin to work with Greek and Latin directly even before they have begun systematic study of those languages.

Curated treebanks are not only useful for precise study and analysis with methods from corpus linguistics; these curated treebanks also provide data from which automated systems can learn to perform morphological and syntactic analysis. In general, the more morphological and syntactic training data available that is relevant to a given corpus, the more accurate the automatic analyses will be.

[80] http://nlp.perseus.tufts.edu/syntax/treebank/.

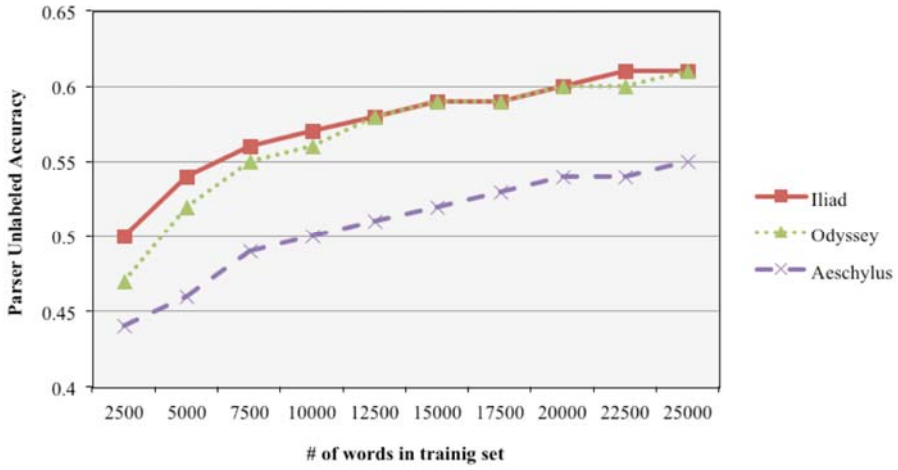


Figure 14: Learning curves for the Iliad*,* Odyssey *and the works of Aeschylus (Saeed Majidi)*

The figure above tracks the growing accuracy of an automatically trained syntactic parser as the training set increases. Saeed Majidi, a PhD candidate in Computer Science at Tufts University computed these figures by using curated syntactic analyses for the Homeric Epics and for Aeschylus, training the parser on part of the curated data and then running the parser against the rest, comparing the parser output with the curated analyses. Two thousand years of students who have worked on Classical Greek would not be surprised to see that Aeschylus is harder than Homer for machines as well as for human beings.

Even noisy syntactic data can be very useful if it is large enough – in effect, errors tend to be random while significant results cluster into significant patterns. In other words, the signal will, in many cases, be visible despite the noise. Relatively modest training sets (10,000-50,000) can generate automatic syntactic analyses that are 50-60% accurate and that provides a great deal of useful data.

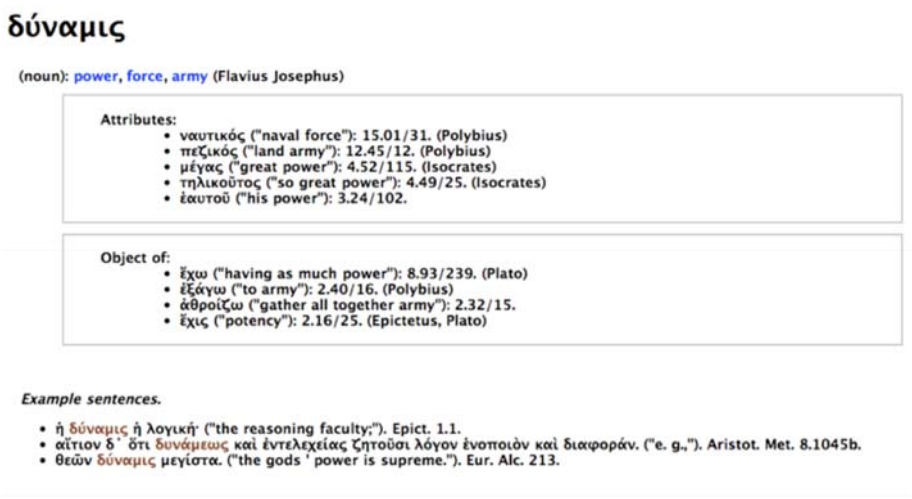


Figure 15: Dynamic Lexicon Entry for the Greek noun δύναμις (David Bamman)

The figure above presents work from the Dynamic Lexicon project,[81] which applied computational methods to extract basic lexical data. The figures above are derived from a corpus of 8 million words of Greek, of which c. 5 million have been aligned with English translations. "While the automatically induced information naturally contains noise (e.g., the misclassification of ἕχις or the mistranslation of the second example sentence), it reveals larger patterns of usage consistent with traditional lexica. In particular, we have automatically induced three categories of information:

- ***Morphology.*** This entry has correctly categorized δύναμις as a noun. Some lexemes have multiple parts of speech – e.g., the very common word καί can be used as a conjunction ("and") and as an adverb ("even") and has different sense and syntactic behavior as a result of this distinction.
- ***Sense.*** By aligning all our Greek source texts with their English translations at the level of individual sentences and then words, we have induced that δύναμις has three predominant senses in all of Greek literature – "power," "force," and "army" – and that "army" itself is an especially dominant sense in the works of Flavius Josephus.
- ***Syntax.*** The availability of syntactically-parsed data allows us to calculate that the most common attributes for δύναμις are ναυτικός

[81] See BAMMAN & CRANE (2008).

("naval") and πεζικός ("on foot") – both especially dominant in the works of Polybius. The alignment of parallel texts lets us present appropriate translations of each (e.g., a naval *force* rather than a naval *army*)

In addition, the availability of Greek/English and Latin/English parallel text that has been aligned at the level of individual sentences also allows us to supplement the lexical entry with several instances of its actual use in text – allowing us to present not only the source text but also its automatically aligned translation."[82]

The Dynamic Lexicon cannot create finished articles on the grammatical usage and meanings of a word but it does provide a starting point – and more importantly it scales to large collections. The *Thesaurus Linguae Latinae* (TLL), begun in 1894, is creating a lexicon for Latin through c. 600CE. Its staff page lists 23 names,[83] including a general editor, four editors, and twelve collaborators. "The work is based on an archive of about 10 million slips which takes account of all surviving texts. In the older texts there is a slip for each occurrence of each word; the later ones are generally covered by a selection of lexicographically relevant examples."[84] As of 2012, published volumes of the TLL had reached the beginning of the letter "r".[85]

There are now billions of words available in Latin. Approaches such as those demonstrated in the Dynamic Lexicon grow more, rather than less, effective as the collection size increases. But the accuracy of those automated processes depends upon the size and quality of the training data. Each digital edition not only serves an immediate circle of human readers but also contributes new data to intelligent services, some already in operation and surely others that we cannot yet predict. The digital edition is distinguished by its ability to support interaction between each individual reader and a growing network of increasingly sophisticated services.

The Greek and Latin Dependency Treebanks available from Perseus represent a basic standard. They encode morphological form and syntactic function but they do not include other features (such as

---

[82] http://nlp.perseus.tufts.edu/lexicon/ -- quoted text and research by D. Bamman.

[83] http://www.thesaurus.badw.de/english/index.htm -- accessed on October 26, 2012.

[84] http://www.thesaurus.badw.de/english/index.htm.

[85] http://www.badw.de/publikationen/kommissionen_publ/thesaurus/index.html Vol. XI 2 Fasc. I: r – rarus. Redaktoren: J. Blundell, S. Clavadetscher, C. G. van Leijenhorst.

co-reference resolution, which specifies who the "he" or "they" are in a given sentence). The Greek and Latin Treebanks represent only a conservative first step, representing only the most obvious annotations that should accompany digital texts. The dominant shape of digital editions will depend upon a social consensus that will evolve over time. The morpho-syntactic analyses reflect a very conservative estimate of what will be expected either a decade or a generation from now.

7. New forms of intellectual production. Wikipedia will almost certainly be remembered as the single most important advance for the humanities from the early twenty-first century. Wikipedia as a particular project may or may not flourish over time but it has nonetheless demonstrated a fundamentally new mode of intellectual production, one that is far more deeply collaborative than any of its immediate print predecessors.[86] Humanists who question the potential of this medium because they find the articles in their area problematic might spend time working with Wikipedia articles on mathematically complex topics (one example of which is shown in the figure below). These cover concepts quite as challenging as any that students of historical languages face. If the articles on Greco-Roman topics are not as impressive as those for various mathematical sciences, then that only means that those of us who advance understanding of the past as a vocation have ourselves not developed the broader community of interest.

[86] The volume of both scholarship and academic commentary on either the importance of or the disaster of Wikipedia as both a collaborative knowledge creation model and as a reference source is far too vast to wade into here, but for some differing perspectives, see the seminal piece on open source history by R. ROSENZWEIG (2006), for an example of using Wikipedia articles as a model to improve student writing (GRAHAM 2012), and for faculty uses of and responses across the disciplines (DOOLEY 2010, WHALLEY 2012).

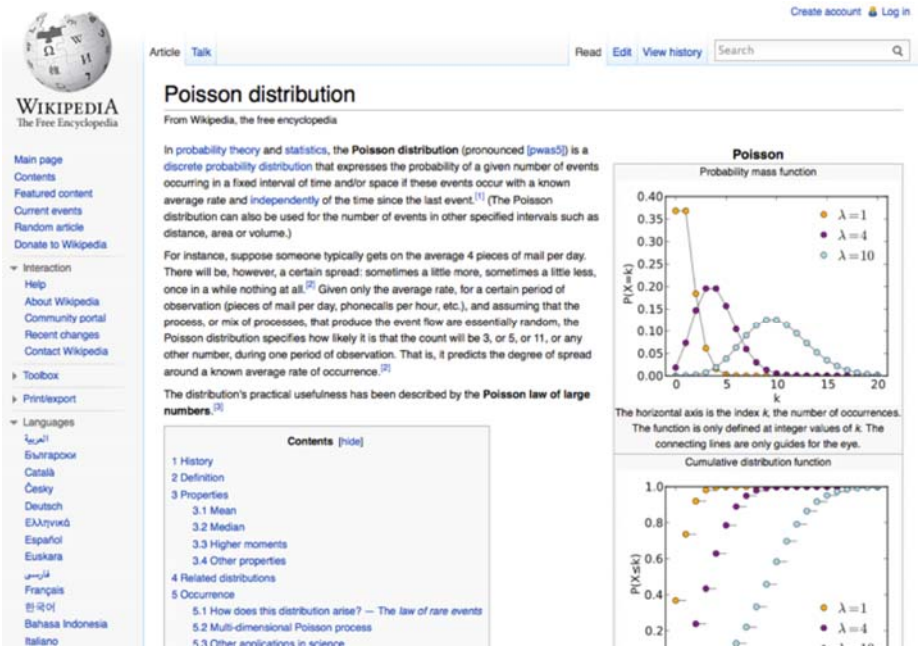


Figure 16: Wikipedia article on the "Poisson distribution" in probability theory (as of October 24, 2012). The decentralized mode of intellectual activity produces an immense amount of clear, accurate exposition on topics quite as complex as those addressed by students of historical languages.

The Homer Multitext Project[87] (HMT) may well be the most important project that has emerged within Classical studies since the beginning of the twenty-first century.[88] Only within that recent time frame have we had the technology to create, store, distribute and license very high-resolution images of manuscripts. The first three changes reflect decreases in the costs of digital cameras, storage and bandwidth. The fourth feature may be less obvious but machine-actionable licenses, such as those available in a growing number of languages, provided by Creative Commons[89] are essential for scalable work with digital sources. In the first generation of digital work, licenses were written in expository prose and could differ in multiple ways. If one wished to create a work with materials from different

[87] http://www.homermultitext.org/.

[88] For more on the history and scholarly future of the HMT, see NAGY (2010), and for an outline of the technical choices, see SMITH (2010).

[89] http://creativecommons.org/.

sources, each source required a separate agreement. Such a procedure does not scale to projects that may draw upon thousands of different sources, especially when projects may dynamically detect and repurpose newly available materials (e.g., a morphological and syntactic analysis engine that generates annotations for Greek and Latin sources as these become available).

The HMT seeks to represent the textual history of the Homeric Iliad and Odyssey in its full complexity. This task is particularly challenging because the Homeric epics emerge from an oral poetic tradition that was formulaic and fluid in nature. Thus the HMT is not attempting to create a single authoritative edition but rather to represent every detectable version of the Homeric epics.[90] To do so requires far more detailed publication of the surviving manuscripts than has ever been feasible before. The general idea behind the HMT is not necessarily new – Milman Parry and Albert Lord articulated models of oral composition for the Homeric epics in the twentieth century. The method behind the HMT represents a sharp departure from recent practices.

Undergraduate researchers play fundamental roles in the HMT.[91] The most knowledgeable experts of particular manuscripts are juniors and seniors who have worked for years on these documents and who publish their findings. The summer of 2012, for example, saw research published by Stephanie Lindeborg on "Catalog of Ships Summary Scholia Part Two: Comparing the Y.1.1 with the Venetus B" and "Catalog of Ships Summary Scholia in the Escorial Y.1.1"[92], Matthew Angiolillo and Christine Roughan on "Scholia to Iliad 14.506 in Two Manuscripts in Venice (Venetus A and Marciana 458)"[93] and Thomas Arralde on "Identifying Aristarchean Commentary in the Venetus A Scholia."[94] The expository form of this research follows the traditions of expository prose that have evolved over millennia.

---

90 For further discussion of these issues, see DUÉ & EBBOTT (2009).

91 To read more about the role of undergraduate researchers and the HMT, see BLACKWELL & MARTIN (2009).

92 http://homermultitext.blogspot.de/2012/08/catalog-of-ships-summary-scholia-part.html; http://homermultitext.blogspot.de/2012/08/catalog-of-ships-summary-scholia-in.html.

93 http://homermultitext.blogspot.de/2012/07/scholia-to-iliad-14506-in-two.html.

94 http://homermultitext.blogspot.de/2012/06/identifying-aristarchean-commentary-in.html.

The relationship between the arguments and the data within the manuscript is radically traditional – it departs from the print conventions by more fully realizing the ideals of scholarly argumentation. These publications explicitly document their arguments with high-resolution images of those sections of the manuscripts upon which they base their arguments. At the same time, these particular images contain the coordinate data that allows automatic linking directly into the archival images, available at high resolution and often in multiple spectra of light.[95] Assertion and evidence are far more tightly – and consistently – linked than was ever feasible in print – especially when arguments depended upon extensive visual imagery. The underlying idea is deeply traditional – footnotes have for centuries allowed us to define our sources. But we can realize that traditional idea much more fully.

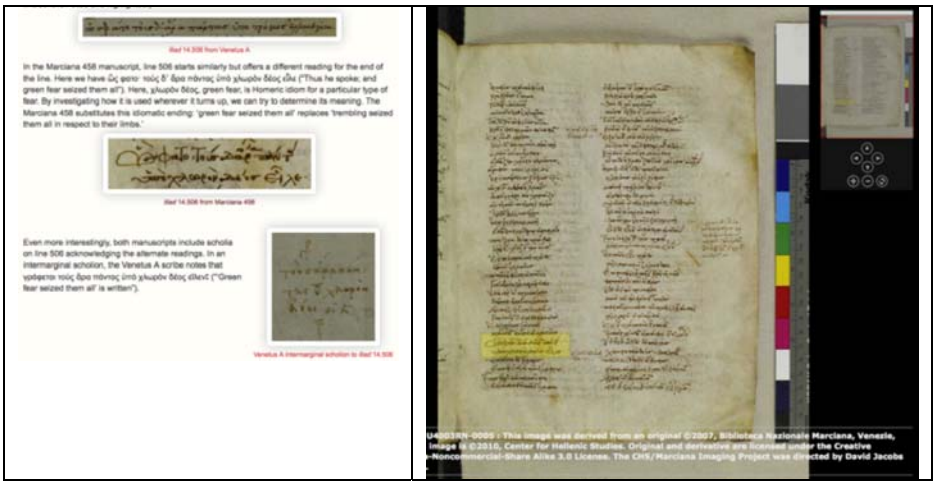


Figure 17: Citations to particular passages in a manuscript include coordinate data that enables dynamic linking into images available at high resolution and in multiple spectra of light.[96]

The HMT demonstrates a new culture of intellectual activity, one in which undergraduates have an opportunity to develop their own

[95] The ability to create tools or programs that can at least semi-automatically link manuscript transcriptions directly to images, particularly at the word level, has been a subject of active research, see for example, FISCHER *et al.* (2011), PORTER *et al.* (2009), and CAYLESS (2008).

[96] Example drawn from http://homermultitext.blogspot.de/2012/07/scholia-to-iliad-14506-in-two.html.

voices and to contribute in substantive ways. The figure below uses different colors to mark different elements and logical relationships within one page of the tenth century Byzantine Venetus A manuscript. There are at least four categories of annotation associated with the text of the Iliad (left of the text, right of the text, interlinear, surrounding the text) and various relationships between the scholia, the text and each other.

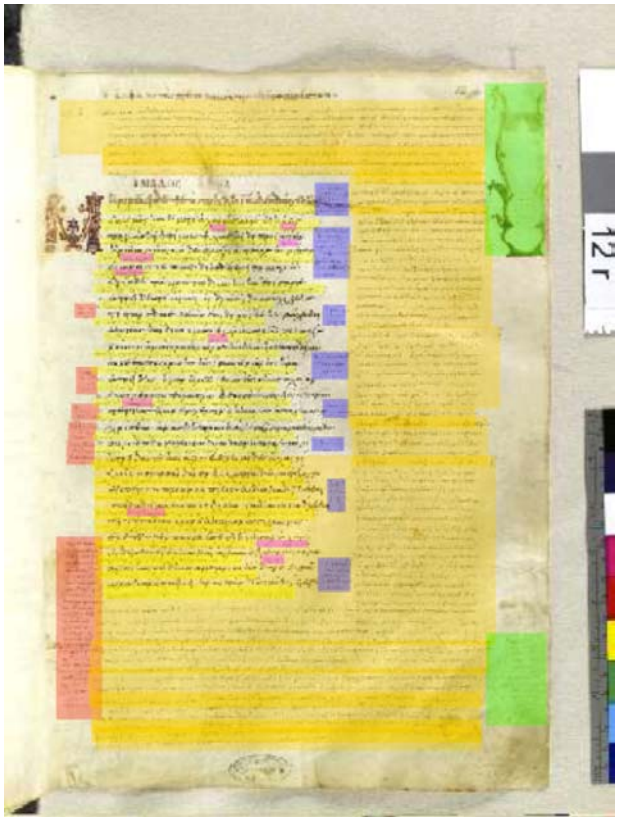


Figure 18: Venetus A, folio 12 recto, with the first 25 lines of the Iliad*; overlays show the location of scholia, color-coded for their class of placement on the folio.*[97] *First year students of Greek were able to create these overlays, providing them with an early opportunity in their careers to use their incipient knowledge of Greek to contribute fundamental data that no machine could provide.*

[97] http://homermultitext.blogspot.de/2012/07/verifying-inventory-of-scholia.html.

No page layout system can identify the regions of the manuscript page above. Nor can existing systems for handwriting analysis determine the first and last lines of the *Iliad* in the central textual section on the page above. These are, however, fundamental tasks for the analysis of the manuscript as a whole. Students of Greek can, however, as early as their first year, begin to contribute such analyses, learning how to interpret the manuscripts as a whole and how to associate the Byzantine script to the characters that they learned in their textbooks and the Greek poetry that they aspire to read.

Ultimately the HMT upon far more detailed transcriptions and representations of the textual data than were ever published in print. In August 2012, the HMT published TEI XML transcriptions of the Iliadic text and scholia from *Iliad* 1-6 in the Venetus A manuscript, and other texts from the first eleven folios of the Venetus A manuscript. Undergraduates at Furman, Holy Cross and the University of Houston produced these transcriptions, working with each other and with their faculty collaborators over several years.

In the twentieth century, the study of manuscripts involved the specialized field of palaeography.[98] Advanced researchers might have an opportunity to take seminars in this subject, working often with facsimiles of the originals produced as large-scale books or as microfilms. Few, if any undergraduates, took such courses – they were expected to focus on learning the standardized Greek and Latin of their critical editions. In the twenty-first century, we find undergraduates energized by access to very high-resolution images of these originals and (like their counterparts in the growing citizen science movement) by the realization that they can contribute to human knowledge. At Holy Cross and Furman, enrollments in Classical Greek have expanded – with 2,898 and 2,951 students each, both schools have more than 25 students in introductory Greek. Undergraduate interest in manuscripts has led to a new open palaeography project.[99] The Holy Cross Manuscripts, Inscriptions and Documents Club – a student organized, volunteer organization – advances "the study of these academic fields: paleography, codicology, epigraphy, as well as the study of languages. We strive for undergraduate

[98] For one perspective on how the study of palaeography is changing with the availability of digital methods, see CIULA (2009).

[99] http://homermultitext.blogspot.de/2012/10/announcing-open-paleography-project.html.

inclusion in work normally reserved for the graduate level."[100] "At the club's first general meeting of the new academic year on Friday, seventeen returning members and three faculty collaborators were joined by twenty newcomers. Six of the club's most active members could not attend Friday's meeting because they are currently studying abroad, but they have already sent back photographs of inscriptions as part of a club project on the epigraphic sources for tribute in fifth-century Athens, just one of an expanded roster of projects the club is hosting this year."[101]

Others have encountered the enthusiasm that students and the general public show when working with original sources.[102] The HMT is important because the Byzantine Greek manuscripts offer great challenges of form (they contain many abbreviations as well as handwriting that is very different from modern Greek fonts) and of content (they contain not only the archaic poetic dialect of the Homeric epics but much later technical prose of commentators writing about grammar, meter, style, and other subjects). The HMT demonstrates the feasibility of a very hard case. If undergraduates working together and with their faculty can produce data about and research on these Homeric manuscripts, they can contribute a wide range of challenging subjects in many languages.

The HMT and the Greek and Latin treebanks each contribute essential components to a mature digital edition. The HMT addresses the challenge of documenting textual witnesses that are inherently complex in form and that cannot be analyzed by methods such as OCR or handwriting recognition. The Greek and Latin treebanks provide the linguistic analyses for the phenomena transcribed from various paper, papyrus or stone sources. Both share a common

---

[100] http://shot.holycross.edu/hcmid/.

[101] http://homermultitext.blogspot.de/2012/09/undergraduate-interest-in-manuscripts.html.

[102] Another example from Classical studies can be found at http://udallasclassics.org/maurer_files/Valla-Intro.htm, which publishes transcriptions of Lorenzo Valla's translation of Thucydides into Latin: "The motive was given by an undergraduate Thucydides course at the University of Dallas, in fall 2008, where at my suggestion, two students chose to transcribe Valla's translation of the Plataean Debate (using Stephanus' text) instead of writing a term paper. I suggested this knowing that it would help both their Latin and their Greek, and give them a glimpse (normally denied to undergraduates) of the rich (in Thucydides' case peculiarly, immensely rich) history of classical scholarship. But when I saw that they did this work with gusto, remarkably carefully and accurately, it occurred to me that it might interest others too; so I added the apparatus, and now put the whole thing online."

philosophy that emphasizes the links between assertions and the data upon which those assertions are based. While the HMT links transcriptions to images, the Greek and Latin treebanks allow us to link assertions about particular linguistic phenomena to the precise places where those phenomena occur.

And like the HMT, the Greek and Latin treebanks depend upon collaboration among students and professional researchers. Two undergraduate or MA-level students independently proposed morphological and syntactic analyses for 230,000 words in the Homeric *Iliad* and *Odyssey*. A professional Homerist, Jack Mitchell, resolved those instances where two different analyses were proposed. The result was a data set in which each sentence has identifiers for the initial annotators and the expert reviewer. Each sentence constitutes a distinct, citable publication that sets out to describe a defensible interpretation.


```
- <sentence id="3044" document_id="Perseus:text:1999.01.0133" subdoc="book=6:card=1" span="pa/ntas0:.4">
    <primary>mpkinn10</primary>
    <primary>millermo</primary>
    <secondary>nicanor</secondary>
    <word id="1" form="pa/ntas" lemma="pa=s1" postag="a-p---ma-" head="3" relation="OBJ"/>
    <word id="2" form="ga\r" lemma="ga/r1" postag="g--------" head="3" relation="AuxY"/>
    <word id="3" form="file/esken" lemma="file/w1" postag="v3siia---" head="0" relation="PRED"/>
    <word id="4" form="o(dw=l" lemma="o(do/s1" postag="n-s---md-" head="5" relation="ADV"/>
    <word id="5" form="e)/pi" lemma="e)pi/1" postag="r--------" head="7" relation="AuxP"/>
    <word id="6" form="oi)ki/a" lemma="oi)ki/on1" postag="n-p---na-" head="7" relation="OBJ"/>
    <word id="7" form="nai/wn" lemma="nai/w2" postag="t-sppamn-" head="3" relation="ADV"/>
    <word id="8" form="." lemma="period1" postag="u--------" head="0" relation="AuxK"/>
  </sentence>
- <sentence id="3045" document_id="Perseus:text:1999.01.0133" subdoc="book=6:card=1" span="a)lla/0::2">
    <primary>mpkinn10</primary>
    <primary>millermo</primary>
    <secondary>nicanor</secondary>
    <word id="1" form="a)lla/" lemma="a)lla/1" postag="d--------" head="14" relation="AuxY"/>
    <word id="2" form="oi(" lemma="e(/1" postag="p-s---md-" head="8" relation="OBJ"/>
    <word id="3" form="ou)/" lemma="ou)1" postag="d--------" head="8" relation="AuxZ"/>
    <word id="4" form="tis" lemma="tis1" postag="p-s---mn-" head="8" relation="SBJ"/>
    <word id="5" form="tw=n" lemma="o(1" postag="l--------" head="4" relation="ATR"/>
    <word id="6" form="ge" lemma="ge1" postag="g--------" head="5" relation="AuxZ"/>
    <word id="7" form="to/t'" lemma="to/te1" postag="d--------" head="8" relation="ADV"/>
    <word id="8" form="h)/rkese" lemma="a)rke/w1" postag="v3saia---" head="14" relation="PRED_CO"/>
    <word id="9" form="lugro\n" lemma="lugro/s1" postag="a-s---ma-" head="10" relation="ATR"/>
    <word id="10" form="o)/leqron" lemma="o)/leqros1" postag="n-s---ma-" head="8" relation="OBJ"/>
```


Figure 19: Morphological and syntactic analyses represented as XML. Each sentence contains a unique identifier for the two annotators (<primary>) and the Homerist (<secondary>) who reviewed their contributions to create the final collaborative entries in the Treebank.

The workflow used to develop the treebank data for Homer was designed to produce data of high accuracy but it was, in its initial form, slow. Months might pass after a first student created an initial annotation before a second annotation was created and the two were compared. There was no mechanism to provide students with significant feedback. The goal was to generate data.

But the treebanking process can be organized to produce data of high accuracy quickly and to give students feedback as they create that data. The figure below illustrates how two different students in a third semester Latin class differently annotated the same sentence. In this scenario, students can independently annotate one or more sentences, then work together to resolve the different interpretations, present the final results (with questions) to the class and instructor and publish, by the end of the class, the results as data for comment. The class can build up their own corpus over a semester, eliciting comments and feedback from the broader community and making such adjustments as they see fit. New interpretations can – and inevitably will – be proposed long after the class. The results can be quite accurate.

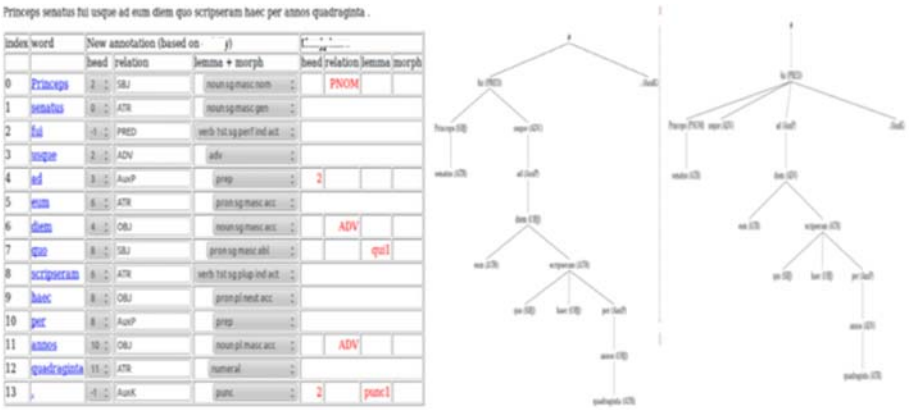


Figure 20: Individual sentences analyzed by third semester Latin students. The left display shows in red where students differed in their analyses. The right display visualizes the interpretations as trees. We will be able to support such dynamic activities, where individuals, whether in the same classroom or in completely different locations, can compare their analyses, revising or defending their choices. In a classroom setting, the instructor can help adjudicate and classroom work can, where a consensus appears, be immediately submitted as a contribution to the Greek or Latin Treebanks, with instructor and students as joint, named contributors.

Students have accounted for the morphological and syntactic functions of words in Greek and Latin since grammatical analysis began in antiquity but this ancient pedagogical practice can now produce much of the linguistic data that we need, both to rebuild our understanding of Greek, Latin and other historical languages on

explicit, evidence-based models and to support a global audience of readers from many different linguistic and cultural backgrounds.

The greatest challenge facing Greek, Latin and other historical languages is social rather than technical. A new intellectual culture has begun to emerge that reflects the strengths and possibilities of a society where ideas circulate primarily in digital, rather than print, forms. The departments that provide doctoral training for new researchers remain, certainly in the field of Greek and Latin, deeply rooted in a traditional print culture that emphasizes single authored, static publications and specialist audiences rather than collaborative research, dynamic knowledge bases (of which a digital edition constitutes a special case) and the relentless effort to use specialized scholarship to advance the general life of society.

A new generation of researchers is increasingly eager to move forward, if only because many realize that fields that do not exploit the strengths of digital culture are at a disadvantage and because students of historical languages have enough disadvantages in the twenty-first century. A NEH-funded three week institute on Working with Texts in a Digital Age[103] attracted almost eighty applications for twenty-five slots. All of the participants – most of them early in their careers and under pressure to complete PhDs or to crank out publications – had agreed to devote a substantial part of their summer time to acquiring new skills and they thus reflected a self-selected group with a stated interest in digital methods. Most expressed profound surprise at how much was, in fact, possible. Even those who were most active already on digital projects had little, if any, exposure to immediately applicable methods from either corpus or computational linguistics.

We are poised for a shift in the intellectual culture of the humanities as a whole and of philology in particular. In the twentieth century, departments of Classics in the United States and elsewhere began, of necessity, to develop curricula for students who studied little or no Greek and Latin. Such a move was necessary because of the decline in the number of students who entered college with background in either of these languages. The APA has even begun serious consideration of changing its name – "the term philology has become so obscure to all but practitioners as to impede

[103] http://sites.tufts.edu/digitalagetext/.

our efforts to gain broader public (even academic) visibility."[104] We have certainly come a long way from 1956, when the mad scientist of the film *Forbidden Planet*[105] was a philologist. But the present obscurity of the term creates an opportunity to reinvent and refashion its meaning and to assert, in fact, a meaning much like that of Friedrich Wolf in eighteenth century Halle and Augustus Boeckh in nineteenth century Berlin, for whom philology aimed at fostering an understanding of antiquity as a whole (*cognitio universae antiquitatis*) and a means to breath life back into the past. As the Greek and Latin sources become accessible to a global audience, the old term for studying these sources directly may reassert itself and become a symbol of a reborn field.

Nevertheless, we see now in the twenty-first century opportunities to re-integrate the language into our curricula, both by making the language more accessible and by making contribution and research feasible for our undergraduates. We have an opportunity in the study of Greek, Latin, and other historical languages to be leaders in fostering a new generation of student researchers and citizen scholars. An opportunity is, however, not inevitability, and no technological determinism will save or overwhelm us. How well we realize the possibilities emerging before us will depend upon decisions that we make as communities and as individuals.

The role of Germany

This paper builds upon a 2011 talk delivered to the Berlin-Brandenburg Academy of Sciences, the twenty-first century successor to the Prussian Academy of Sciences founded by Gottfried Wilhelm Leibniz in 1700 more than 300 years before. In that period, Germany became, for many years, the primary center for scholarship on Greek and Latin. Early in the twenty-first century, Germany has a unique opportunity to build upon this tradition of scholarship and to advance a global dialogue among civilizations.

First, Germany now occupies a unique position within the world. The strongest economic power within the European Union, Germany also lacks the complicating background in global affairs that color perceptions of the geopolitically active Anglo-American nations.

---

[104] Jeff Henderson, APA president: http://apaclassics.org/index.php/apa_blog/apa_blog_entry/request_for_comments_on_possible_name_change_for_association/.

[105] http://www.imdb.com/title/tt0049223/.

Within the diplomatic conditions of the early twenty-first century, no country in Europe or North America is better situated to advance a dialogue among civilizations than is Germany.

Second, in the period between 1700 and the present, more editions of Greek and Latin may have been produced in the area of contemporary Germany than in the rest of the world – Leipzig, in particular, was the greatest center for the publication of Greek and Latin print editions through the Second World War. And German authors produced an immense stream of original Latin in virtually every written genre and on every topic from the medieval period through the twentieth century. This immense body of Greek and Latin represents a major component of German cultural heritage and well deserves digital publication. A library of Greek and Latin produced in the German speaking lands would be of immense value to those interested not only in the texts themselves but also in the intellectual and cultural history of Europe.

Third, German academic traditions do not separate computer science from the humanities – both are instances of *Wissenschaft*, where the English term "science" is used exclusively for the natural and, when qualified, social sciences. The semantic distinction has immense practical consequences in the Anglo-American world. In the United States, for example, the NEH[106] (with a 2010 budget of around US$167 million) and the NSF[107] (with a 2010 budget of around US$6.89 billion) are officially separate organizations that serve different communities. The NSF can support computer scientists working on applications in biology, physics, earth sciences, or any other NSF-supported discipline but the NSF cannot readily support computer science research on subjects that belong to the NEH. With a budget 40 times smaller than the NSF, the NEH simply cannot provide significant support for computer science research, however important that may be for the humanities.[108] Efforts such as

[106] http://www.neh.gov.

[107] http://www.nsf.gov.

[108] The NEH Preservation and Access Research and Development track can provide up to $350,000 (http://www.neh.gov/grants/preservation/preservation-and-access-research-and-development) -- a very large sum for NEH grants but well below the $500,000 cap for small grants awarded for Computer Science research by the NSF: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12765&org=CISE&sel_org=CISE&from=fund.

the Digging into Data Program[109] depend upon ad hoc collaborations to bring NEH and NSF funded research together.

In Germany, computer scientists face no structural barriers if they wish to focus their research upon problems from the humanities. In 2012, the German Ministry of Education[110] announced that it had provided 19.5 million Euros to support research projects that involved computer science and the humanities. In April of 2012, the Humboldt Foundation announced my own election to a Humboldt Chair of Digital Humanities, a chair situated in a Department of Computer Science at Leipzig and bringing with it support of 5,000,000 Euros over five years. Leipzig was already hosting projects with joint humanist and computer scientist teams with aggregate support of c. 1 million Euros a year. Other such collaborations between humanists and computer scientists can be found around Germany. The overall consequence of this for the humanities in a digital world could be profound in the long run. In Germany, emerging researchers in computer science can explicitly build a career on collaboration with humanists. If the 2012 19.5-million euro BMBF investment draws promising computer scientists into long-term research agendas relevant to the humanities, that one program can shape development for decades.

Fourth, Germany passed in 1965 an explicit law to define the rights status of editions. Sophocles and Vergil may be long gone, but German law provides protection to scientific editions for a period of 25 years after publication.[111] Textual notes on the bottom of the page in many editions may be considered a separate original work and qualify for the regular European protection of the life of the author + 70 years. This complicates redistribution of the text as scanned image book because the textual notes on the bottom of the page would have to be excluded. Nevertheless, the reconstructed text can be manually marked before or after the books are scanned, and methods exist to identify the text automatically. The reconstructed texts of editions published through 1987 can be redistributed in 2012, with a moving wall freeing a year's worth of editions with each new calendar year.

The German situation does not reflect the full needs of scholarship. Scholars who handed over their introductions and

---

[109] http://www.diggingintodata.org/.

[110] www.bmbf.de.

[111] http://de.wikipedia.org/wiki/Schutz_wissenschaftlicher_Ausgaben.

textual notes to publishers can expect that, under current law, their work will not be able to circulate freely for scholarly analysis until all of their immediate colleagues are long dead – a grandchild ten years old at the editor's death would be eighty before the editorial data was available. But, of course, even if the printed editions were released, they do not represent their data in a machine actionable format (e.g., you can't use a digitized apparatus to compare dynamically the contributions of multiple witnesses) and they do not include the full range of data for a true digital edition (e.g., commas, periods, and other annotations from print culture are imposed upon the original text but print editions do not record the morphological, syntactic and other analyses behind punctuation and page layout in any form, machine- or human-readable).

Nevertheless, recently printed books lend themselves to OCR better than do older books. OCR software could be applied to a library of page images from editions whose authors have not been dead 70 years but that were published 25+ years ago. The OCR-generated text can then be aligned to other editions and the scholarly community can then quickly see how individual passages in this edition relate to others that are available online. Because editors worked on Greek and Latin sources from the fifteenth century through the present, one or more complete editions – including introduction and textual notes – is available for digitization for virtually every Greek and Latin source printed from manuscript sources.

Conclusions: what is to be done?

If in creating digital editions we wish to foster a dialogue among civilizations – and not all editors may share this goal – we need to work from the two convergent directions of breadth and depth. First, we need to make very large bodies of linguistic sources accessible with methods that are not only scalable but that become more effective as collections grow larger. Second, we need to build upon methods by which to represent our textual sources and linguistic data more precisely, with dense and growing webs of machine actionable annotations that either perfect print practice (e.g., back-of-the book indices of people and places become links to authority lists) or represent a major step forward (e.g., encoding morpho-syntactic analyses, co-reference resolution etc). In effect, students of historical languages must draw upon the results of computational linguistics to account for phenomena at scale and corpus linguistics for intensive analysis. Our goal must be to serve dozens, if not

hundreds, of historical languages, but Greek and Latin provide a starting point: they are big enough and complicated enough for us to develop methods for working with historical languages embedded in much larger collections of modern language materials.

First, to address breadth, we need to put as much of the human textual record as possible online for computational analysis and for the results of that analysis to be shared freely. A great amount of the underlying scanning has already been done. The Internet Archive offers 3.6 million books for public download, HathiTrust currently has 3.2 million public domain books, and Gallica offers more than 1 million books and manuscripts. The original scans of these books should be made available where researchers can apply OCR software customized for particular languages. Such aggregation requires storage as well as computational power.

Second, one can begin by focusing on subsets such as the 65,000 public domain titles out of c. 90,000 that the HathiTrust lists as being in Ancient Greek or Latin. But the real challenge is to find not only the Ancient Greek and Latin in such obvious places but to also track all the quotations of Greek and Latin scattered throughout the other three million plus books. Such tracking includes recognizing passages written primarily in some other language (e.g., English or German) that have quoted shorter passages in Greek or Latin so that we can run customized OCR on the relevant chunks of those pages. Such tracking also includes the ability to recognize as many instances of text reuse as possible, including quotations of a modern language translation of a Greek or Latin work, paraphrases, citations (e.g., Th. 1.32 refers to Thucydides book 1, chapter 32) and names (distinguishing Aristotle the philosopher from Aristotle Onassis).

The HathiTrust Research Center[112] has provided an initial approach to solving this problem for researchers in the United States. This approach is itself evolving but even if perfected for users in the United States, work needs to be done for researchers in Europe, where copyright laws are different and different materials are in the public domain. Germany has a real opportunity to lead in this case because it can provide funding for computer science and humanities collaborations and because of its special copyright laws for editions, which create a moving wall that brings 25-year old editions into the public domain each year.

[112] http://www.hathitrust.org/htrc.

Third, we need to not only educate philologists about new, more intensive, machine actionable methods of representing textual data (such as providing not only punctuation but the morphological and syntactic analyses that punctuation assumes) but also enable them to make informed decisions about how to fashion their work for a rapidly changing intellectual world.

In this we need to engage not only advanced researchers in editing, and library professionals in documenting, historical sources, but we must also involve a generation of student researchers and citizen scholars upon whom we must rely if we are to make the individual documents within the vast and growing digital collections intellectually accessible. Here the means is also the end – at least, insofar as we believe that the end of our work is to advance the intellectual life of humanity and engage society as broadly and deeply as possible.

Two hundred years ago, Augustus Boeckh saw already that the true aim of philology was to understand the ancient world as fully as possible but he also understood that the study of the past was important because it contributed to the lived experience of society as a whole. And one could find such statements from scholars for hundreds and thousands of years before Boeckh, in every corner of Europe, in Baghdad and Cairo, and, of course, in Alexandria. In the end, our methods may change but our goals do not. We honor in the present those values of the past that we most admire by re-imagining those values to serve the future.

REFERENCES

AGOSTI, M. & N. FERRO, 2007: A Formal Model of Annotations of Digital Content, in: *ACM Transactions on Information Systems* 26(1), Article No. 3, 1-57.

ALMAS, B. *et al.*, 2011: What Did We Do With A Million Books: Rediscovering the Greco-Ancient world and reinventing the Humanities, White Paper Submitted to the NEH, National Endowment for the Humanities. http://hdl.handle.net/10427/75558.

ANTONIADIS, G. *et al.*, 2009: Integrated Digital Language Learning, in: BALACHEFF, N. *et al.* (ed.), *Technology-Enhanced Learning*, Dordrecht, Chapter 6, 89-103.

BABEU, A., 2008: Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience. Technical report. http://www.perseus.tufts.edu/~ababeu/PerseusFRBRExperiment.pdf.

BABEU, A., 2011: *Rome Wasn't Digitized in a Day: Building a Cyberinfrastructure for Digital Classicists*, Council on Library and Information Resources. http://www.clir.org/pubs/abstract/reports/pub150.

BAMMAN, D. & G. CRANE, 2008: Building a Dynamic Lexicon from a Digital Library, in: *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, New York, 11-20. http://hdl.handle.net/10427/42686.

BAMMAN, D. & D. SMITH, 2012: Extracting Two Thousand Years of Latin from a Million Book Library, in: *Journal on Computer and Cultural Heritage* 5(1), Article No. 2. http://dx.doi.org/10.1145/2160165.2160167.

BAMMAN, D. *et al.*, 2009: An Ownership Model of Annotation: The Ancient Greek Dependency Treebank, in: *TLT 2009-Eighth International Workshop on Treebanks and Linguistic Theories*. http://hdl.handle.net/10427/70399.

BAMMAN, D. *et al.*, 2010: Transferring Structural Markup Across Translations Using Multilingual Alignment and Projection, in: *JCD '10: Proceedings of the 10th annual joint conference on Digital libraries*, New York, 11-20. http://hdl.handle.net/10427/70398.

BARKER, E. *et al.*, 2010: Mapping An Ancient Historian In A Digital Age: The Herodotus Encoded Space-Text-Image Archive (HESTIA), in: *Leeds International Classical Studies* 9. http://www.leeds.ac.uk/classics/lics/2010/201001.pdf.

BEAULIEU, M.-C. & B. ALMAS, 2012: Digital Humanities in the Classroom: Introducing a New Editing Platform for Source Documents in Classics, in: *Digital Humanities*. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/digital-humanities-in-the-classroom-introducing-a-new-editing-platform-for-source-documents-in-classics/.

BERTI, M. *et al.*, 2009: Collecting Fragmentary Authors in a Digital Library, in: *JCDL '09: Proceedings of the 2009 joint international conference on Digital libraries*, New York, 259-262. http://hdl.handle.net/10427/70401.

BLACKWELL, C. & T. R. MARTIN, 2009: Technology, Collaboration, and Undergraduate Research, in: *Changing the Center of Gravity*, Vol. 3 No. 1.

BODARD, G., 2009: Digital Classicist: Re-use of Open Source and Open Access. Publications in Ancient Studies, in: *Digital Humanities 2009 Conference Abstracts*, 2. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf.

BODARD, G. & J. GARCÉS, 2009: Open Source Critical Editions: A Rationale, in: DEEGAN, M. & K. SUTHERLAND (eds.), *Text Editing, Print and the Digital World*, Farnham, Surrey, 83-98.

BOSCHETTI, F., 2007: Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses Onto Reference Text, in: *CL 2007: Proceedings of the Corpus Linguistics Conference*, Birmingham. http://ucrel.lancs.ac.uk/publications/CL2007/paper/150_Paper.pdf.

BOSCHETTI, F. *et al.*, 2009: Improving OCR Accuracy for Classical Critical Editions, in: AGOSTI, M. *et al.* (ed.), *Research and Advanced Technology for Digital Libraries*, Lecture notes in computer science 5714, Berlin / Heidelberg, Chapter 17, 156-167. http://hdl.handle.net/10427/70402.

BÜCHLER, M. *et al.*, 2010: Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts, in: *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1(2). http://letterpress.uchicago.edu/index.php/jdhcs/article/viewArticle/60.

BÜCHLER, M. *et al.*, 2012: Increasing Recall for Text Re-use in Historical Documents to Support Research in the Humanities Theory and Practice of Digital Libraries, in: *Theory and Practice of Digital Libraries (TPDL)*, Lecture Notes in Computer Science 7489, Berlin / Heidelberg, Chapter 11, 95-100.

CAUSER, T. *et al.*, 2012: Transcription Maximized; Expense Minimized? Crowdsourcing and Editing The Collected Works of Jeremy Bentham, in: *Literary and Linguistic Computing* 27, 119-137.

CAYLESS, H. A., 2008: Linking Page Images to Transcriptions with SVG, in: *Balisage: The Markup Conference 2008*, 12-15. http://www.balisage.net/Proceedings/vol1/html/Cayless01/BalisageVol1-Cayless01.html.

CAYLESS, H. A., 2010: Ktêma es aiei: Digital Permanence from an Ancient Perspective, in: BODARD, G. & S. MAHONY (eds.), *Digital Research in the Study of Classical Antiquity*, Burlington, 139-150. http://philomousos.com/papers/Cayless_DRSCA.pdf.

CAYLESS, H. A. *et al.*, 2009: Epigraphy in 2017, in: *Digital Humanities Quarterly* 3. http://www.digitalhumanities.org/dhq/vol/3/1/000030.html.

CISNE, J. L. *et al.*, 2010: Mathematical Philology: Entropy Information in Refining Classical Texts' Reconstruction, and Early Philologists' Anticipation of Information Theory, in: *PloS one* 5: e8661 + . http://dx.doi.org/10.1371/journal.pone.0008661.

CIULA, A., 2009: The Palaeographical Method Under the Light of a Digital Approach, in: *Kodikologie und Paläographie im digitalen Zeitalter-Codicology and Palaeography in the Digital Age*. Norderstedt, 219-237. http://kups.ub.uni-koeln.de/volltexte/2009/2971/.

CLEMENT, T., 2012: Methodologies In The Digital Humanities For Analyzing Aural Patterns In Texts, in: *Proceedings of the 2012 iConference*, New York, 287-293.

CLOUGH, P. D. *et al.*, 2009: Extending Domain-Specific Resources to Enable Semantic Access to Cultural Heritage Data, in: *Journal of Digital Information* 10.

http://journals.tdl.org/jodi/article/view/698.

CRANE, G. *et al.*, 2012: Student Researchers, Citizen Scholars And The Trillion Word Library, in: *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, New York, 213-222. http://hdl.handle.net/10427/75559.

DIEKEMA, A. R., 2012: Multilinguality in the Digital Library: A Review, in: *The Electronic Library* 30, 165-181.

DOOLEY, P. L., 2010: Wikipedia and the Two-Faced Professoriate, in: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, New York, 1-2.

DUÉ, C. & M. EBBOTT, 2009: Digital Criticism: Editorial Standards for the Homer Multitext, in: *Digital Humanities Quarterly* 3. http://www.digitalhumanities.org/dhq/vol/3/1/000029.html#.

FISCHER, A. *et al.*, 2011: HMM-Based Alignment of Inaccurate Transcriptions for Historical Documents, in: *International Conference on Document Analysis and Recognition (ICDAR), 2011*, Piscataway, NJ, 53-57.

GIBBS, F. W., 2011: New Textual Traditions from Community Transcription, in: *Digital Medievalist* 7. http://www.digitalmedievalist.org/journal/7/gibbs/.

GRAHAM, S., 2012: The Wikiblitz: A Wikipedia Editing Assignment in a First Year Undergraduate Class, in: DOUGHERTY, J. & K. NAWROTZKI (eds.), *Writing History in the Digital Age*, Forthcoming from the University of Michigan Press, web-book edition. http://WritingHistory.trincoll.edu.

GRAHAM, S. & G. RUFFINI, 2007: Network Analysis and Greco-Roman Prosopography, in: Keats-Rohan, K. S. B. (ed.), *Prosopography Approaches and Applications: A Handbook*, Oxford, 325-336.

HUNTER, J. & A. GERBER, 2012: Towards Annotopia—Enabling the Semantic Interoperability of Web-Based Annotations, in: *Future Internet* 4, 788-806.

IFLA 1998: *Functional Requirements for Bibliographic Records: Final Report*, UBCIM Publications N.S. 19, München. http://www.ifla.org/VII/s13/frbr/frbr.pdf.

ISAKSEN, L. *et al.*, 2012: GAP: A NeoGeo Approach to Classical Resources, in: *Leonardo* 45, 82-83.

KÜSTER, M. W. *et al.*, 2011: TextGrid Provenance Tools for Digital Humanities Ecosystems, in: *Digital Ecosystems and Technologies Conference (DEST), 2011 Proceedings of the 5th IEEE International Conference on*, IEEE May 31 – June 3 2011, 317-323.

LÜDELING, A. & A. ZELDES, 2009: Three Views on Corpora: Corpus Linguistics, Literary Computing, and Computational Linguistics, in: *Jahrbuch für Computerphilologie* 9, 149-178. http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html.

MCCARTY, W. (ed.), 2010: *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, Cambridge. http://www.openbookpublishers.com/reader/64.

MICHEL, J.-B. *et al.*, 2011: Quantitative Analysis of Culture Using Millions of Digitized Books, in: *Science* 331, 176-182.

MIHALCEA, R. & M. SIMARD, 2005: Parallel Texts, in: *Natural Language Engineering* 11, 239-246.

MIMNO, D. *et al.*, 2005: Hierarchical Catalog Records Implementing a FRBR Catalog, in: *D-Lib Magazine*. http://www.dlib.org/dlib/october05/crane/10crane.html.

MONELLA, P., 2008: Towards a Digital Model to Edit the Different Paratextuality Levels within a Textual Tradition, in: *Digital Medievalist*. http://www.digitalmedievalist.org/journal/4/monella/.

NAGY, G., 2010: Homer Multitext project, in: MCGANN, J. *et al.* (ed.), *Online Humanities Scholarship: The Shape of Things to Come. Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia March 26-28, 2010*, Houston, 87-112. http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=4087.

O'DONNELL, D. P., 2009: Back to the Future: What Digital Editors Can Learn From Print Editorial Practice, in: *Literary and Linguistic Computing* 24, 113-125.

ODIJK, D. *et al.*, 2012: Semantic Document Selection Theory and Practice of Digital Libraries, in: *Theory and Practice of Digital Libraries (TPDL 2012)*, Lecture Notes in Computer Science 7489, Berlin / Heidelberg 2012, Chapter 24, 215-221.

PEURSEN, W. T. (ed.) *et al.*, 2010: *Text Comparison and Digital Creativity: The Production of Presence and Meaning in Digital Text scholarship*, Leiden [u. a.].

PIERAZZO, E., 2011: A Rationale of Digital Documentary Editions, in: *Literary and Linguistic Computing* 26, 463-477.

PIOTROWSKI, M., 2010: Leveraging Back-Of-The-Book Indices To Enable Spatial Browsing Of A Historical Document Collection, in: *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, New York, 1-2.

PIOTROWSKI, M., 2012: Natural Language Processing for Historical Texts, in: *Synthesis Lectures on Human Language Technologies* 5, 1-157. http://dx.doi.org/10.2200/S00436ED1V01Y201207HLT017.

PORTER, D. *et al.*, 2009: Text-Image Linking Environment (TILE), in: *Digital Humanities 2009: Conference Abstracts*, 388-390. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf.

ROBINSON, P., 2010a: Editing Without Walls, in: *Literature Compass* 7, 57-61.

ROBINSON, P., 2010b: Electronic Editions for Everyone, in: *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, 145-163.

ROMANELLO, M. *et al.*, 2009: When Printed Hypertexts Go Digital: Information Extraction From The Parsing Of Indices, in: *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, New York, 357-358.

ROSENZWEIG, R., 2006: Can History be Open Source: Wikipedia and the Future of the Past? in: *Journal of American History* 93, 117-146.

SALERNO, E. *et al.*, 2007: Digital Image Analysis to Enhance Underwritten Text in the Archimedes Palimpsest, in: *International Journal on Document Analysis and Recognition* 9, 79-87.

SCHMIDT, D. & R. COLOMB, 2009: A Data Structure for Representing Multi-Version Texts Online, in: *International Journal of Human-Computer Studies* 67, 497-514.

SCHMITZ, P., 2009: Using Natural Language Processing and Social Network Analysis to Study Ancient Babylonian Society, in: *UC Berkeley iNews*.
http://inews.berkeley.edu/articles/Spring2009/BPS.

SIEMENS, R. *et al.*, 2012: Toward Modeling The Social Edition: An Approach To Understanding The Electronic Scholarly Edition In The Context Of New And Emerging Social Media, in: *Literary and Linguistic Computing* 27, 445-461.

SMITH, D. A. *et al.*, 2001: Management of XML Documents in an Integrated Digital Library.
http://xml.coverpages.org/perseus-hopperExtreme2000.pdf.

SMITH, D. A. *et al.*, 2011: Mining Relational Structure from Millions of Books: Position Paper, in: *Proceedings of the 4th ACM workshop on Online books, Complementary Social Media and Crowdsourcing*, New York, 49-54.

SMITH, D. N., 2009: Citation in Classical Studies, in: *Digital Humanities Quarterly* 3.
http://www.digitalhumanities.org/dhq/vol/003/1/000028.html#

SMITH, D. N., 2010: Digital Infrastructure and the Homer Multitext Project, in: BODARD, G. & S. MAHONY (eds.), *Digital Research in the Study of Classical Antiquity*, Burlington, 121-137.

SOSIN, J., 2010: Digital Papyrology, in: *26th Congress of the International Association of Papyrologists (19 August 2010)*.
http://www.stoa.org/archives/1263.

SPORLEDER, C., 2010: Natural Language Processing for Cultural Heritage Domains, in: *Language and Linguistics Compass* 4, 750-768.

STEWART, G. *et al.*, 2007: A New Generation Of Textual Corpora: Mining Corpora From Very Large Collections, in: *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, New York, 356-365.
http://hdl.handle.net/10427/14853.

TOBIN, R. *et al.*, 2008: Named Entity Recognition for Digitised Historical Texts, in: *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC '08)*.
http://www.ltg.ed.ac.uk/np/publications/ltg/papers/bopcris-lrec.pdf.

UNSWORTH, J., 2011: Computational Work with Very Large Text Collections, in: *Journal of the Text Coding Initiative* 1. http://jtei.revues.org/215.

VERTAN, C., 2010: Towards the Integration of Language Tools Within Historical Digital Libraries, in: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC '10)*, European Language Resources Association (ELRA). http://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/811_Paper.pdf.

WHALLEY, B., 2012: Wikipedia: Reflections on Use and Acceptance in Academic Environments, in: *Ariadne* 69. http://www.ariadne.ac.uk/issue69/whalley.

YALNIZ, I. Z. & R. MANMATHA, 2012: Finding Translations in Scanned Book Collections, in: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, New York, 465-474.

ZHANG, Z. *et al.*, 2010: A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality, in: CIMIANO, P. & H. PINTO (eds.), *Knowledge Engineering and Management by the Masses*, Lecture Notes in Computer Science 6317, Berlin / Heidelberg, Chapter 21, 301-315.