DHQ: Digital Humanities Quarterly
Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure
2009
Volume 3 Number 1

Classics in the Million Book Library

Gregory Crane <gregory_dot_crane_at_tufts_dot_edu>, Tufts University
Alison Babeu <Alison_dot_Jones_at_tufts_dot_edu>, Tufts University
David Bamman <David_dot_Bamman_at_tufts_dot_edu>, Tufts University
Thomas Breuel <tmb_at_informatik_dot_uni-kl_dot_de>, Technical University of Kaiserslautern
Lisa Cerrato <lcerrato_at_perseus_dot_tufts_dot_edu>, Tufts University
Daniel Deckers <daniel_dot_deckers_at_uni-hamburg_dot_de>, Hamburg University
Anke Lüdeling <anke_dot_luedeling_at_rz_dot_hu-berlin_dot_de>, Humboldt-University, Berlin
David Mimno <mimno_at_cs_dot_umass_dot_edu>, University of Massachusetts, Amherst
Rashmi Singhal <rashmi_dot_s_dot_singhal_at_gmail_dot_com>, Tufts University
David A. Smith <dasmiq_at_gmail_dot_com>, University of Massachusetts, Amherst
Amir Zeldes <amir_dot_zeldes_at_rz_dot_hu-berlin_dot_de>, Humboldt-University, Berlin

Abstract

In October 2008, Google announced a settlement that will provide access to seven million scanned books while the number of books freely available under an open license from the Internet Archive exceeded one million. The collections and services that classicists have created over the past generation place them in a strategic position to exploit the potential of these collections. This paper concludes with research topics relevant to all humanists on converting page images to text, one language to another, and raw text into machine actionable data.

Introduction

> In a long span of time it is possible to see many things that you do not want to, and to suffer them, too. I set the limit of a man's life at seventy years; these seventy years have twenty-five thousand, two hundred days, leaving out the intercalary month. But if you make every other year longer by one month, so that the seasons agree opportunely, then there are thirty-five intercalary months during the seventy years, and from these months there are one thousand fifty days. Out of all these days in the seventy years, all twenty-six thousand, two hundred and fifty of them, not one brings anything at all like another. (Herodotus, *Histories* 1.32, tr. Godley)

In the first book of Herodotus' *Histories*, the Athenian statesman Solon calculates that an average human life of seventy years contains roughly 25,000 days. If we could read a book every day of our lives, it would take a thousand years — almost forty generations — to work our way through one million books. It would take 10,000 years or four hundred generations to work through the ten million or so unique books that the original Google library partners contained in their collections.[1] On October 28, 2008, Google announced an agreement with publishers that would allow libraries to provide, largely under a subscription basis, access through Google book search to some seven million books, including copyrighted materials.[2] Google is providing immense scale but the scholarly significance is not so great as it might be: there is at present no way to understand what subset of the world's knowledge that seven million volumes represents. Even if there were, scholars have no way of understanding in more than the most general way how the services that extract information from that collection work — what is missed? What biases are embedded in the system? Scholarship depends upon transparency, and we must be careful that we do not, in pursuing our immediate research projects, compromise our fundamental commitment to transparency. 1

A day before the momentous Google announcement, another and arguably even more important milestone was crossed. On October 27, 2008, the number of books available from the Internet Archive exceeded 1,000,000. While the million books is only a fraction of the size of the seven million that Google boast, the million books available from the Internet Archive are freely downloadable — anyone can analyze them and publish the results. The collection available from the Internet Archive provides the foundation for transparent services and, even more important, transparent discourse. Open source services, carefully evaluated and publicly documented, applied to open content, freely downloadable by anyone without restrictions, embody the goals of scholarly and scientific practice. 2

A million books alone would support a book-of-the-day club for almost 3000 years. Thus, even if we restrict ourselves to digitized printed books available for public download in a single location, the scale of content available has already passed that which any single human mind could comprehend. As a physical collection, a million books is hardly remarkable. As a store of knowledge for human analysis, the scale of 1,000,000 books has already passed human scale and is as abstract as the distance between galaxies or the number of insects in the world. Only machines can process the collections to which we already in late 2008 have access. What can we do with a million books with the tools now at our disposal and which we could build? What are the research questions that emergent huge collections raise for the historians, literary critics, and other humanists who study their contents and for the computer and information scientists who develop methods with which to process digital information in general? 3

This paper summarizes research, supported by a grant from the Mellon Foundation, into the challenges and opportunities confronting the humanities in general and classical studies in particular as we shift from small, carefully edited and curated digital collections to very large, industrially produced collections that, in their fullest instantiation, aim to subsume whole libraries. We view classical studies as a special case with the more general question that we have termed “what do you do with a million books?” The authors of this paper come from Europe as well as North America, from classical philology, computer science, corpus linguistics and library and information science, but many others contributed substantively to the work reported here during workshops that we conducted between November 2006 and March 2008.[3] Previous publications have addressed some of the general issues that the humanities as a whole face.[4] This paper explores the particular case of classics within the million-book library. 4

When we began this study in 2006, we planned to focus upon materials related to the Greco-Roman world, early modern Europe, and the 19th century United States so that we could examine the varying problems and opportunities associated with print materials relating to each. As our work progressed, the advantages of focusing on classical studies became progressively clear. Classics includes not only the Greco-Roman world but the subsequent scholarship about the Greco-Roman world and a vast body of material written in Latin on virtually every subject. Beginning with early printed editions in the 1470s and continuing through the present, Classical scholarship brings us to every corner of Europe, North America, and the Middle East. 5

Classical studies do not, of course, touch the same audiences as Shakespeare or the American Civil War, and there are not nearly so many Classicists as there are experts in English Literature and American History, but Classics has produced the largest coherent community of scholars engaged with the digital infrastructure for their field. Classical studies became a logical focus for our work: if we could understand how to build a comprehensive collection of classical scholarship from the beginning of print culture to the present, we would know how to work with centuries of print publications on every aspect of human society and in every discipline and from every corner of Europe and North America. 6

This paper begins by stressing that we have moved beyond islands of digital content in a vast sea of print. Where our first generation collections were autonomous, carefully curated, discipline-specific islands, we now see emerging a world where we dynamically generate collections of heterogeneous materials from vast and constantly expanding digital libraries over which no individual discipline or project exercises control. We cannot thus rely upon a centralized editorial structure to guarantee for us the consistency of what we find. We need tools that can help us assess how representative our automatically extracted corpus is (e.g., what biases are there in the distribution of Latin texts available for searching?) and the accuracy of our analytical tools (e.g., the precision and recall of named entity systems that search for Salamis in Cyprus vs. the Salamis near Athens, the error rates in Latin text that Optical Character Recognition 7

(OCR) extracts from various editions printed in different places and times).

Our discussion then moves to the services that humanists need to exploit very large collections. These include not only advanced services for information extraction, multilingual technologies, and visualization but simple access to the scanned page images with which to support domain-optimized document analysis. These services require the rise of a new, fourth generation of digital corpora. Our first digital corpora included accurate transcriptions with markup of surface features (e.g., we simply indicate that a word is in italics). A second generation began to add semantic markup (e.g., a phrase is in italics because it is the title of a work or a Latin quotation). The third generation created much larger collections by shifting the focus of manual labor from carefully edited typing to industrial scanning of page images. We need fourth-generation collections that can seamlessly integrate image-books, accurate transcriptions, and machine actionable knowledge in various formats. 8

These fourth generation collections are a qualitatively new phenomenon. They allow us to design collections that are not only more comprehensive but more diverse than we could ever produce in print culture. These collections are unbounded and can include not only texts but every category of data about their subjects — high resolution images, three-dimensional models, geographic data sets, and anything that we can represent in digital form. Even if we restrict ourselves to linguistic data, fourth generation collections are a qualitative advance over print: we can include not only images of neatly printed modern books but non-print representations of language such as three dimensional models of words engraved on stone and digital sound recordings. 9

For classics, the most important such project is what we have termed the apographeme of classical Greek and Latin — an analogy to the *genome*, representing the complete record of all Greek and Latin textual knowledge preserved from antiquity, ultimately including every inscription, papyrus, graffito, manuscript, printed edition and any writing bearing medium. This *apographeme* constitutes a superset of the capabilities and data that we inherit from print culture but it is a qualitatively different intellectual space. In the mature *apographeme*, every canonical text is a multitext, with dynamic editions linked to visual representations of the manuscripts, inscriptions, papyri and other sources. In the mature *apographreme*, each source is linked to the background data that we need to understand it — a transcription, information about the particular type of Greek or Latin script and its abbreviations, about the monastery, print shop or Egyptian village that produced it, etc. 10

From Curated Collections to Dynamic Corpora

The methods whereby we assemble digital content are very different now from those that were available when, a generation ago, the first pioneers began designing digital tools for the humanities. In the 1970s and 1980s, most scholars considered digital resources — insofar as they considered them at all — as instruments with which to navigate a paper sea of information. The *Thesaurus Linguae Graecae (TLG)* and the *Dictionary of Classical Bibliography (DCB)*, two of the pioneering efforts within classical studies, were, in effect, indices and depended upon the ability to pay human beings to read and to type. The *Thesaurus Linguae Graecae* began work in 1972 and can boast in 2008 almost 100,000,000 words of cleanly transcribed Greek text.[5] 11

By the opening of the twenty-first century, of course, the technologies available began to open up very different approaches. Between 2001 and the end of 2005, one of the authors of this paper, Gregory Crane, developed a 55,000,000 word collection of 19th century American English.[6] He personally scanned 400 volumes, applied OCR to the scanned page images, applied automated post-processing and shipped the results to a data entry firm. A handful of reference works with complicated formatting required traditional manual data entry but for the vast majority of this corpus the OCR-generated text provided a solid foundation and avoided the need for typing. The contractor checked for errors and added basic structural markup, tagging such elements as footnotes, quotations, figures at a cost of under $100,000 for 500,000,000 characters or about $200/book. The corpus was not an end in itself but rather an instrument with which to study problems of automatically analyzing large collections. Most of Crane’s effort went to the production of a system that could automatically identify people, places, organizations, and other named entities in unstructured text.[7] That research became the foundation for a project entitled “Scalable Named Entity Services for Classical Studies” that would, with support from the National Endowment for the Humanities (NEH) and the Institute for Museum 12 and Library Services, adapt this system for use with documents from classical studies.

13 The situation has changed even further in the past several years. The primary medium for human intellectual life is now irrevocably digital. The most heavily funded academic disciplines use paper to print digital resources on demand. Print-only resources are now archival materials. Consider the following developments that have intensified since Crane began work on the Perseus American Collection in 2001:

1. *Massive scanning*. In December 2004, Google seized the initiative to create a vast library of scanned books, with text generated by OCR, but the library community has the resources to convert its print holdings into digital form: the 123 North American libraries who belong to the Association of Research Libraries spent more than a billion dollars on their collections in the 2005-06 academic year.[8] Of course, most libraries will claim that they are under-funded and cannot maintain their existing collections, much less consider a major new initiative. Some of us are old enough to remember hearing that the costs of print collections would never allow for libraries to make digital materials accessible.
2. *Scanning on demand*. The OCA has created an infrastructure whereby individuals could, by 2007, select particular books for scanning and then inclusion in the larger OCA collection. The quality is high and the cost is low: $.10US per page plus handling costs of $5US per book — about $40 for a standard book with about 300 pages. It costs about the same to create a high resolution scan that anyone attached to the internet can scan than it does to buy a single printed book ($52 in 2005-06). Support from the Mellon Foundation has allowed the Cybereditions Project at Tufts University to begin creating within the OCA an open source library of Greek and Latin that will contain at least one text or fragment of every major surviving Greek and Latin author and a range of reference materials, commentaries and core publications.
3. *The growth of open access and open source licensing*. In 2008, open access publication became the dominant model for academic publication. The US National Institutes of Health (NIH) have established a public access policy. As of April 2008, this policy “requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central upon acceptance for publication.” [9] The policy further requires scientists funded by the NIH to include in their papers citations to the open access copies of previous publications in the open access PubMed Central web site. The NIH provides more than 22 billion dollars in funding for medical research.[10] Publishers in the most heavily funded research area in all of academia must now develop business models that assume open access that precedes publication. The US NEH, by contrast, requested just under $145 million dollars in funding for 2009 — less than 1% as much as the NIH invests.
4. *Improved OCR*. Traditionally, classical Greek has been a huge barrier for classicists — there was no useful OCR. All classical Greek required manual data entry and such specialized work usually cost much more than data entry for English. By early 2008, Google had begun to generate initial OCR text from page images of classical Greek. The software is evidently based on OCR designed for modern Greek and contains errors, but clever search software could ameliorate this problem. If classicists have access to the scanned page images and can optimize OCR software for classical Greek, we can achieve character level accuracy (99.94%) comparable to the standard quality for manual data entry (99.95%). In a preliminary analysis of printed scholarly editions, we found that 13% of the unique Greek words on a page, on the average, only appeared in the textual notes. Restricting our analysis to older volumes from the Loeb Classical Library (which traditionally provided a minimal number of textual variants), we found that only 97% of the unique Greek words on a given page appeared in the main text. Thus, collections that contain perfect transcriptions of the reconstructed text but no textual notes offer at most 97% — and, if we use fuller editions, 86% — of the relevant data. The worst OCR error that we measured (98%) matches the overall recall rate of perfect transcriptions of text alone.[11] Once we enter multiple editions of the same text, we can begin using each scanned edition to identify OCR errors and intentional textual variants.
5. *A new generation of text mining and quantitative analysis*. The DCB contains 600,000 bibliographic entries from 1949 to 2005 and adds 12,500 new items each year.[12] By contrast, the CiteSeer system, upon which computer scientists depend, was developed more than a decade ago in 1997 at the NEC Research Institute

in Princeton, NJ, and offers an automatically generated index of 767,000 publications, including automatically extracted bibliographic citations.[13] Research continues and new generations of automated bibliographic systems, based on the automated analysis of on-line publications, have begun to appear. The Rexa System, developed at the University of Massachusetts, had assembled a collection of almost 1,000,000 publications in 2005 [Mimno 2007]. David Mimno, one of the authors of this paper, is a member of that research group and has support from the Cybereditions project at Tufts University to begin in 2008-09 applying that research to publications from classics.

We might summarize the current situation as follows: Google has begun creating on-line a digital collection that would be more comprehensive than the greatest university libraries ever produced — and the university libraries themselves control the resources needed to do the job were Google to falter: our retrospective collections are being digitized. The OCA has created a public, scalable infrastructure whereby we can, in fact, build high quality collections within the existing library infrastructure: if massive projects miss anything, smaller efforts can fill in the gaps and create curated collections. The US government, under a conservative, pro-business administration, has made the most profitable monopolies on which publishers had depended illegal and declared open access a condition of its most generous funding agency: the richest publishers must learn to make money under open access. Advances in OCR technology have made it possible for scholars in fields such as Classics to generate very serviceable searchable text for non-standard character sets such as Greek: once we scan editions, we can more comprehensively search primary sources and, for the first time, secondary sources that quote Greek. A new generation of text mining can provide new methods with which to trace ideas and research topics that appear in millions of publications: we can design bibliographic databases that incorporate features of particular interest to classics (e.g., the ability to determine whether “Th. 1.38” designates line 38 of Theocritus’ first *Idyll*, or chapter 38 of book 1 of Thucydides) with the common features of academic publication (e.g., footnotes and bibliographic citations).

New services feasible in such an environment include 14

- **Multitexts:** Scholars have grown accustomed to finding whatever single edition a particular collection has chosen to collect. In large digital collections, we can begin to collate and analyze generations of scholarly editions, generating dynamically produced diagrams to illustrate the relationships between editions over time. We can begin to see immediately how and where each edition varies from every other published edition.
- **Chronologically deeper corpora:** We can locate Greek and Latin passages that appear anywhere in the library, not just in those publications classicists are accustomed to reading. We can identify and analyze quotations of earlier authors as these appear embedded in texts of various genres.
- **New forms of textual bibliographic research:** We can automatically identify key words and phrases within scholarship, cluster and classify existing publications, generate indices of particular people (e.g., Antonius the triumvir vs. one of the many other figures of that name, Salamis on Cyprus vs. the Salamis near Athens). Such searches can go beyond the traditional disciplinary boundaries, allowing students of Thucydides, for example, to analyze publications from international relations and political philosophy as well as classics.

In this world, we need to recognize that we are — as indeed classicists have always in large measure been — corpus linguists. All classicists can articulate, in some measure, the relationship between the texts that survive and the subject that we are studying. If we work with Sophocles, we know that only seven plays survive and we have only fragments and even titles for the rest. If we study Alexander the Great, we must first understand the fact that our most comprehensive surviving Greek sources were composed centuries later and depend upon earlier histories that are now lost.

Consider three topics that we might pursue in a very large collection: the usage of a Latin word over two thousand years (e.g., *oratio*, which can, for example, in different contexts designate a speech or a prayer), the reception of Euclid’s *Elements*, and the reputation of Alexander the Great. In each case, we can assemble far more information that we could ever collect in print culture. The next section will touch upon some of the services with which we can make the sprawling 15 corpora relevant to each of these topics intellectually accessible. But even before we begin our analysis, we need to understand the limitations of the corpus that we have assembled:

- **How representative is the corpus?** Is all of a given corpus available on-line? (e.g., have all the published volumes of a series been scanned?) Can we estimate the percentage of the corpus that survives? (e.g., what percentage of Sophocles' do the seven plays and other fragments constitute?) What biases are inherent in our data? (e.g., do we have any accounts produced by women or by members of every national/ethnic group involved in a topic? If we find 100,000 instances of the Latin word *oratio*, what are the periods, genres, locations, and (in the case of later Latin) original languages of the authors?)[14] And, are there correlations between these parameters? These may in fact be automatically discernable from the data, even if the human eye doesn't notice them in the forest of data.
- **How accurate are the digital surrogates for each object?** We may have a satisfactory corpus of print materials but these materials may yield very different results to automated services such as OCR, named entity identification, cross-language information retrieval, etc. Readers of Jeff Rydberg-Cox's contribution to this collection will realize that OCR software will, at least in the immediate future, extract much less usable text from early modern printed editions than from editions printed in the early twentieth century. We need automatically generated metrics for the precision and accuracy of each automated process on which we depend.

Services for the humanities in very large collections

If humanists are to exploit large new collections to their fullest, they need, as a minimum, the following services: 16

- **Access to images of the physical sources:** This includes access to particular copies of a document, any pagination or naming scheme with which to address the individual pages, and a coordinate system with which to describe regions of interest on a given page. Many born-digital publications do not provide such access — logical "page 12" of a report (as printed as a page number) may physically be page 21 of the PDF document (after adjusting for front matter, a table of contents etc.). Coordinate systems must have sufficient abstraction so that they can address relationships of the printed page even if the paper has been cropped or varies from one printing to another: coordinates for one First Folio should be useful with others. [15]
- **Access to transcriptional data:** At the least, we need to be able to analyze the words and symbols that are encoded on the physical page.[16] The rough "bag-of-words" approach, where systems ignore the location of words on the page and even their word order, has proven remarkably useful. This level of service is fundamental to everything that follows. Conventional OCR software has traditionally provided no useful data from historical writing systems such as classical Greek. Latin is much more tractable but OCR software expecting English will introduce errors (e.g., converting *t-u-m*, Latin "then," into English *t-u-r-n*). Even earlier books with clear print will contain features that confuse contemporary OCR (e.g., the long 's' which looks like an 'f,' such that words such as l-e-s-t become l-e-f-t).[17]
- **Access to basic areas of a page such as header, main text, notes, marginalia:** Even transcription depends upon basic page layout if it is to achieve high accuracy: we cannot transcribe individual words unless we can automatically resolve hyphenization and this in turn implies that we can distinguish multi- from single column text, footnotes, headers marginalia, etc. from main text, etc.[18] We need, however, to recognize basic scholarly document layouts: thus, we should be able to search for either the reconstructed notes or the textual notes at the bottom of the page. This stage corresponds roughly to WYSIWYG markup. At this stage, the system can distinguish the main text from the notes in Figure 5 and Figure 10 but it does not recognize that one set of notes are commentary and the other constitute an apparatus criticus.
- **Access to visually labeled structures within the text:** Explicit labeling in this case includes headwords of dictionaries and encyclopedias and canonical citations such as book/chapter/verse/line. These structures draw upon typographical conventions: e.g., bold and indenting to show headwords, numbers in the margins or embedded in the text with brackets to illustrate citations. This stage would recognize where index entries

begin along with their headwords and easily recognized citations. This stage corresponds to semantically meaningful structural markup, e.g., descriptive structures about the text.[19] At this stage, the system recognizes that the notes in Figure 5 are a commentary and contain comments on *agro vectigali, cum et maxima … ageretur*, and *tibique … indicium* within the text above.

- **Access to knowledge dynamically generated from analysis of explicitly labeled knowledge:** This process can begin with very coarse analysis: if we recognize when various encyclopedia articles describing several dozen figures named Antonius or Alexander begin and end, then we can analyze the vocabulary of each article to begin deciding which Alexander is meant in running text. This stage includes the lemmatization and morphological analysis to support the lookups and searches familiar to classicists for more than a decade (e.g., query *fecisset* and learn that it is the pluperfect subjunctive of *facio*, “to do, make”; query *facio* and retrieve inflected forms such as *fecisset*).[20] We also need at this stage translation services (e.g., a service that determines whether a given instance of the Latin word *oratio* more likely corresponds to “oration,” “prayer” or some other usage). At this point, knowledge based services augment general text mining (e.g., being able to cluster usages of the dictionary entry, *facio*, as a whole — or of *facio* as it is used in the subjunctive etc. — rather than treating each form of *facio* as a separate entity.)[21]
- **Access to linguistically labeled, machine actionable knowledge:** This overlaps with the analysis of visual structures but implies a greater emphasis on the analysis of natural language, e.g., “Y, son of Z,” “perf. feci” → the perfect stem is fec-., “b. July 2, 1887” → the subject of this encyclopedia was born in 1887 and any references to people by the same name that predate 1887 cannot describe this person,” etc. This stage corresponds to encoding information for particular ontologies, i.e., prescriptive structures separate from the text.[22] At this point, we should be able to pose queries such as “encyclopedia entries for Thucydides who is son_of Olorus or has_occupation historian, etc.”, “dictionary word senses is_cited_in Homer or has_voice passive;” “Book 1, lines 11-21 from all translations_of Homer_Iliad that have_language German.”

Techniques exist to address all of the services outlined above. Computer scientists strive for completely general approaches and are willing to accept error rates as a cost to achieve the benefit of scalability. Traditional humanists by contrast manually analyze and, where they feel it necessary, justify the results (i.e., results that may be controversial but for which experts can make reasonable arguments) and are willing to accept labor as a cost to achieve a level of transparency. The grand challenge lies in integrating these two sources of energy: scholars need to be able to build on the results of automated processes but automated processes need to be able to build on scholarly data as well.

Fourth-Generation Collections

We need collections that can support a core set of interlocking services. Core services such as morphological and syntactic analysis, citation identification, word sense disambiguation, word sense discovery, cross-language information retrieval, and named entity identification are, however, data-driven and, for optimal performance, require substantial amounts of carefully encoded knowledge and the largest possible bodies of unstructured data. To support these services, we need a new generation of collections. Within the humanities, we need a new, fourth generation of digital collections. 17

While classicists have digitized texts for a generation and accurate transcriptions exist for selected editions of almost every author, we do not have the developed, scalable, sustainable knowledge base with which to represent the core primary sources that have survived to us in textual form from Greco-Roman antiquity. 18

We have already touched upon the first generation of digital primary sources. Classicists still depend primarily upon the *Thesaurus Linguae Graecae* and *Packard Humanities Institute* collections of Greek and Latin texts, which follow designs from the 1970s. These first generation digital collections concentrated on accurate transcription of the reconstructed text with structural markup showing where works begin and end. They capture general page layout and approximate citation information: if a number in the margin of the original print edition indicates a line or section begins somewhere in the adjacent line, the human reader is left to determine where the break occurs. They do not contain any 19 of the introductory materials, back matter such as indices and appendices or any textual notes.

Second generation collections (such as those available within the Perseus Digital Library) also emphasize carefully produced transcriptions but include explicit semantic markup that follows the Text Encoding Initiative (TEI) Guidelines. These collections reflect the conditions of the late 1980s where image capture and storage remained expensive. They thus do not include page images of the original source texts and only occasionally include textual notes. Second generation collections may apply more sophisticated techniques to automate transcription and tagging but their design still assumes an expensive initial, centralized editing process with small fixes for residual errors after the initial production phase. 20

Third generation collections, popularized by projects such as the Making of America and JSTOR, emerged in the 1990s when storage costs had declined to the point where page images for large collections of books could be kept on-line.[23] Third generation systems minimize manual labor and emphasize automatic analysis of page images — especially the use of OCR software to generate searchable text. As OCR software increases in accuracy, texts can be rescanned and the searchable text can improve. First and second generation collections worked from the inside of the book outwards, focusing on subsets of printed books for digital conversion. Third generation collections by contrast, work from the outside of the book, starting with book-level library metadata that may be extended with analytical cataloguing for articles within books. 21

All of the features that characterize the fourth-generation have existed in one form or another — our group at the Perseus Project has been developing some aspects of this plan for more than twenty years. What distinguishes fourth generation collections is the integration of a small body of data, carefully curated and laboriously structured by semi-automated or even wholly manual methods with an arbitrarily large collection for which automated analysis alone is feasible. The semantically encoded data of second generation digital collections becomes the machine actionable reference rooms from which automated systems learn how to structure the vast third generation collections of page images: 22

- **Fourth-generation collections contain images of all source writings, whether these are on paper, stone or any other medium:** Like third generation collections, the Cybereditions project sets out to incorporate page images of all print originals. Our goal is to help classical scholars shift the center of gravity for textual scholarship to a networked, digital environment. Scholars should not have to consult paper originals of scanned print editions to see what was on the original page.
- **Fourth-generation collections manage legacy structures derived from physical books and pages but focus primarily upon logical structures that exist within and across pages and books:** Even when fourth-generation collections depend upon page images, they exploit legacy book-page citations but they are fundamentally oriented towards the underlying logical structures within the documents. A great deal of emphasis is placed upon page layout analysis so that we can isolate not only tables of contents, bibliographic references and indices but dictionary and encyclopedia entries and critical scholarly document types such as commentary notes and textual apparatus. Cross language information retrieval hunts for translations of primary sources. Alignment services align OCR generated text to XML editions of the same works with established structural metadata. Quotation identification services spot commentaries by recognizing sequences of quotations from the same text at the start of paragraphs.
- **Fourth-generation collections integrate XML transcriptions of original print data as these become available: All digital editions are, at the least, re-born digital:** The best work published so far cannot convert the elliptical and abbreviated conventions by which scholars represent textual data in print into machine actionable data — we cannot even reliably link the textual notes to the chunks of text which they cover, much less convert these notes into machine actionable formats so that we could automatically compare the readings from one MS against those of another. Fourth generation collections naturally integrate page images with XML representations of varying sophistication. XML representations may, like first generation collections, capture basic page layout and they may have advanced structural and basic semantic markup (e.g., careful tagging for each speaker in a play). They may encode no textual notes, textual notes as simple footnotes (free text associated with a point in the reconstructed text) or as fully

machine actionable variants (e.g., variants associated with spans of source text, such that we can, among other things, compare the text in various editions or witnesses).

- **Fourth-generation collections contain machine actionable reference materials:** Our digital collections should be tightly and automatically embedded in a growing web of machine actionable reference materials. If a new prosopography or lexicon appears, links should appear between its articles and references to the people or words in the primary sources. Commentaries should align themselves automatically to multiple editions of their subject work. To the extent possible, these links should bear human readable and machine actionable information: humans should be able to see from a link what the destination is about (e.g., “Thucydides the Historian” rather than “Thucydides-3,” ἀρχή-“empire” rather than “ἀρχή-sense2”). Equally important, these links should point to machine actionable information: a named entity system should be able to mine the entries in the biographical encyclopedia to distinguish Thucydides the Historian, Thucydides the mid-fifth century Athenian politician and various other people by that name; a word sense disambiguation system should be able to use the lexicon entries to find untagged instances where ἀρχή corresponds to “empire” or “beginning.” Editions should be self-collating — when a new edition of a text comes on-line, we should see immediately how it differs from its predecessors.
- **Fourth generation collections learn from themselves:** Even the simplest digital collection depends upon automated processes to generate text from page images or indices from text. Clustering and other text mining techniques discover meaning in unstructured textual data. Fourth-generation collections, however, can also learn from the machine actionable reference materials that they contain so that they apply increasingly more sophisticated analytical and visualization services to their content. In effect, they use a small body of structured data — training sets, machine actionable dictionaries, linguistic databases, encyclopedias and gazetteers with heuristics for classification to find structure within the much larger body of content for which only OCR-generated text and catalogue level metadata is available. In a fourth generation collection, structured documents are programs that services compile into machine actionable code: *Aeneid*, book 2, line 48 in a dozen different editions already on-line as image books with OCR generated text.
- **Fourth generation collections learn from their users:** Even third generation systems depend upon the ability of OCR software to classify markings into distinct letters and words. Fourth generation systems include an increasing number of classification systems such as named entity analysis, word sense disambiguation, syntactic analysis, morphological analysis, citation and quotation identification. Where there are simple decidable answers (e.g., to which Alexandria does a particular text refer?) we want users to be able to submit corrections. Where the answers are less well-defined (e.g., expert annotators do not agree on word sense assignment and some passages are simply open to multiple interpretation), we need to be able to manage multiple annotations. Human annotators need to be able to own their contributions and readers should be able to form conclusions about their confidence in individual contributors. Automated systems need to be able to make intelligent use of human annotation, determining how much weight to apply to various contributions, especially where these conflict. We therefore need a multi-layer system that can track contributions, by both humans and automated systems, through different versions of the same texts.
- **Fourth generation collections adapt themselves to their readers, both according to specific recommendations (customization) and by making inferences from observed user behavior (personalization):** Fourth-generation collections can process knowledge profiles that model the backgrounds of particular users: e.g., one user may be an expert in early Modern Italian, who has read extensively in Machiavelli, but only have a few semesters of classical Greek with which to read Thucydides and Plato. The fourth-generation collection can determine with tolerable accuracy what words in a new Italian or Greek text will be new and/or of interest, given the differing backgrounds but consistent research interests of the professor. At the same time, the system can infer from the reader’s behavior what other resources may be of interest.
- **Fourth-generation collections enable deep computation, with as many services applied to their content as possible:** No monolithic system can provide the best version of every advanced service upon

which scholarship depends. Google, for example, has a growing number of publications about ancient Greece but currently produces only limited searchable text from classical Greek. Different groups should be able to apply various systems for morphological and syntactic analysis, named entity identification, and various text mining and visualization techniques with minimal, if any, restrictions. These groups should include both commercial service providers as well as individual scholars and scholarly teams.

The Classical *Apographeme*

23 Fourth-generation collections allow us to design corpora that go far beyond limitations that we internalized in print culture. To describe comprehensive fourth-generation collections we use the term *apographeme*, derived from the Greek word for copy (*apographê*). The *apographeme* echoes the term genome because an *apographeme* contains, in its mature form, a complete record of every surviving linguistic source for a particular corpus. For classicists, an *apographeme* of Greek and Latin would contain representations of every written version of every piece of writing from Greco-Roman antiquity. This includes images of every page of every inscription, papyrus, graffito, manuscript, and printed edition — the entire surviving record of the linguistic output for classical Greece and Rome — and the knowledge base whereby machines can intelligently process and humans productively decipher, insofar as existing knowledge and probing intellect can, every written word in every witness. In a library grounded on images of writing, there is no fundamental reason not to integrate, at the base level, images of writing from all surfaces. Inscriptions, papyri, and manuscripts may not be suitable fodder with which OCR software can generate useful text, but neither Google nor OCA can produce much useable output for even the best printed classical Greek and little, if any, useable text from early modern books.

24 The Cybereditions project at Tufts has begun preliminary work for this massive task, focusing on the texts that have survived from Greco-Roman antiquity through manuscript tradition. These literary texts are, however, designed from the start to become part of a larger collection that will include documentary materials that survive on stone and papyrus (see Hugh Cayless's article in this collection) as well as manuscripts (see Casey Dué and Mary Ebbott's article in this collection). While developing the underlying bibliography is a major and on-going task of the Cybereditions project, we currently estimate that this apographeme would contain the following (because page images would be the first stage of collection, we use “books” as a rough initial unit of measure). Major work for the Cybereditions project will be (1) to complete a first cut of the bibliography below, (2) to begin creating the apographeme, with particular attention to the published editions, and (3) to make progress on the services that will convert these image pages into machine actionable data, with particular attention to the problem of high accuracy OCR for Greek and Latin.

25 We will not be able to create a comprehensive *apographeme* for classical Greek and Latin for many years but we can establish a solid foundation from that portion of the *apographeme* represented by texts that have survived in manuscript tradition. The figures associated with each element reflect very preliminary estimates for broad, illustrative coverage sufficient to model a more mature system that can evolve over time.

- **c. 500 “book-length” authors/collections.** Hundreds and thousands of ancient Greek and Latin authors survive as names or with a small number of fragments preserved in quotations of later authors or on papyrus. F. W. Hall’s *Handbook to Classical Texts* lists 133 entries in its survey of the “chief classical writers” — including portmanteau works with many authors (e.g., the *Greek Anthology*) and authors with very large corpora (e.g., Aristotle and Cicero).[24] The Loeb Classical Library does not contain comprehensive editions for massive authors such as Galen or the early Church Fathers but its 500 volumes contain Greek and Latin texts as well as English translations for most surviving authors and works. If we assume that Galen and early church fathers would double the size of the Loeb, then we would have c. 500 volumes worth of Greek and Latin source text. Measured by word count, the corpora of classical Greek and Latin are closer to 100 and 20 million words respectively.[25]
- **c. 1,000 manuscripts (MS) and an undetermined number of papyri, many very small fragments of literary works.** Based on a survey of summary data from Richard and Olivier's *Repertoire des bibliothèques et des catalogues de manuscrits grecs* (1995), we possess more than 30,000 medieval manuscripts that contain at least parts of Greek classical texts (there are nearly 1,200 manuscripts for

Aristotle alone). Since the number of extant Latin manuscripts is conventionally assumed to be 5 to 10 times that of Greek manuscripts, there might be as many as 150,000 to 300,000 manuscripts for Classical Latin. Nevertheless, a small subset of these provide most of the textual information relevant to the authors and editions of the most commonly studied authors. Hall’s early twentieth-century Handbook to Classical Texts summarized the major MS sources for major classical authors and contains c. 650 readily identifiable MS sigla (e.g., patterns of the form “A = Parisinus 7794”) — while editors have since added additional manuscripts of importance for most authors, Hall provides a reasonable estimate for the number of the manuscripts on which our editions primarily depend. Some authors do not have a few very authoritative MSS and editors must examine large numbers of MSS of roughly equal authority, and these will inflate the total. Assuming that this list underestimates the whole by 50-100%, we are still left with the evidence that a database of 1,000 MSS would represent the majority of textual knowledge preserved for us by MS transmission.

- **c. 5,000 major editions over the five centuries extending from the editiones principes of the early modern period to the start of the twenty-first century.** Assuming at the high end that each author has c. 10 volumes worth of major editions. Multi-work canonical authors will have many editions of individual and selected works. At the very high end, the New Variorum Shakespeare series chooses c. 50 editions of each play as worth collation and this may represent an upper bound for canonical texts outside the Bible.
- **c. 5,000 translations in European languages such as English, French, German and Italian.** These are important because we can use parallel text analysis to infer translation equivalents and word senses and then use advanced language services (e.g., syntactic analysis, named entity analysis) on the translations and then project this backwards onto the original. Such a technique can, for example, add 15% to our current ability to analyze Latin syntax (e.g., from 54% to 70%).
- **c. 5,000 modern commentaries, author lexica etc.** These are useful for human readers and may lend machine actionable data as well.
- **c. 1,000 general reference works such as lexica, grammars, encyclopedias, indices** and other entry/labeled paragraph reference works with high concentrations of citations and, in some cases, elaborate knowledge bearing hierarchical structures.
- **c. 1000 specialized studies of Greek and Latin language** in a sufficiently structured format for high precision information extraction.

Three Technical Challenges

The implications of very large collections for the humanities are profound. We can transform existing research agendas, 26
render content physically and intellectually accessible to new audiences and make human inquiry possible over barriers of language, culture and sheer volume. An immense amount must be — and is being — done. Within this context, we offer three strategic areas of development that are both essential for the humanities and are not, to our knowledge, currently covered by industrially driven research. These areas of interest include the need to transform page images into machine-readable text, machine readable text into machine actionable knowledge, and text from one language into another. Each of these areas of development reflects the particular needs of humanities scholarship and would benefit from targeted support.

- **Leverage the fact that many historical texts quote documents for which excellent transcriptions exist in machine-readable form**

  Thus, the tenth century Venetus A manuscript (Figure 1 and Figure 2) and Jensen’s 1475 incunabulum (Figure 3 and Figure 4) contain texts of Homer and Augustine. We need systems that can use their knowledge that a given document represents texts for which transcriptions exist to decode the writing system of the document, to separate text from headings, notes, and others annotations, to recognize and expand idiosyncratic abbreviations of words within the text, to distinguish variants from errors, and to provide alignments between the transcribed text and their probable equivalents on the written page. Even if we only succeed in general alignments between a canonical text and sources such as early modern printed books and manuscripts, the results will be significant.[26] If we can improve our ability to collate manuscripts or extract useful text from otherwise intractable sources, the results will be powerful.

This task requires very different OCR technology from that currently in use. In this case, we assume that our texts contain many passages for which we possess good transcriptions. The problem becomes (1) finding those quotations, (2) learning what written symbols correspond to various components of transcription, and (3) comparing multiple versions of the same passage to distinguish variants and errors. The OCR system uses a library of known texts to learn new fonts, idiosyncratic abbreviations and even handwriting.

There are two measures for this category of OCR. First, there is the overall character accuracy of transcriptional output from documents that the OCR software produces by training itself with recognized quotations. Second, the ability to locate quotations of existing texts is an important scholarly task in and of itself.[27] Two of the prime tasks in the German eAqua Classics Text Mining Project focus on identifying undiscovered quotations of Plato and of Greek Fragmentary Historians.[28] The apparatus criticus for the Ahlberg *Sallust* (Figure 9 and Figure 10), for example, includes not only textual variants but *testimonia* — places where later authors have quoted Sallust. Such manually constructed lists of testimonia provide us with instruments with which to measure precision and recall for automated methods.

- **Use propositional data already available to decode the formats in which unrecognized knowledge has been stored.**

  Printed reference works contain an immense body of information that can be converted into machine actionable knowledge. The Perseus Digital Library, to take one example, has tagged hundreds of thousands of propositional data within reference works originally published on paper. Thus, the Liddell-Scott-Jones Greek-English (Figure 13 and Figure 14) and Lewis and Short Latin-English lexica, for example, contain tagged citations to 422,000 Greek and 303,000 Latin authors (i.e., citations tagged with author numbers from the *TLG* and *PHI* canons of Greek and Latin authors). Since the structure of the dictionary articles has also been tagged, many of these citations represent propositional statements of the form SENSE-M of DICTIONARY-WORD-N appears in CITATION-P of AUTHOR-Q.

  The works of many Greek and Roman authors survive only insofar as other authors have quoted or described them. These fragmentary texts are published as lists of excerpts (Figure 12). Thus, fragment 116 of the historian Ephorus in Mueller's edition contains an excerpt from chapter 12 of Plutarch's *Life of Cimon*. Each of which represents the propositional statement "EXCERPT-A frm CITATION-C of AUTHOR-D refers to (fragmentary) AUTHOR-X." Note that not all citations refer to the author: thus, fragment 113 of Ephorus includes a cross-reference for background information on a historical event in Herodotus, who wrote before Ephorus.

  Grammars also contain well-structured information: citations within a section on contrary to fact conditionals, for example, (Figure 15 and Figure 16 through Figure 18) can be converted into propositional form: e.g., GRAMMATICAL-STRUCTURE CONTRARY-TO-FACT occurs at Xenophon's *Cyropaedia*, book 1, chapter 2, section 16. Fine-grained analysis of the print content can also extract quotations and their English translations that appear throughout reference grammars and lexica. Smyth's Greek Grammar, the German Kühner-Gerth reference *Greek Grammar*, and the Allen and Greenough Latin Grammar contain 5,300, 21,000 citations and 2,000 tagged citations within labeled sections

  Citations in indices of proper names and in encyclopedias about people and places provide similar propositional data to disambiguate references to ambiguous names: thus, the print index to Rawlinson's Herodotus (Figure 20) distinguishes passages where Herodotus cites Alexander, a king of Macedon, from Alexander, the son of Priam who appears in the Trojan War. Encyclopedias (Figure 22) contain citations from many different sources and many different people and places with the same name. By converting the citations to links and then extracting the contexts in which different Alexanders appear, machine learning algorithms can be used to find patterns with which to distinguish one Alexander from another elsewhere. The Smith's biographical and geographical dictionaries contain 37,000 tagged citations for 20,000 entries on people and 26,000 tagged citations for 10,000 entries on places. The Perseus Encyclopedia, integrating entries from originally separate print indices contains 69,000 citations for 13,000 entries.

A great deal of information remains to be mined from the print record and we need to be able to leverage the information already extracted to extract even more from the much larger body of reference materials available only as page images.

Extraction contains at least two dimensions. In each case, we need more scalable methods.

- **Parsing the structure of individual documents:** Even if we can recognize that “Th. 1.33” represents a citation to a text, we need to determine whether this cites book 1, chapter 33 of Thucydides’ *Peloponnesian War* or Idyll 1, line 33 of Theocritus. The indices shown at Figure 20, Figure 21, Figure 15, etc. illustrate some of the varying formats with which different works encode similar information
- **Aligning information from different documents:** Author indices distinguish different people and places with the same name in the same document, but aligning information from multiple author indices is not easy. Is Alexander the son of Amyntas in Herodotus the same person as Alexander the father of Perdiccas in Thucydides?

- **Use existing translations of source texts to generate multi-lingual services such as cross language information retrieval, word sense disambiguation and other searching/translation services.**

  There are already English translations aligned by canonical citation to more than 5,000,000 words of Greek and Latin available in the Perseus Digital Library. These provide enough parallel text to support basic multi-lingual services such as contextualized word glossing (e.g., recognizing in a given context whether oratio is more likely to correspond to “prayer,” “oration” or some other word sense), cross language information retrieval (e.g., being able to generate “prayer” and “oration” as possible English equivalents of Latin *oratio*), and semantic searching (e.g., find all Latin and Greek words that probably correspond to the English word “prayer” in particular passages).

  The larger our collections of parallel text and translation, the more powerful the services can become. We need methods to locate more translations of Greek and Latin and then to align these with their sources. In some cases, library metadata will allow us to identify translations of particular Greek and Latin works. In other cases, however, we will need to depend upon cross-language information retrieval to find translations where no machine actionable cataloguing exists (e.g., anthologies, quotations of excerpts or smaller works).

  Once we have identified a translation, we need automated methods to align translation and text. Figure 11 shows a best case scenario: a book where the modern translation and classical source text are printed side by side. In this case, the modern translation shares the chapter number of the Latin source text (both have “LXIV” to indicate that they include chapter 64), but the English translation does not include the finer grained section numbers in the Latin text. We need automated methods to align the many translations now appearing in large image book collections.

Conclusion

Comprehensive collections of industrially scanned written materials provide historic new instruments with which to better understand and to make intellectually accessible the record of human existence. These comprehensive collections of scanned print materials are, however, not an end in themselves but instead provide the foundation on which new collections, integrating images of writing with machine actionable data, will support a new generation of services for a new generation of intellectual projects.

Appendix: Sample Page Images

Primary Sources

The 10th Century Venetus A MS of Homer

40

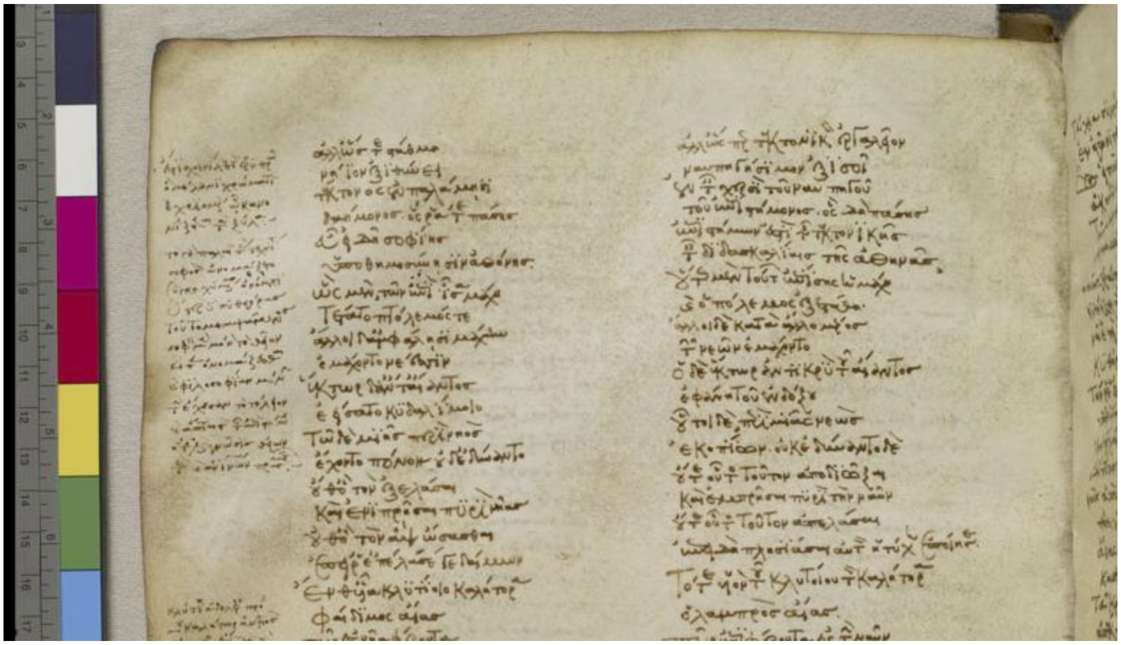


Figure 1. The 10th Century Venetus A MS of Homer: U4 (Allen): Marcianus Graecus Z. 458 (= 841) - the back (verso) of folio 15 (available under a Creative Commons license from Harvard's Center for Hellenic Studies: http://chs.harvard.edu/chs/manuscript_images) The knowledge based OCR project recommended in this report would allow us to work with manuscripts as well as printed materials.

41

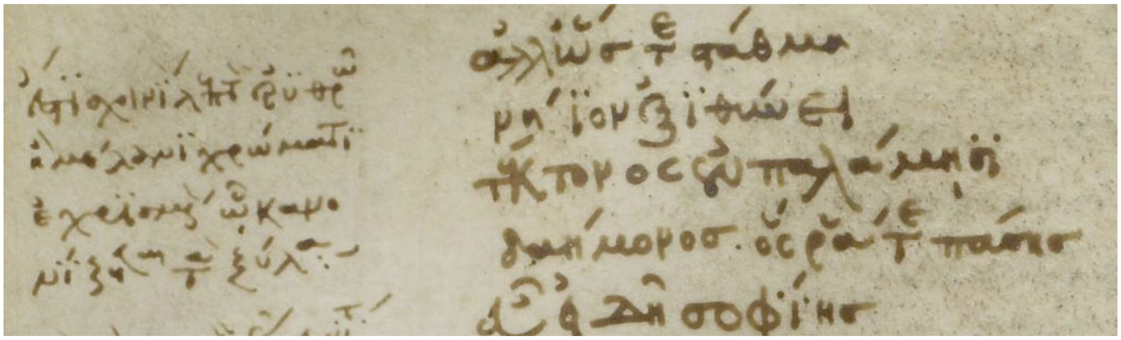
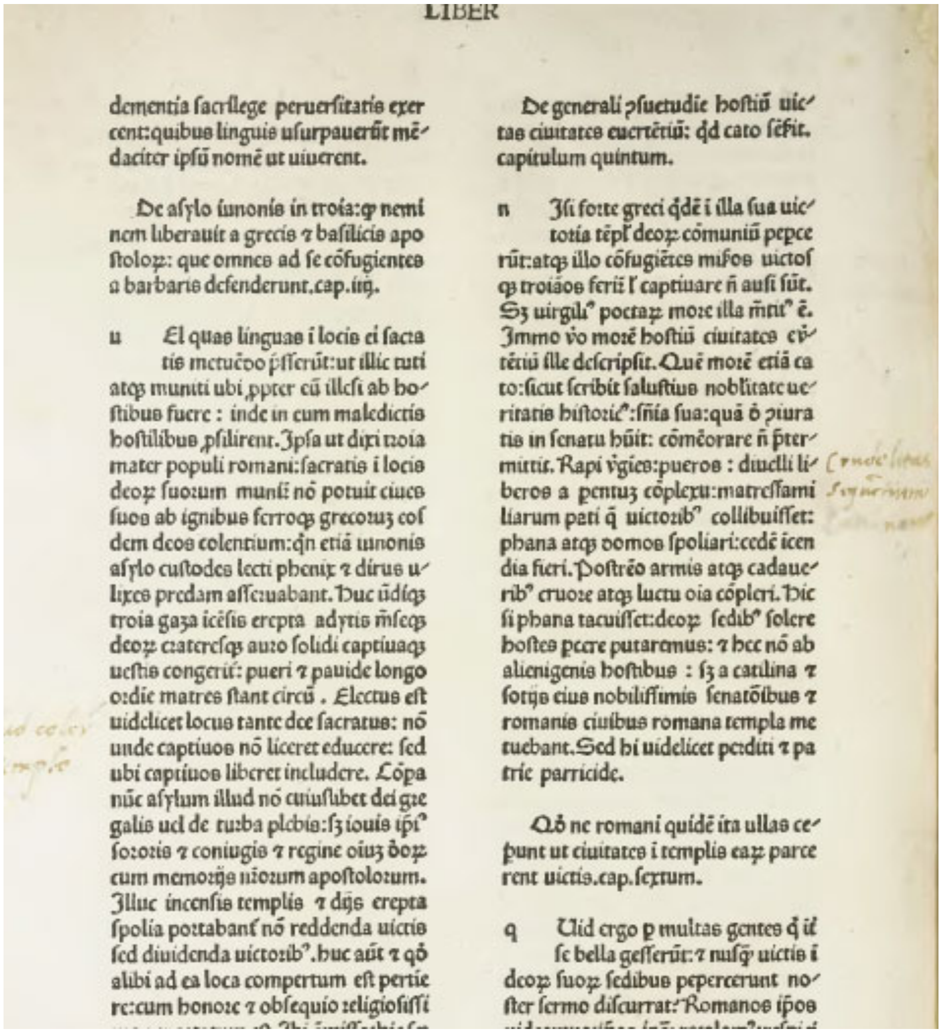


Figure 2. Detail of the Venetus A showing scholia and text

The 1475 Jensen printing of Augustine's *De Civitate Dei*

42



LIBER

dementia ſacrílege peruerſitatis exer-
cent: quibus linguis uſurpauerũt mẽ-
daciter ipſũ nomẽ ut uiuerent.

De aſylo iunonis in troia: ꝙ nemi-
nem liberauit a grecis ⁊ baſilicis apo-
ſtoloꝝ: que omnes ad ſe cõfugientes
a barbaris defenderunt. cap. iiij.

u El quas linguas i locis ei ſacra-
tis metuẽdo ꝓfferũt: ut illic tuti
atqꝫ muniti ubi ꝓpter eũ illeſi ab ho-
ſtibus fuere: inde in eum maledictis
hoſtilibus ꝓſilirent. Ipſa ut dixi troia
mater populi romani: ſacratis i locis
deoꝝ ſuorum munĩ nõ potuit ciues
ſuos ab ignibus ferroqꝫ grecorũ eoſ-
dem deos colentium: q̃n etiã iunonis
aſylo cuſtodes lecti phenix ⁊ dirus u-
lixes predam aſſeruabant. Huc ũdiqꝫ
troia gaza icẽſis erepta adytis mẽſeqꝫ
deoꝝ crateresqꝫ auro ſolidi captiuaqꝫ
ueſtis congerit̃: pueri ⁊ pauide longo
ordie matres ſtant circũ. Electus eſt
uidelicet locus tante dee ſacratus: nõ
unde captiuos nõ liceret educere: ſed
ubi captiuos liberet includere. Cõpa
nũc aſylum illud nõ cuiuſlibet dei gre-
galis uel de turba plebis: ſꝫ iouis ipſi᷒
ſororis ⁊ coniugis ⁊ regine oĩuꝫ doꝝ
cum memorijs nr̃orum apoſtolorum.
Illuc incenſis templis ⁊ dijs erepta
ſpolia portabant nõ reddenda uictis
ſed diuidenda uictorib᷒. huc aũt ⁊ qd̃
alibi ad ea loca compertum eſt pertie-
re: cum honore ⁊ obſequio religioſiſſi

De generali ꝯſuetudie hoſtiũ uic-
tas ciuitates euertẽtiũ: q̃d cato ſẽſit.
capitulum quintum.

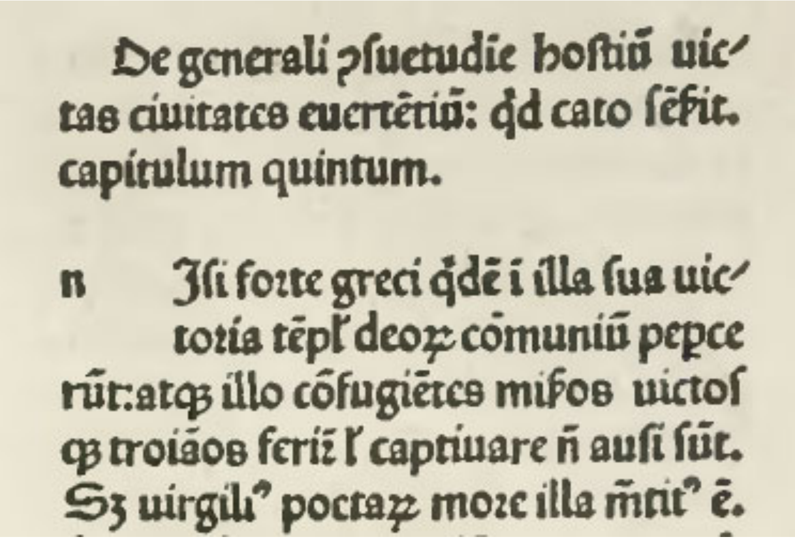
n Iſi forte greci q̃dẽ i illa ſua uic-
toria tẽpl' deoꝝ cõmuniũ pepce-
rũt: atqꝫ illo cõfugiẽtes miſeros uictoſ-
qꝫ troiãos ferirẽ l' captiuare ñ auſi ſũt.
Sꝫ uirgili᷒ poetaꝝ more illa mẽtit᷒ ẽ.
Immo uo morẽ hoſtiũ ciuitates eu-
ẽtiũ ille deſcripſit. Quẽ morẽ etiã ca-
to: ſicut ſcribit ſaluſtius nobĩitate ue-
ritatis hiſtoric᷒: ſñia ſua: quã ð ꝯiura-
tis in ſenatu hũit: cõmẽorare ñ ꝑter-
mittit. Rapi ṽgies: pueros: diuelli li-
beros a parẽtuꝫ cõplexu: matreſfami-
liarum pati q̃ uictorib᷒ collibuiſſet:
phana atqꝫ domos ſpoliari: cedẽ icen-
dia fieri. Poſtrẽo armis atqꝫ cadaue-
rib᷒ cruore atqꝫ luctu oĩa cõpleri. Hic
ſi phana tacuiſſet: deoꝝ ſedib᷒ ſolere
hoſtes parcere putaremus: ⁊ hec nõ ab
alienigenis hoſtibus: ſꝫ a catilina ⁊
ſocijs eius nobiliſſimis ſenatorĩbus ⁊
romanis ciuibus romana templa me-
tuebant. Sed hi uidelicet perditi ⁊ pa-
trie parricide.

Crudelitas [illegible]

Qd̃ ne romani quidẽ ita ullas ce-
ꝑunt ut ciuitates i templis eaꝝ parce-
rent uictis. cap. ſextum.

q Uid ergo ꝑ multas gentes q̃ iſ-
ſe bella geſſerũt: ⁊ nuſq̃ uictis i
deoꝝ ſuoꝝ ſedibus pepercerunt no-
ſter ſermo diſcurrat? Romanos ipos

Figure 3. A page from Nicholas Jensen's 1475 printing of Augstine's *De Civitate Dei* available for public download from the Open Content Alliance (http://www.archive.org/details/augustinidecivitatedei00jensuoft/)



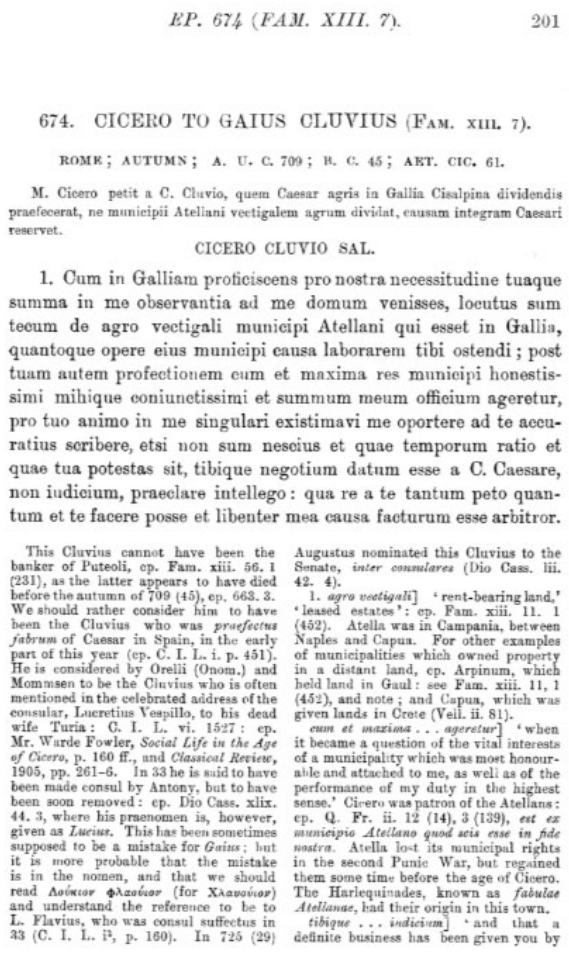
De generali ꝯſuetudie hoſtiũ uic-
tas ciuitates euertẽtiũ: q̃d cato ſẽſit.
capitulum quintum.

n Iſi forte greci q̃dẽ i illa ſua uic-
toria tẽpl' deoꝝ cõmuniũ pepce-
rũt: atqꝫ illo cõfugiẽtes miſeros uictoſ-
qꝫ troiãos ferirẽ l' captiuare ñ auſi ſũt.
Sꝫ uirgili᷒ poetaꝝ more illa mẽtit᷒ ẽ.

Figure 4. Detail of Jensen's Augustine

Tyrrell's Edition of Cicero's Letters

44



EP. 674 (FAM. XIII. 7). 201

674. CICERO TO GAIUS CLUVIUS (Fam. xiii. 7).

ROME; AUTUMN; A. U. C. 709; B. C. 45; AET. CIC. 61.

M. Cicero petit a C. Cluvio, quem Caesar agris in Gallia Cisalpina dividendis praefecerat, ne municipii Atellani vectigalem agrum dividat, causam integram Caesari reservet.

CICERO CLUVIO SAL.

1. Cum in Galliam proficiscens pro nostra necessitudine tuaque summa in me observantia ad me domum venisses, locutus sum tecum de agro vectigali municipi Atellani qui esset in Gallia, quantoque opere eius municipi causa laborarem tibi ostendi; post tuam autem profectionem cum et maxima res municipi honestissimi mihique coniunctissimi et summum meum officium ageretur, pro tuo animo in me singulari existimavi me oportere ad te accuratius scribere, etsi non sum nescius et quae temporum ratio et quae tua potestas sit, tibique negotium datum esse a C. Caesare, non iudicium, praeclare intellego: qua re a te tantum peto quantum et te facere posse et libenter mea causa facturum esse arbitror.

This Cluvius cannot have been the banker of Puteoli, cp. Fam. xiii. 56. 1 (231), as the latter appears to have died before the autumn of 709 (45), cp. 663. 3. We should rather consider him to have been the Cluvius who was *praefectus fabrum* of Caesar in Spain, in the early part of this year (cp. C. I. L. i. p. 451). He is considered by Orelli (Onom.) and Mommsen to be the Cluvius who is often mentioned in the celebrated address of the consular, Lucretius Vespillo, to his dead wife Turia: C. I. L. vi. 1527: cp. Mr. Warde Fowler, *Social Life in the Age of Cicero*, p. 160 ff., and *Classical Review*, 1905, pp. 261–6. In 33 he is said to have been made consul by Antony, but to have been soon removed: cp. Dio Cass. xlix. 44. 3, where his praenomen is, however, given as *Lucius*. This has been sometimes supposed to be a mistake for *Gaius*; but it is more probable that the mistake is in the nomen, and that we should read *Λούκιον Φλαούιον* (for Χλαουούιον) and understand the reference to be to L. Flavius, who was consul suffectus in 33 (C. I. L. i^{2}, p. 160). In 725 (29) Augustus nominated this Cluvius to the Senate, *inter consulares* (Dio Cass. lii. 42. 4).

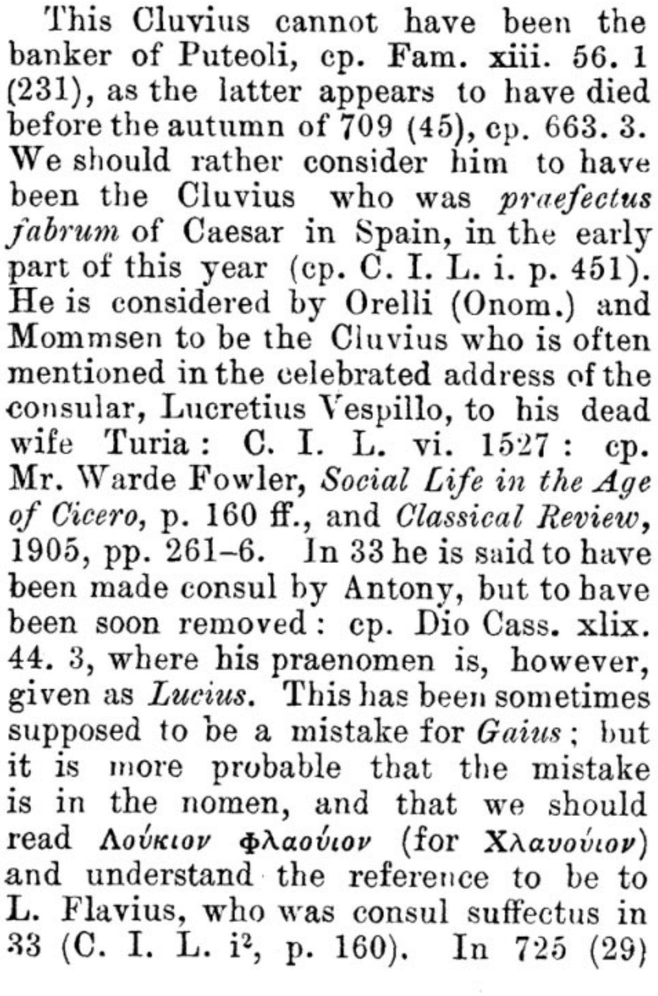
1. *agro vectigali*] 'rent-bearing land,' 'leased estates': cp. Fam. xiii. 11. 1 (452). Atella was in Campania, between Naples and Capua. For other examples of municipalities which owned property in a distant land, cp. Arpinum, which held land in Gaul: see Fam. xiii. 11, 1 (452), and note; and Capua, which was given lands in Crete (Vell. ii. 81).

cum et maxima . . . ageretur] 'when it became a question of the vital interests of a municipality which was most honourable and attached to me, as well as of the performance of my duty in the highest sense.' Cicero was patron of the Atellans: cp. Q. Fr. ii. 12 (14), 3 (139), *est ex municipio Atellano quod scis esse in fide nostra*. Atella lost its municipal rights in the second Punic War, but regained them some time before the age of Cicero. The Harlequinades, known as *fabulae Atellanae*, had their origin in this town.

tibique . . . iudicium] 'and that a definite business has been given you by

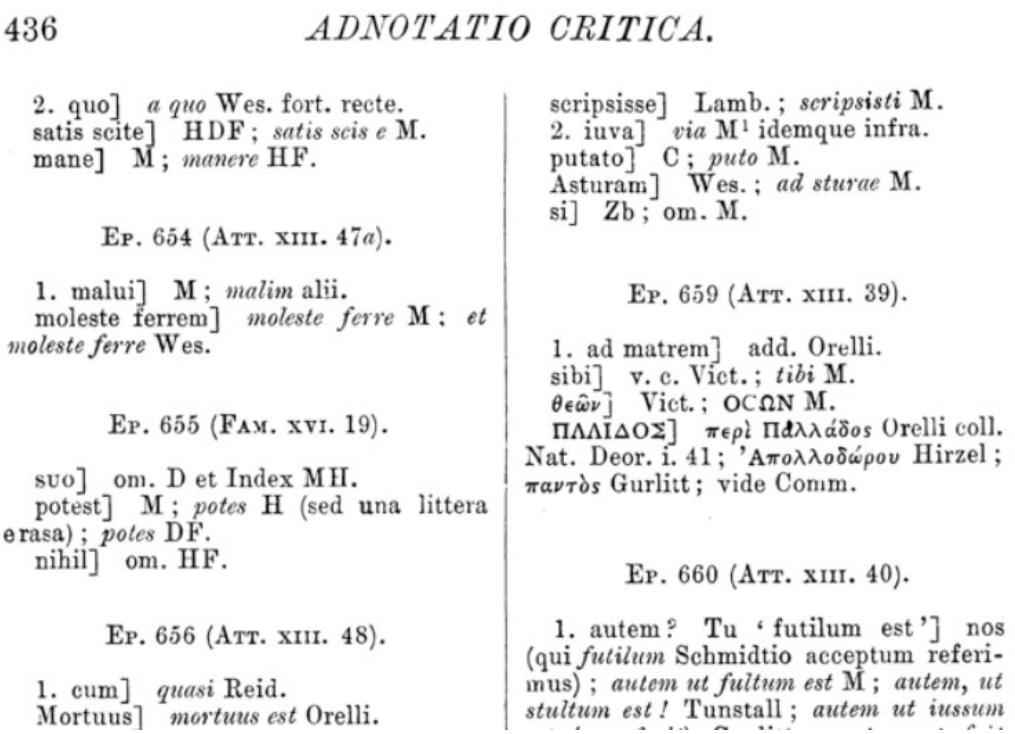
Figure 5. Tyrell's text and commentary of Cicero's Letters.

45



This Cluvius cannot have been the banker of Puteoli, cp. Fam. xiii. 56. 1 (231), as the latter appears to have died before the autumn of 709 (45), cp. 663. 3. We should rather consider him to have been the Cluvius who was *praefectus fabrum* of Caesar in Spain, in the early part of this year (cp. C. I. L. i. p. 451). He is considered by Orelli (Onom.) and Mommsen to be the Cluvius who is often mentioned in the celebrated address of the consular, Lucretius Vespillo, to his dead wife Turia: C. I. L. vi. 1527: cp. Mr. Warde Fowler, *Social Life in the Age of Cicero*, p. 160 ff., and *Classical Review*, 1905, pp. 261–6. In 33 he is said to have been made consul by Antony, but to have been soon removed: cp. Dio Cass. xlix. 44. 3, where his praenomen is, however, given as *Lucius*. This has been sometimes supposed to be a mistake for *Gaius*; but it is more probable that the mistake is in the nomen, and that we should read Λούκιον Φλαούιον (for Χλαυούιον) and understand the reference to be to L. Flavius, who was consul suffectus in 33 (C. I. L. i^2, p. 160). In 725 (29)

Figure 6. A detail showing Tyrrell’s commentary on the page above.



436 *ADNOTATIO CRITICA.*

2. quo] *a quo* Wes. fort. recte.
satis scite] HDF; *satis scis e* M.
mane] M; *manere* HF.

Ep. 654 (Att. xiii. 47*a*).

1. malui] M; *malim* alii.
moleste ferrem] *moleste ferre* M: *et moleste ferre* Wes.

Ep. 655 (Fam. xvi. 19).

suo] om. D et Index MH.
potest] M; *potes* H (sed una littera erasa); *potes* DF.
nihil] om. HF.

Ep. 656 (Att. xiii. 48).

1. cum] *quasi* Reid.
Mortuus] *mortuus est* Orelli.

scripsisse] Lamb.; *scripsisti* M.
2. iuva] *via* M^1 idemque infra.
putato] C; *puto* M.
Asturam] Wes.; *ad sturae* M.
si] Zb; om. M.

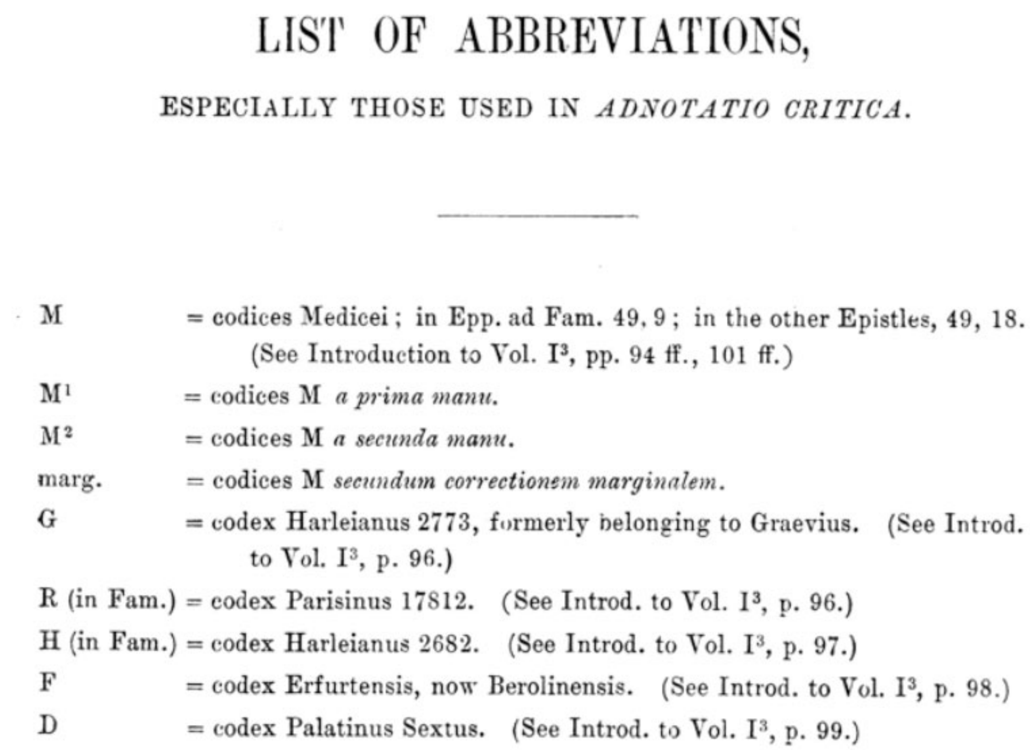
Ep. 659 (Att. xiii. 39).

1. ad matrem] add. Orelli.
sibi] v. c. Vict.; *tibi* M.
θεῶν] Vict.; OCΩN M.
ΠΛΛΙΔΟΣ] *περὶ Παλλάδος* Orelli coll. Nat. Deor. i. 41; *'Απολλοδώρου* Hirzel; *παντὸς* Gurlitt; vide Comm.

Ep. 660 (Att. xiii. 40).

1. autem? Tu 'futilum est'] nos (qui *futilum* Schmidtio acceptum referimus); *autem ut fultum est* M; *autem, ut stultum est!* Tunstall; *autem ut iussum*

Figure 7. Textual notes in Tyrrell stored in an appendix rather than at the bottom of the page.



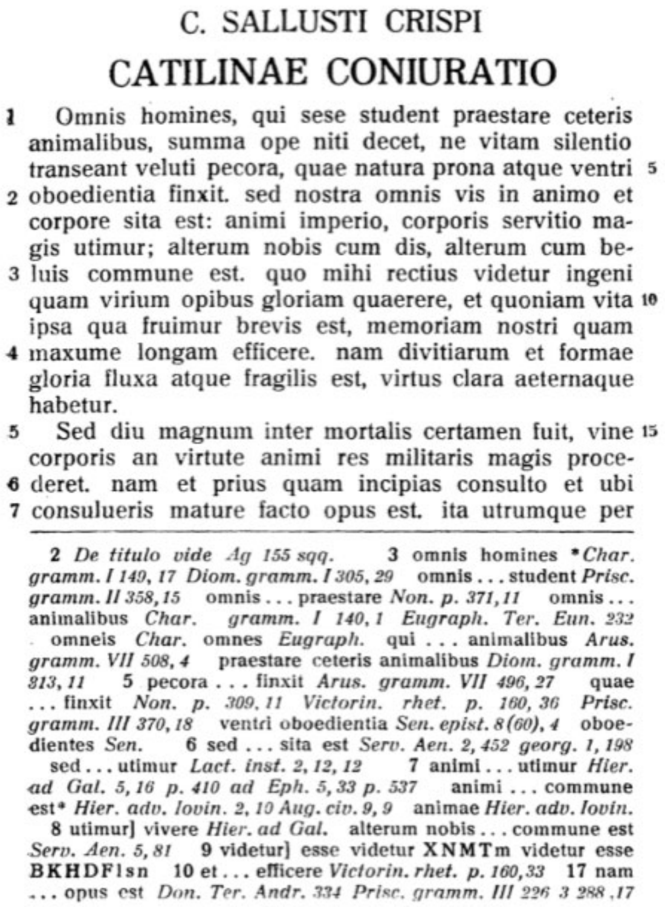
LIST OF ABBREVIATIONS,

ESPECIALLY THOSE USED IN *ADNOTATIO CRITICA.*

M	= codices Medicei; in Epp. ad Fam. 49, 9; in the other Epistles, 49, 18. (See Introduction to Vol. I^3, pp. 94 ff., 101 ff.)
M^1	= codices M *a prima manu.*
M^2	= codices M *a secunda manu.*
marg.	= codices M *secundum correctionem marginalem.*
G	= codex Harleianus 2773, formerly belonging to Graevius. (See Introd. to Vol. I^3, p. 96.)
R (in Fam.)	= codex Parisinus 17812. (See Introd. to Vol. I^3, p. 96.)
H (in Fam.)	= codex Harleianus 2682. (See Introd. to Vol. I^3, p. 97.)
F	= codex Erfurtensis, now Berolinensis. (See Introd. to Vol. I^3, p. 98.)
D	= codex Palatinus Sextus. (See Introd. to Vol. I^3, p. 99.)

Figure 8. Abbreviations used in the textual notes and commentary.

Ahlberg's 1919 Edition of Sallust



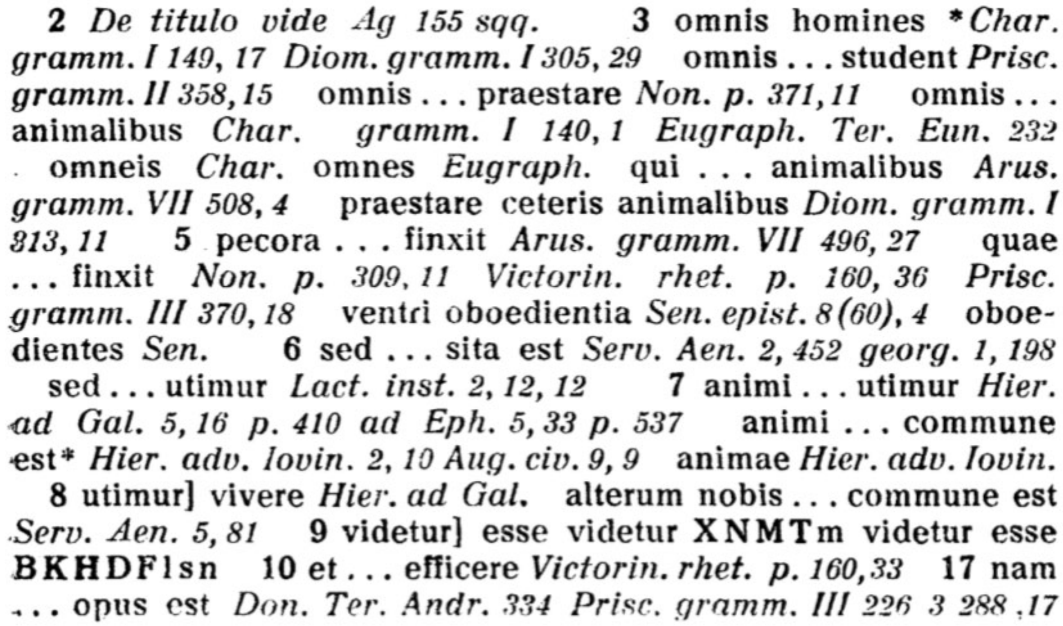
C. SALLUSTI CRISPI

CATILINAE CONIURATIO

1 Omnis homines, qui sese student praestare ceteris
animalibus, summa ope niti decet, ne vitam silentio
transeant veluti pecora, quae natura prona atque ventri
2 oboedientia finxit. sed nostra omnis vis in animo et
corpore sita est: animi imperio, corporis servitio magis
utimur; alterum nobis cum dis, alterum cum beluis
3 commune est. quo mihi rectius videtur ingeni
quam virium opibus gloriam quaerere, et quoniam vita
ipsa qua fruimur brevis est, memoriam nostri quam
4 maxume longam efficere. nam divitiarum et formae
gloria fluxa atque fragilis est, virtus clara aeternaque
habetur.
5 Sed diu magnum inter mortalis certamen fuit, vine
corporis an virtute animi res militaris magis procederet.
6 nam et prius quam incipias consulto et ubi
7 consulueris mature facto opus est. ita utrumque per

2 *De titulo vide Ag 155 sqq.* 3 omnis homines **Char. gramm. I 149, 17 Diom. gramm. I 305, 29* omnis . . . student *Prisc. gramm. II 358, 15* omnis . . . praestare *Non. p. 371, 11* omnis . . . animalibus *Char. gramm. I 140, 1 Eugraph. Ter. Eun. 232* omneis *Char.* omnes *Eugraph.* qui . . . animalibus *Arus. gramm. VII 508, 4* praestare ceteris animalibus *Diom. gramm. I 813, 11* 5 pecora . . . finxit *Arus. gramm. VII 496, 27* quae . . . finxit *Non. p. 309, 11 Victorin. rhet. p. 160, 36 Prisc. gramm. III 370, 18* ventri oboedientia *Sen. epist. 8 (60), 4* oboedientes *Sen.* 6 sed . . . sita est *Serv. Aen. 2, 452 georg. 1, 198* sed . . . utimur *Lact. inst. 2, 12, 12* 7 animi . . . utimur *Hier. ad Gal. 5, 16 p. 410 ad Eph. 5, 33 p. 537* animi . . . commune est* *Hier. adv. Iovin. 2, 10 Aug. civ. 9, 9* animae *Hier. adv. Iovin.* 8 utimur] vivere *Hier. ad Gal.* alterum nobis . . . commune est *Serv. Aen. 5, 81* 9 videtur] esse videtur XNMTm videtur esse BKHDFlsn 10 et . . . efficere *Victorin. rhet. p. 160, 33* 17 nam . . . opus est *Don. Ter. Andr. 334 Prisc. gramm. III 226 3 288, 17*

Figure 9. The opening of Sallust's *Catiline* in Axel Ahlberg's 1919 *Editio Major*.

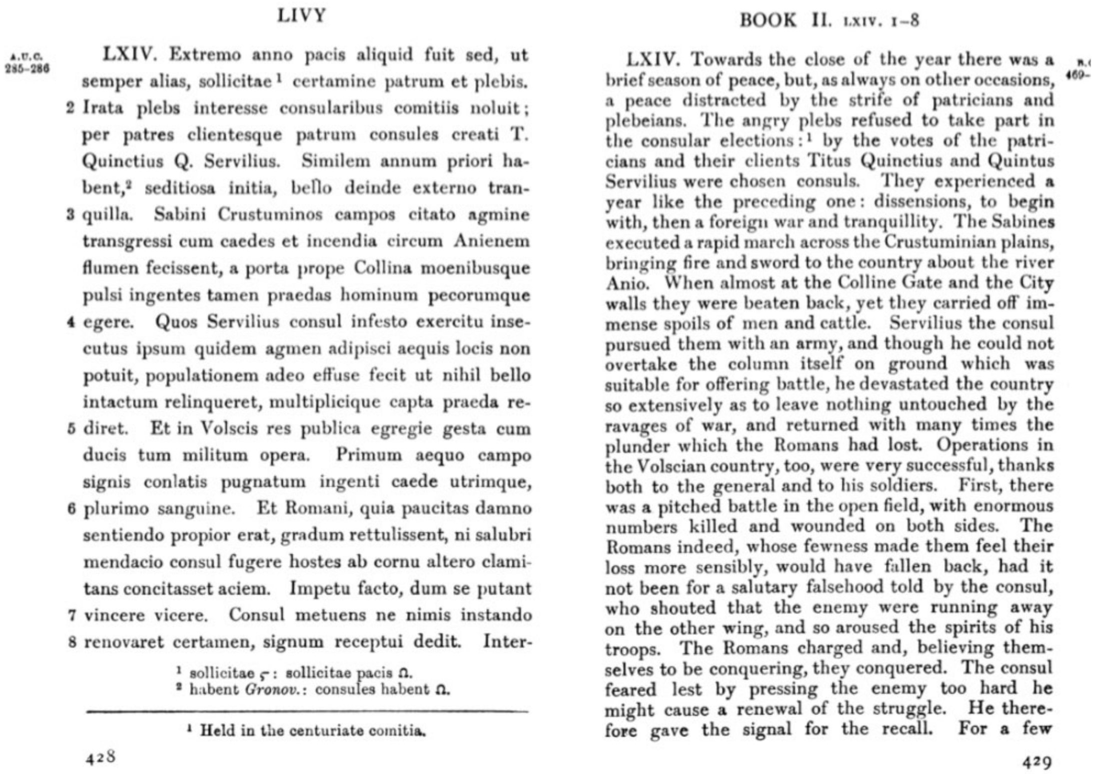


2 *De titulo vide Ag 155 sqq.* 3 omnis homines **Char. gramm. I 149, 17 Diom. gramm. I 305, 29* omnis . . . student *Prisc. gramm. II 358, 15* omnis . . . praestare *Non. p. 371, 11* omnis . . . animalibus *Char. gramm. I 140, 1 Eugraph. Ter. Eun. 232* omneis *Char.* omnes *Eugraph.* qui . . . animalibus *Arus. gramm. VII 508, 4* praestare ceteris animalibus *Diom. gramm. I 813, 11* 5 pecora . . . finxit *Arus. gramm. VII 496, 27* quae . . . finxit *Non. p. 309, 11 Victorin. rhet. p. 160, 36 Prisc. gramm. III 370, 18* ventri oboedientia *Sen. epist. 8 (60), 4* oboedientes *Sen.* 6 sed . . . sita est *Serv. Aen. 2, 452 georg. 1, 198* sed . . . utimur *Lact. inst. 2, 12, 12* 7 animi . . . utimur *Hier. ad Gal. 5, 16 p. 410 ad Eph. 5, 33 p. 537* animi . . . commune est* *Hier. adv. Iovin. 2, 10 Aug. civ. 9, 9* animae *Hier. adv. Iovin.* 8 utimur] vivere *Hier. ad Gal.* alterum nobis . . . commune est *Serv. Aen. 5, 81* 9 videtur] esse videtur XNMTm videtur esse BKHDFlsn 10 et . . . efficere *Victorin. rhet. p. 160, 33* 17 nam . . . opus est *Don. Ter. Andr. 334 Prisc. gramm. III 226 3 288, 17*

Figure 10. The apparatus criticus from the page above.

Translations

50



LIVY

A.U.C.
285–286

LXIV. Extremo anno pacis aliquid fuit sed, ut
semper alias, sollicitae[1] certamine patrum et plebis.
2 Irata plebs interesse consularibus comitiis noluit;
per patres clientesque patrum consules creati T.
Quinctius Q. Servilius. Similem annum priori ha-
bent,[2] seditiosa initia, bello deinde externo tran-
3 quilla. Sabini Crustuminos campos citato agmine
transgressi cum caedes et incendia circum Anienem
flumen fecissent, a porta prope Collina moenibusque
pulsi ingentes tamen praedas hominum pecorumque
4 egere. Quos Servilius consul infesto exercitu inse-
cutus ipsum quidem agmen adipisci aequis locis non
potuit, populationem adeo effuse fecit ut nihil bello
intactum relinqueret, multiplicique capta praeda re-
5 diret. Et in Volscis res publica egregie gesta cum
ducis tum militum opera. Primum aequo campo
signis conlatis pugnatum ingenti caede utrimque,
6 plurimo sanguine. Et Romani, quia paucitas damno
sentiendo propior erat, gradum rettulissent, ni salubri
mendacio consul fugere hostes ab cornu altero clami-
tans concitasset aciem. Impetu facto, dum se putant
7 vincere vicere. Consul metuens ne nimis instando
8 renovaret certamen, signum receptui dedit. Inter-

[1] sollicitae ς: sollicitae pacis Ω.
[2] habent *Gronov.*: consules habent Ω.

428

BOOK II. LXIV. 1–8

B.
469-

LXIV. Towards the close of the year there was a brief season of peace, but, as always on other occasions, a peace distracted by the strife of patricians and plebeians. The angry plebs refused to take part in the consular elections:[1] by the votes of the patricians and their clients Titus Quinctius and Quintus Servilius were chosen consuls. They experienced a year like the preceding one: dissensions, to begin with, then a foreign war and tranquillity. The Sabines executed a rapid march across the Crustuminian plains, bringing fire and sword to the country about the river Anio. When almost at the Colline Gate and the City walls they were beaten back, yet they carried off immense spoils of men and cattle. Servilius the consul pursued them with an army, and though he could not overtake the column itself on ground which was suitable for offering battle, he devastated the country so extensively as to leave nothing untouched by the ravages of war, and returned with many times the plunder which the Romans had lost. Operations in the Volscian country, too, were very successful, thanks both to the general and to his soldiers. First, there was a pitched battle in the open field, with enormous numbers killed and wounded on both sides. The Romans indeed, whose fewness made them feel their loss more sensibly, would have fallen back, had it not been for a salutary falsehood told by the consul, who shouted that the enemy were running away on the other wing, and so aroused the spirits of his troops. The Romans charged and, believing themselves to be conquering, they conquered. The consul feared lest by pressing the enemy too hard he might cause a renewal of the struggle. He therefore gave the signal for the recall. For a few

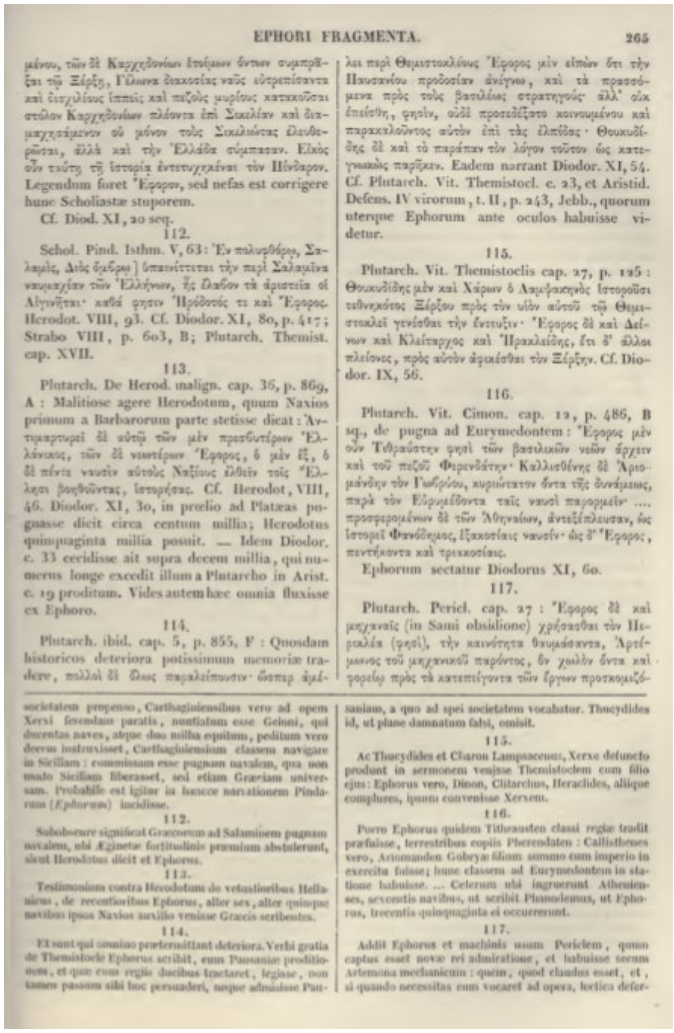
[1] Held in the centuriate comitia.

429

Figure 11. Facing Latin text and English translation (R. O. Foster, from the first volume of the Loeb Classical Library Livy series (Cambridge 1919)).

Editions of Fragmentary Authors and Works

51



EPHORI FRAGMENTA. 265

μένου, τῶν δὲ Καρχηδονίων ἑτοίμων ὄντων συμπρᾶξαι τῷ Ξέρξῃ, Γέλωνα διακοσίας ναῦς εὐτρεπίσαντα καὶ δισχιλίους ἱππεῖς καὶ πεζοὺς μυρίους κατακοῦσαι στόλον Καρχηδονίων πλέοντα ἐπὶ Σικελίαν καὶ διαμαχησάμενον οὐ μόνον τοὺς Σικελιώτας ἐλευθερῶσαι, ἀλλὰ καὶ τὴν Ἑλλάδα σύμπασαν. Εἰκὸς οὖν ταύτῃ τῇ ἱστορίᾳ ἐντετυχηκέναι τὸν Πίνδαρον. Legendum foret Ἔφορον, sed nefas est corrigere hunc Scholiastæ stuporem.

Cf. Diod. XI, 20 seq.

112.

Schol. Pind. Isthm. V, 63: Ἐν πολυφθόρῳ, Σαλαμὶς, Διὸς ὄμβρῳ] ὑπαινίττεται τὴν περὶ Σαλαμῖνα ναυμαχίαν τῶν Ἑλλήνων, ἧς ἔλαβον τὰ ἀριστεῖα οἱ Αἰγινῆται· καθά φησιν Ἡρόδοτός τε καὶ Ἔφορος. Herodot. VIII, 93. Cf. Diodor. XI, 80, p. 417; Strabo VIII, p. 603, B; Plutarch. Themist. cap. XVII.

113.

Plutarch. De Herod. malign. cap. 36, p. 869, A : Malitiose agere Herodotum, quum Naxios primum a Barbarorum parte stetisse dicat : Ἀντιμαρτυρεῖ δὲ αὐτῷ τῶν μὲν πρεσβυτέρων Ἑλλάνικος, τῶν δὲ νεωτέρων Ἔφορος, ὁ μὲν ἕξ, ὁ δὲ πέντε ναυσὶν αὐτοὺς Ναξίους ἐλθεῖν τοῖς Ἕλλησι βοηθοῦντας, ἱστορήσας. Cf. Herodot, VIII, 46. Diodor. XI, 30, in prœlio ad Platæas pugnasse dicit circa centum millia; Herodotus quinquaginta millia posuit. — Idem Diodor. c. 33 cecidisse ait supra decem millia, qui numerus longe excedit illum a Plutarcho in Arist. c. 19 proditum. Vides autem hæc omnia fluxisse ex Ephoro.

114.

Plutarch. ibid. cap. 5, p. 855, F : Quosdam historicos deteriora potissimum memoriæ tradere, πολλοὶ δὲ ὅλως παραλείπουσιν· ὥσπερ ἀμέλει περὶ Θεμιστοκλέους Ἔφορος μὲν εἰπὼν ὅτι τὴν Παυσανίου προδοσίαν ἀνέγνω, καὶ τὰ πρασσόμενα πρὸς τοὺς βασιλέως στρατηγούς· ἀλλ' οὐκ ἐπείσθη, φησὶν, οὐδὲ προσεδέξατο κοινουμένου καὶ παρακαλοῦντος αὐτὸν ἐπὶ τὰς ἐλπίδας· Θουκυδίδης δὲ καὶ τὸ παράπαν τὸν λόγον τοῦτον ὡς κατεγνωκὼς παρῆκεν. Eadem narrant Diodor. XI, 54. Cf. Plutarch. Vit. Themistocl. c. 23, et Aristid. Defens. IV virorum, t. II, p. 243, Jebb., quorum uterque Ephorum ante oculos habuisse videtur.

115.

Plutarch. Vit. Themistoclis cap. 27, p. 125 : Θουκυδίδης μὲν καὶ Χάρων ὁ Λαμψακηνὸς ἱστοροῦσι τεθνηκότος Ξέρξου πρὸς τὸν υἱὸν αὐτοῦ τῷ Θεμιστοκλεῖ γενέσθαι τὴν ἔντευξιν· Ἔφορος δὲ καὶ Δείνων καὶ Κλείταρχος καὶ Ἡρακλείδης, ἔτι δ' ἄλλοι πλείονες, πρὸς αὐτὸν ἀφικέσθαι τὸν Ξέρξην. Cf. Diodor. IX, 56.

116.

Plutarch. Vit. Cimon. cap. 12, p. 486, B sq., de pugna ad Eurymedontem : Ἔφορος μὲν οὖν Τιθραύστην φησὶ τῶν βασιλικῶν νεῶν ἄρχειν καὶ τοῦ πεζοῦ Φερενδάτην· Καλλισθένης δὲ Ἀριομάνδην τὸν Γωβρύου, κυριώτατον ὄντα τῆς δυνάμεως, παρὰ τὸν Εὐρυμέδοντα ταῖς ναυσὶ παρορμεῖν· .... προσφερομένων δὲ τῶν Ἀθηναίων, ἀντεξέπλευσαν, ὡς ἱστορεῖ Φανόδημος, ἑξακοσίαις ναυσίν· ὡς δ' Ἔφορος, πεντήκοντα καὶ τριακοσίαις.

Ephorum sectatur Diodorus XI, 60.

117.

Plutarch. Pericl. cap. 27 : Ἔφορος δὲ καὶ μηχαναῖς (in Sami obsidione) χρήσασθαι τὸν Περικλέα (φησὶ), τὴν καινότητα θαυμάσαντα, Ἀρτέμωνος τοῦ μηχανικοῦ παρόντος, ὃν χωλὸν ὄντα καὶ φορείῳ πρὸς τὰ κατεπείγοντα τῶν ἔργων προσκομιζό-

societatem propenso, Carthaginiensibus vero ad opem Xerxi ferendam paratis, nuntiatum esse Geloni, qui ducentas naves, atque duo millia equitum, peditum vero decem instruxisset, Carthaginiensium classem navigare in Siciliam : commissam esse pugnam navalem, qua non modo Siciliam liberasset, sed etiam Græciam universam. Probabile est igitur in hancce narrationem Pindarum (*Ephorum*) incidisse.

112.

Subobscure significat Græcorum ad Salaminem pugnam navalem, ubi Æginetæ fortitudinis præmium abstulerunt, sicut Herodotus dicit et Ephorus.

113.

Testimonium contra Herodotum de vetustioribus Hellanicus, de recentioribus Ephorus, alter sex, alter quinque navibus ipsos Naxios auxilio venisse Græcis scribentes.

114.

Et sunt qui omnino prætermittant deteriora. Verbi gratia de Themistocle Ephorus scribit, eum Pausaniæ proditionem, et quæ cum regiis ducibus tractaret, legisse, non tamen passum sibi hoc persuaderi, neque admisisse Pausaniam, a quo ad spei societatem vocabatur. Thucydides id, ut plane damnatum falsi, omisit.

115.

Ac Thucydides et Charon Lampsacenus, Xerxe defuncto produnt in sermonem venisse Themistoclem cum filio ejus: Ephorus vero, Dinon, Clitarchus, Heraclides, aliique complures, ipsum convenisse Xerxem.

116.

Porro Ephorus quidem Tithrausten classi regiæ tradit præfuisse, terrestribus copiis Pherendaten : Callisthenes vero, Ariomanden Gobryæ filium summo cum imperio in exercitu fuisse; hunc classem ad Eurymedontem in statione habuisse. ... Ceterum ubi ingruerunt Athenienses, sexcentis navibus, ut scribit Phanodemus, ut Ephorus, trecentis quinquaginta ei occurrerunt.

117.

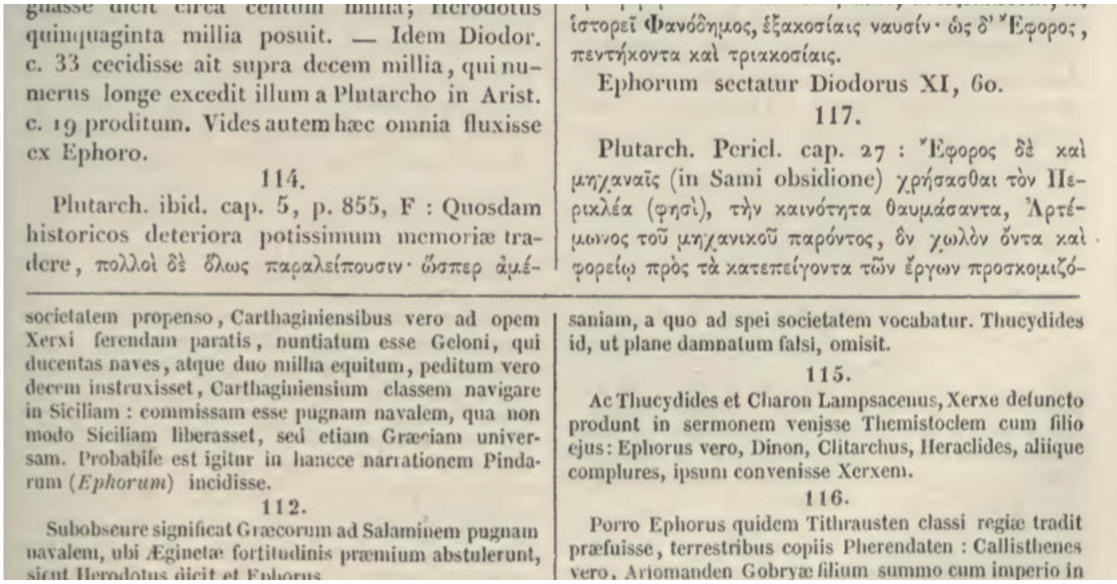
Addit Ephorus et machinis usum Periclem, quum captus esset novæ rei admiratione, et habuisse secum Artemona mechanicum : quem, quod claudus esset, et, si quando necessitas eum vocaret ad opera, lectica defer-

Figure 12. Typical page from Mueller's *Fragmenta Graecorum Historicorum*. Above we see an edition of a fragmentary Greek author — quotations of and allusions to the Greek historian Ephorus, whose works have been lost. Each fragment contains one or more citations to works that provide information about a particular passage in Ephorus. The format is Fragment number — Citation — Excerpt. Latin translations of the Greek excerpts appear at the bottom of the page.

Reference works

Lexica

Lidell Scott Greek-English Lexicon



gnasse dicit circa centum millia; Herodotus quinquaginta millia posuit. — Idem Diodor. c. 33 cecidisse ait supra decem millia, qui numerus longe excedit illum a Plutarcho in Arist. c. 19 proditum. Vides autem hæc omnia fluxisse ex Ephoro.

114.

Plutarch. ibid. cap. 5, p. 855, F : Quosdam historicos deteriora potissimum memoriæ tradere, πολλοὶ δὲ ὅλως παραλείπουσιν· ὥσπερ ἀμέ-

ἱστορεῖ Φανόδημος, ἑξακοσίαις ναυσίν· ὡς δ' Ἔφορος, πεντήκοντα καὶ τριακοσίαις.

Ephorum sectatur Diodorus XI, 60.

117.

Plutarch. Pericl. cap. 27 : Ἔφορος δὲ καὶ μηχαναῖς (in Sami obsidione) χρήσασθαι τὸν Περικλέα (φησὶ), τὴν καινότητα θαυμάσαντα, Ἀρτέμωνος τοῦ μηχανικοῦ παρόντος, ὃν χωλὸν ὄντα καὶ φορείῳ πρὸς τὰ κατεπείγοντα τῶν ἔργων προσκομιζό-

societatem propenso, Carthaginiensibus vero ad opem Xerxi ferendam paratis, nuntiatum esse Geloni, qui ducentas naves, atque duo millia equitum, peditum vero decem instruxisset, Carthaginiensium classem navigare in Siciliam : commissam esse pugnam navalem, qua non modo Siciliam liberasset, sed etiam Græciam universam. Probabile est igitur in hancce narrationem Pindarum (*Ephorum*) incidisse.

112.

Subobscure significat Græcorum ad Salaminem pugnam navalem, ubi Æginetæ fortitudinis præmium abstulerunt, sicut Herodotus dicit et Ephorus

saniam, a quo ad spei societatem vocabatur. Thucydides id, ut plane damnatum falsi, omisit.

115.

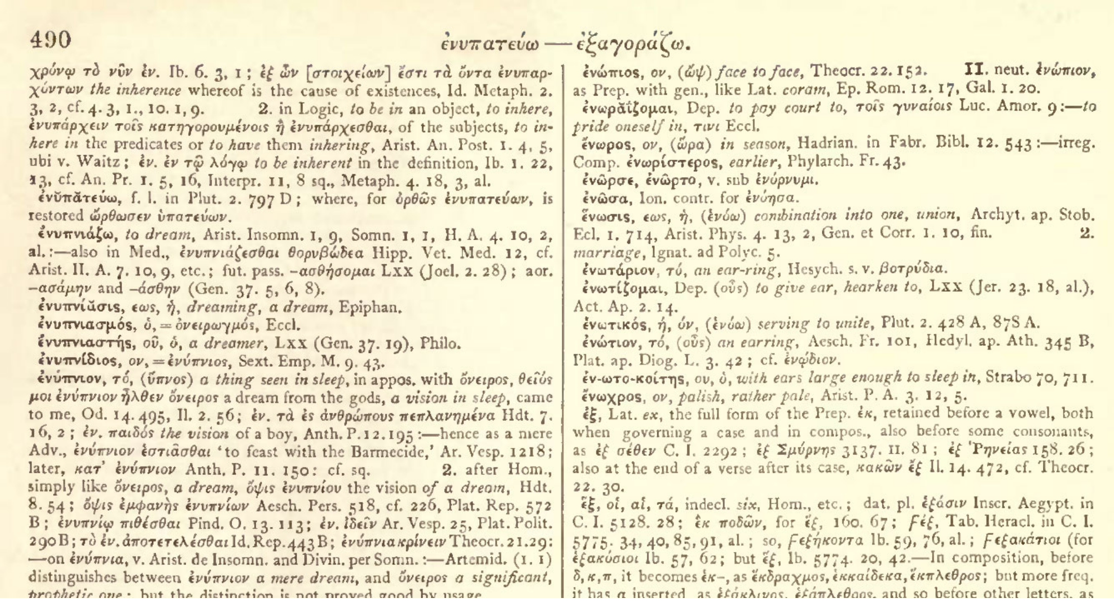
Ac Thucydides et Charon Lampsacenus, Xerxe defuncto produnt in sermonem venisse Themistoclem cum filio ejus: Ephorus vero, Dinon, Clitarchus, Heraclides, aliique complures, ipsum convenisse Xerxem.

116.

Porro Ephorus quidem Tithrausten classi regiæ tradit præfuisse, terrestribus copiis Pherendaten : Callisthenes vero, Ariomanden Gobryæ filium summo cum imperio in

Figure 13. A typical page from an edition of the Liddell Scott Greek English Lexicon (available from the Open Content Alliance: http://www.archive.org/details/greekenglishlex00liddrich/).

53



490 ἐνυπατεύω — ἐξαγοράζω.

χρόνῳ τὸ νῦν ἐν. Ib. 6. 3, 1 ; *ἐξ ὧν* [*στοιχείων*] *ἔστι τὰ ὄντα ἐνυπαρχόντων the inherence* whereof is the cause of existences, Id. Metaph. 2. 3, 2, cf. 4. 3, 1., 10. 1, 9. **2.** in Logic, *to be in* an object, *to inhere*, *ἐνυπάρχειν τοῖς κατηγορουμένοις ἢ ἐνυπάρχεσθαι*, of the subjects, *to inhere in* the predicates or *to have* them *inhering*, Arist. An. Post. 1. 4, 5, ubi v. Waitz ; *ἐν. ἐν τῷ λόγῳ to be inherent* in the definition, Ib. 1. 22, 13, cf. An. Pr. 1. 5, 16, Interpr. 11, 8 sq., Metaph. 4. 18, 3, al.

ἐνῠπᾰτεύω, f. l. in Plut. 2. 797 D ; where, for *ὀρθῶς ἐνυπατεύων*, is restored *ὤρθωσεν ὑπατεύων*.

ἐνυπνιάζω, *to dream*, Arist. Insomn. 1, 9, Somn. 1, 1, H. A. 4. 10, 2, al. :—also in Med., *ἐνυπνιάζεσθαι θορυβώδεα* Hipp. Vet. Med. 12, cf. Arist. H. A. 7. 10, 9, etc. ; fut. pass. *-ασθήσομαι* LXX (Joel. 2. 28) ; aor. *-ασάμην* and *-άσθην* (Gen. 37. 5, 6, 8).

ἐνυπνίᾰσις, *εως, ἡ*, *dreaming*, *a dream*, Epiphan.

ἐνυπνιασμός, *ὁ*, = *ὀνειρωγμός*, Eccl.

ἐνυπνιαστής, *οῦ, ὁ*, *a dreamer*, LXX (Gen. 37. 19), Philo.

ἐνυπνίδιος, *ον*, = *ἐνύπνιος*, Sext. Emp. M. 9. 43.

ἐνύπνιον, *τό*, (*ὕπνος*) *a thing seen in sleep*, in appos. with *ὄνειρος*, *θεῖός μοι ἐνύπνιον ἦλθεν ὄνειρος* a dream from the gods, *a vision in sleep*, came to me, Od. 14. 495, Il. 2. 56 ; *ἐν. τὰ ἐς ἀνθρώπους πεπλανημένα* Hdt. 7. 16, 2 ; *ἐν. παιδός the vision* of a boy, Anth. P. 12. 195 :—hence as a mere Adv., *ἐνύπνιον ἑστιᾶσθαι* 'to feast with the Barmecide,' Ar. Vesp. 1218 ; later, *κατ' ἐνύπνιον* Anth. P. 11. 150 : cf. sq. **2.** after Hom., simply like *ὄνειρος*, *a dream*, *ὄψις ἐνυπνίου* the vision *of a dream*, Hdt. 8. 54 ; *ὄψις ἐμφανὴς ἐνυπνίων* Aesch. Pers. 518, cf. 226, Plat. Rep. 572 B ; *ἐνυπνίῳ πιθέσθαι* Pind. O. 13. 113 ; *ἐν. ἰδεῖν* Ar. Vesp. 25, Plat. Polit. 290 B ; *τὸ ἐν. ἀποτετελέσθαι* Id. Rep. 443 B ; *ἐνύπνια κρίνειν* Theocr. 21. 29 : —on *ἐνύπνια*, v. Arist. de Insomn. and Divin. per Somn. :—Artemid. (1. 1) distinguishes between *ἐνύπνιον a mere dream*, and *ὄνειρος a significant*, *prophetic one* : but the distinction is not proved good by usage.

ἐνώπιος, *ον*, (*ὤψ*) *face to face*, Theocr. 22. 152. **II.** neut. *ἐνώπιον*, as Prep. with gen., like Lat. *coram*, Ep. Rom. 12. 17, Gal. 1. 20.

ἐνωραΐζομαι, Dep. *to pay court to*, *τοῖς γυναίοις* Luc. Amor. 9 :—*to pride oneself in*, *τινι* Eccl.

ἔνωρος, *ον*, (*ὥρα*) *in season*, Hadrian. in Fabr. Bibl. 12. 543 :—irreg. Comp. **ἐνωρίστερος**, *earlier*, Phylarch. Fr. 43.

ἐνῶρσε, ἐνῶρτο, v. sub *ἐνόρνυμι*.

ἐνῶσα, Ion. contr. for *ἐνόησα*.

ἕνωσις, *εως, ἡ*, (*ἑνόω*) *combination into one*, *union*, Archyt. ap. Stob. Ecl. 1. 714, Arist. Phys. 4. 13, 2, Gen. et Corr. 1. 10, fin. **2.** *marriage*, Ignat. ad Polyc. 5.

ἐνωτάριον, *τό*, *an ear-ring*, Hesych. s. v. *βοτρύδια*.

ἐνωτίζομαι, Dep. (*οὖς*) *to give ear*, *hearken to*, LXX (Jer. 23. 18, al.), Act. Ap. 2. 14.

ἑνωτικός, *ή, όν*, (*ἑνόω*) *serving to unite*, Plut. 2. 428 A, 87S A.

ἐνώτιον, *τό*, (*οὖς*) *an earring*, Aesch. Fr. 101, Hedyl. ap. Ath. 345 B, Plat. ap. Diog. L. 3. 42 ; cf. *ἐνῴδιον*.

ἐν-ωτο-κοίτης, *ου, ὁ*, *with ears large enough to sleep in*, Strabo 70, 711.

ἔνωχρος, *ον*, *palish*, *rather pale*, Arist. P. A. 3. 12, 5.

ἐξ, Lat. *ex*, the full form of the Prep. *ἐκ*, retained before a vowel, both when governing a case and in compos., also before some consonants, as *ἐξ σέθεν* C. I. 2292 ; *ἐξ Σμύρνης* 3137. 11. 81 ; *ἐξ Ῥηνείας* 158. 26 ; also at the end of a verse after its case, *κακῶν ἔξ* Il. 14. 472, cf. Theocr. 22. 30.

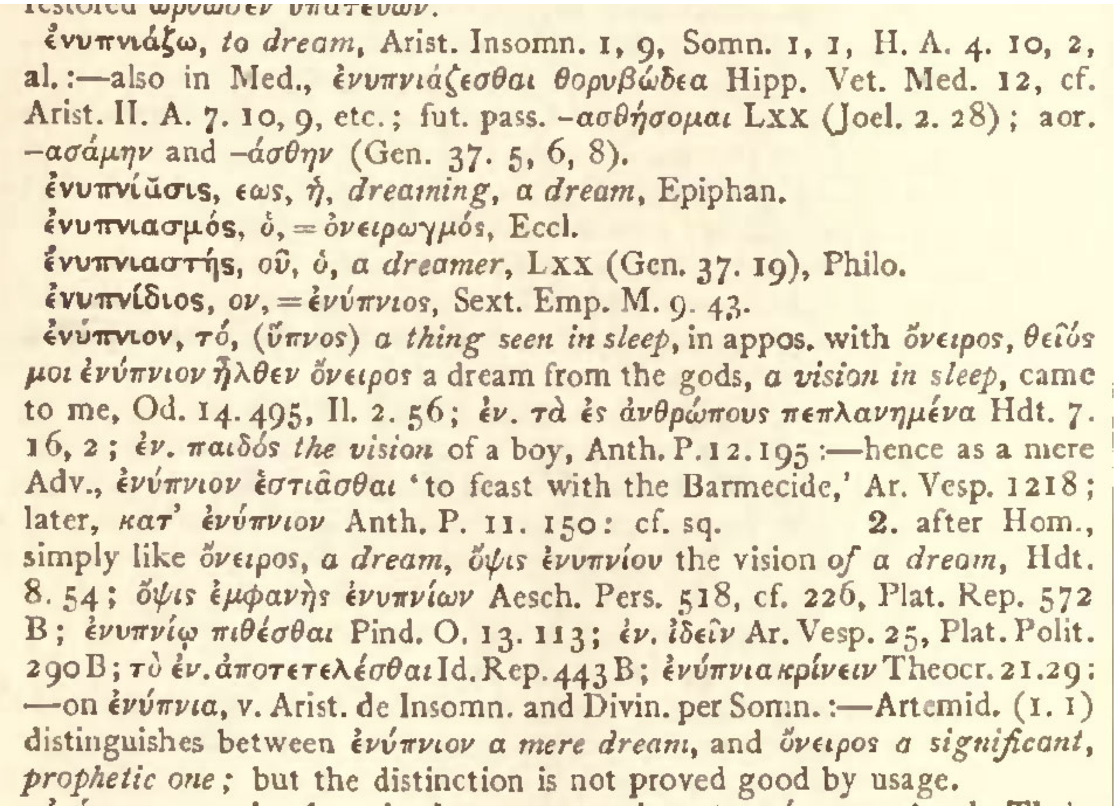
ἕξ, *οἱ, αἱ, τά*, indecl. *six*, Hom., etc. ; dat. pl. *ἑξάσιν* Inscr. Aegypt. in C. I. 5128. 28 ; *ἒκ ποδῶν*, for *ἕξ*, 160. 67 ; *Ϝέξ*, Tab. Heracl. in C. I. 5775. 34, 40, 85, 91, al. ; so, *Ϝεξήκοντα* Ib. 59, 76, al. ; *Ϝεξακάτιοι* (for *ἑξακόσιοι* Ib. 57, 62 ; but *ἕξ*, Ib. 5774. 20, 42.—In composition, before *δ, κ, π*, it becomes *ἐκ-*, as *ἔκδραχμος*, *ἐκκαίδεκα*, *ἔκπλεθρος* ; but more freq. it has *α* inserted, as *ἐξάκλινος*, *ἐξάπλεθρος*, and so before other letters, as

Figure 14. Detail from the Liddell-Scott Greek-English Lexicon

Grammars

Goodwin and Gulick's *Greek Grammar*

54



ἐνυπνιάζω, *to dream*, Arist. Insomn. 1, 9, Somn. 1, 1, H. A. 4. 10, 2, al. :—also in Med., *ἐνυπνιάζεσθαι θορυβώδεα* Hipp. Vet. Med. 12, cf. Arist. II. A. 7. 10, 9, etc.; fut. pass. *-ασθήσομαι* LXX (Joel. 2. 28); aor. *-ασάμην* and *-άσθην* (Gen. 37. 5, 6, 8).

ἐνυπνίᾰσις, *εως*, *ἡ*, *dreaming*, *a dream*, Epiphan.

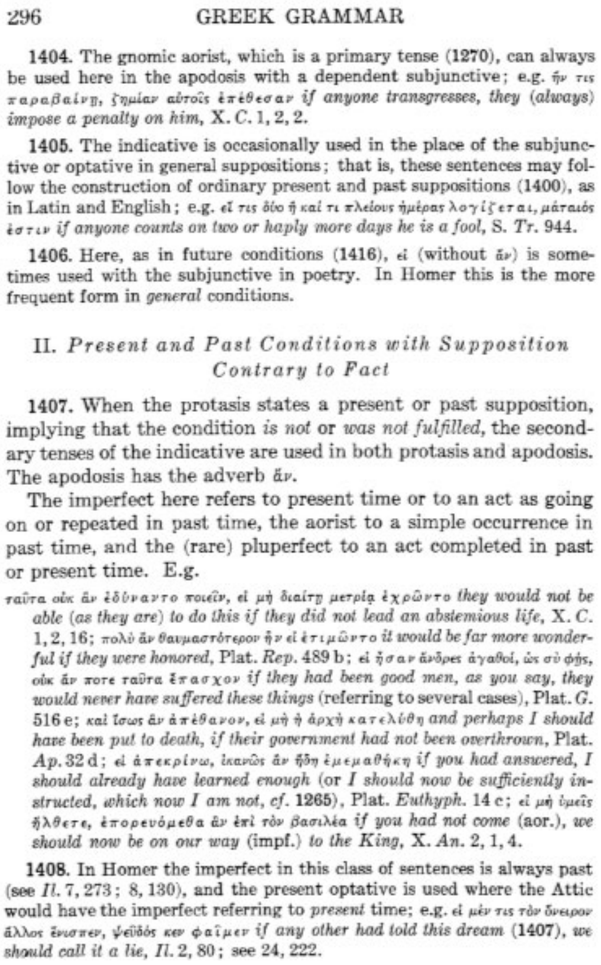
ἐνυπνιασμός, *ὁ*, = *ὀνειρωγμός*, Eccl.

ἐνυπνιαστής, *οῦ*, *ὁ*, *a dreamer*, LXX (Gen. 37. 19), Philo.

ἐνυπνίδιος, *ον*, = *ἐνύπνιος*, Sext. Emp. M. 9. 43.

ἐνύπνιον, *τό*, (*ὕπνος*) *a thing seen in sleep*, in appos. with *ὄνειρος*, *θεῖός μοι ἐνύπνιον ἦλθεν ὄνειρος* a dream from the gods, *a vision in sleep*, came to me, Od. 14. 495, Il. 2. 56; *ἐν. τὰ ἐς ἀνθρώπους πεπλανημένα* Hdt. 7. 16, 2; *ἐν. παιδός the vision* of a boy, Anth. P. 12. 195 :—hence as a mere Adv., *ἐνύπνιον ἑστιᾶσθαι* 'to feast with the Barmecide,' Ar. Vesp. 1218; later, *κατ' ἐνύπνιον* Anth. P. 11. 150: cf. sq. **2.** after Hom., simply like *ὄνειρος*, *a dream*, *ὄψις ἐνυπνίου* the vision *of a dream*, Hdt. 8. 54; *ὄψις ἐμφανὴς ἐνυπνίων* Aesch. Pers. 518, cf. 226, Plat. Rep. 572 B; *ἐνυπνίῳ πιθέσθαι* Pind. O. 13. 113; *ἐν. ἰδεῖν* Ar. Vesp. 25, Plat. Polit. 290 B; *τὸ ἐν. ἀποτετελέσθαι* Id. Rep. 443 B; *ἐνύπνια κρίνειν* Theocr. 21. 29 :—on *ἐνύπνια*, v. Arist. de Insomn. and Divin. per Somn. :—Artemid. (1. 1) distinguishes between *ἐνύπνιον a mere dream*, and *ὄνειρος a significant, prophetic one*; but the distinction is not proved good by usage.

Figure 15. A typical page from Goodwin and Gulick



1404. The gnomic aorist, which is a primary tense (**1270**), can always be used here in the apodosis with a dependent subjunctive; e.g. ἤν τις παραβαίνῃ, ζημίαν αὐτοῖς ἐπέθεσαν *if anyone transgresses, they (always) impose a penalty on him*, X. *C.* 1, 2, 2.

1405. The indicative is occasionally used in the place of the subjunctive or optative in general suppositions; that is, these sentences may follow the construction of ordinary present and past suppositions (**1400**), as in Latin and English; e.g. εἴ τις δύο ἢ καὶ τι πλείους ἡμέρας λογίζεται, μάταιός ἐστιν *if anyone counts on two or haply more days he is a fool*, S. *Tr.* 944.

1406. Here, as in future conditions (**1416**), εἰ (without ἄν) is sometimes used with the subjunctive in poetry. In Homer this is the more frequent form in *general* conditions.

II. *Present and Past Conditions with Supposition Contrary to Fact*

1407. When the protasis states a present or past supposition, implying that the condition *is not* or *was not fulfilled*, the secondary tenses of the indicative are used in both protasis and apodosis. The apodosis has the adverb ἄν.

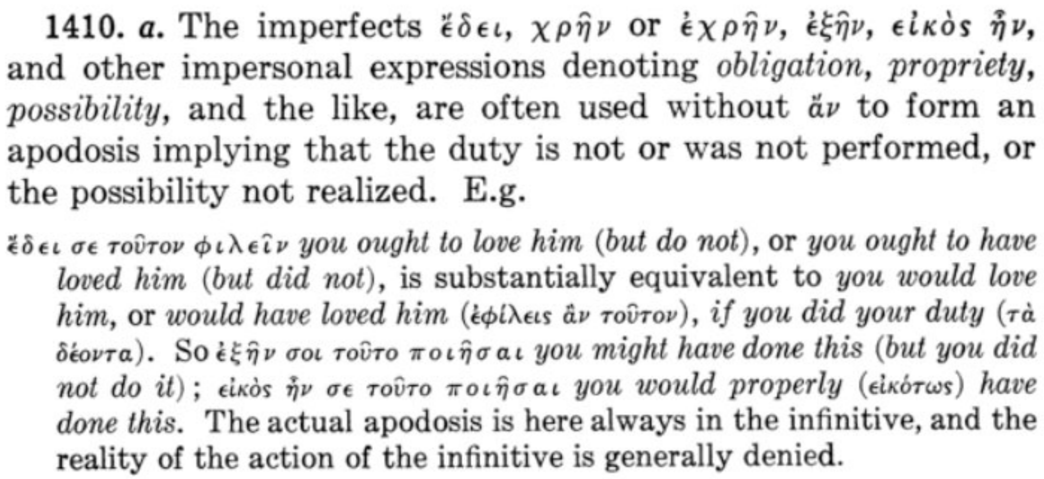
The imperfect here refers to present time or to an act as going on or repeated in past time, the aorist to a simple occurrence in past time, and the (rare) pluperfect to an act completed in past or present time. E.g.

ταῦτα οὐκ ἂν ἐδύναντο ποιεῖν, εἰ μὴ διαίτῃ μετρίᾳ ἐχρῶντο *they would not be able* (*as they are*) *to do this if they did not lead an abstemious life*, X. *C.* 1, 2, 16; πολὺ ἂν θαυμαστότερον ἦν εἰ ἐτιμῶντο *it would be far more wonderful if they were honored*, Plat. *Rep.* 489 b; εἰ ἦσαν ἄνδρες ἀγαθοί, ὡς σὺ φῄς, οὐκ ἄν ποτε ταῦτα ἔπασχον *if they had been good men, as you say, they would never have suffered these things* (referring to several cases), Plat. *G.* 516 e; καὶ ἴσως ἂν ἀπέθανον, εἰ μὴ ἡ ἀρχὴ κατελύθη *and perhaps I should have been put to death, if their government had not been overthrown*, Plat. *Ap.* 32 d; εἰ ἀπεκρίνω, ἱκανῶς ἂν ἤδη ἐμεμαθήκη *if you had answered, I should already have learned enough* (or *I should now be sufficiently instructed, which now I am not, cf.* **1265**), Plat. *Euthyph.* 14 c; εἰ μὴ ὑμεῖς ἤλθετε, ἐπορευόμεθα ἂν ἐπὶ τὸν βασιλέα *if you had not come* (aor.), *we should now be on our way* (impf.) *to the King*, X. *An.* 2, 1, 4.

1408. In Homer the imperfect in this class of sentences is always past (see *Il.* 7, 273; 8, 130), and the present optative is used where the Attic would have the imperfect referring to *present* time; e.g. εἰ μέν τις τὸν ὄνειρον ἄλλος ἔνισπεν, ψεῦδός κεν φαῖμεν *if any other had told this dream* (**1407**), *we should call it a lie*, *Il.* 2, 80; see 24, 222.

Figure 16. A paragraph with number and alphabetic section from the section on contrary to fact conditionals. This paragraph happens to appear on page 297, but the proper reference would be to paragraph 1410a.

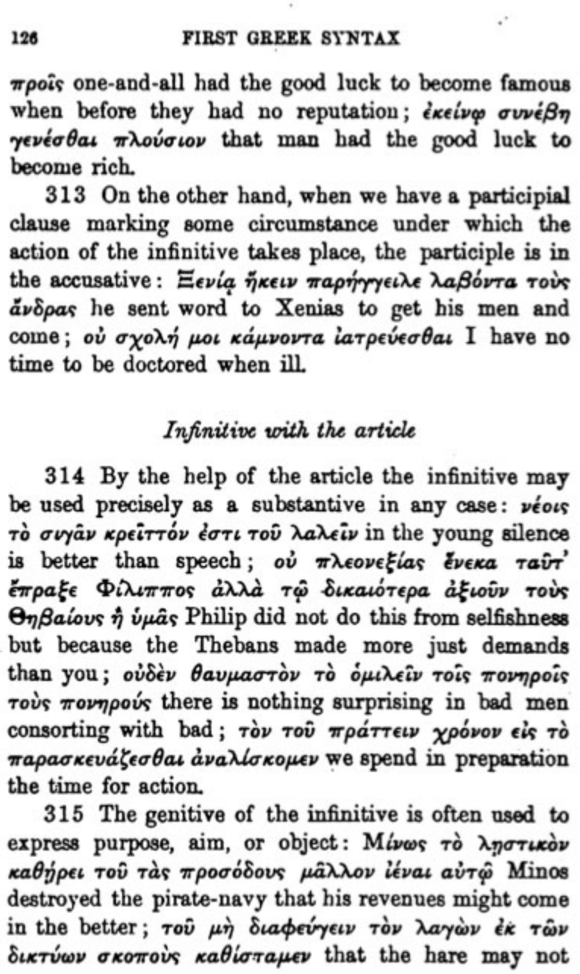
Rutherford's *First Greek Syntax*



1410. ***a.*** The imperfects ἔδει, χρῆν or ἐχρῆν, ἐξῆν, εἰκὸς ἦν, and other impersonal expressions denoting *obligation, propriety, possibility*, and the like, are often used without ἄν to form an apodosis implying that the duty is not or was not performed, or the possibility not realized. E.g.

ἔδει σε τοῦτον φιλεῖν *you ought to love him* (*but do not*), or *you ought to have loved him* (*but did not*), is substantially equivalent to *you would love him*, or *would have loved him* (ἐφίλεις ἂν τοῦτον), *if you did your duty* (τὰ δέοντα). So ἐξῆν σοι τοῦτο ποιῆσαι *you might have done this* (*but you did not do it*); εἰκὸς ἦν σε τοῦτο ποιῆσαι *you would properly* (εἰκότως) *have done this.* The actual apodosis is here always in the infinitive, and the reality of the action of the infinitive is generally denied.

Figure 17. A page from Rutherford's *First Greek Grammar* (downloaded from Google Books).



126 FIRST GREEK SYNTAX

προῖς one-and-all had the good luck to become famous when before they had no reputation; *ἐκείνῳ συνέβη γενέσθαι πλούσιον* that man had the good luck to become rich.

313 On the other hand, when we have a participial clause marking some circumstance under which the action of the infinitive takes place, the participle is in the accusative: *Ξενίᾳ ἥκειν παρήγγειλε λαβόντα τοὺς ἄνδρας* he sent word to Xenias to get his men and come; *οὐ σχολή μοι κάμνοντα ἰατρεύεσθαι* I have no time to be doctored when ill.

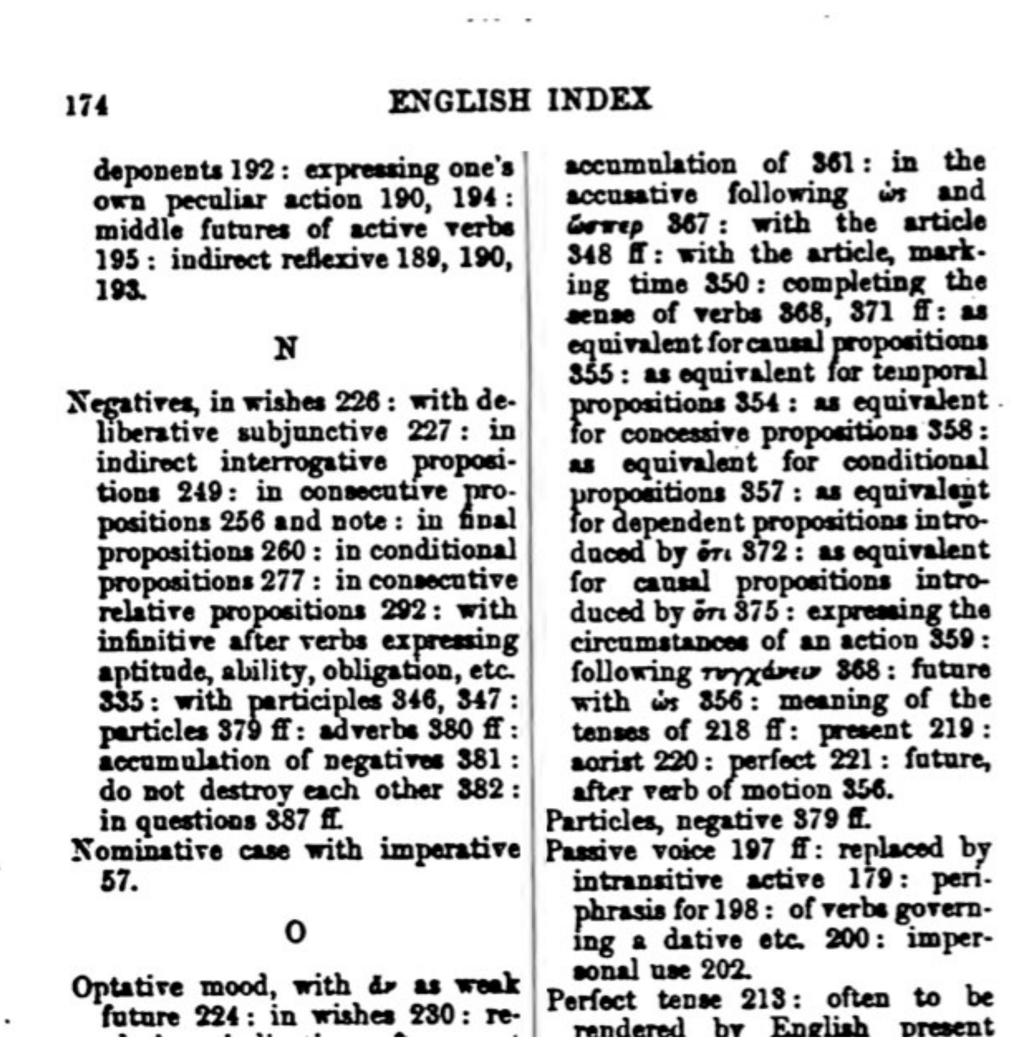
Infinitive with the article

314 By the help of the article the infinitive may be used precisely as a substantive in any case: *νέοις τὸ σιγᾶν κρεῖττόν ἐστι τοῦ λαλεῖν* in the young silence is better than speech; *οὐ πλεονεξίας ἕνεκα ταῦτ᾽ ἔπραξε Φίλιππος ἀλλὰ τῷ δικαιότερα ἀξιοῦν τοὺς Θηβαίους ἢ ὑμᾶς* Philip did not do this from selfishness but because the Thebans made more just demands than you; *οὐδὲν θαυμαστὸν τὸ ὁμιλεῖν τοῖς πονηροῖς τοὺς πονηρούς* there is nothing surprising in bad men consorting with bad; *τὸν τοῦ πράττειν χρόνον εἰς τὸ παρασκευάζεσθαι ἀναλίσκομεν* we spend in preparation the time for action.

315 The genitive of the infinitive is often used to express purpose, aim, or object: *Μίνως τὸ λῃστικὸν καθῄρει τοῦ τὰς προσόδους μᾶλλον ἰέναι αὐτῷ* Minos destroyed the pirate-navy that his revenues might come in the better; *τοῦ μὴ διαφεύγειν τὸν λαγὼν ἐκ τῶν δικτύων σκοποὺς καθίσταμεν* that the hare may not

Figure 18. The index to Rutherford's *First Greek Grammar*: note that citations point to the numbered paragraphs rather than the page numbers. The index appears at the end of the book and an automated system could infer that pages were not the citation scheme because almost all of the numbers in the text above are greater than the current page (174).

Information about People, Places, Organizations and other Named Entities



174 ENGLISH INDEX

deponents 192: expressing one's own peculiar action 190, 194: middle futures of active verbs 195: indirect reflexive 189, 190, 193.

N

Negatives, in wishes 226: with deliberative subjunctive 227: in indirect interrogative propositions 249: in consecutive propositions 256 and note: in final propositions 260: in conditional propositions 277: in consecutive relative propositions 292: with infinitive after verbs expressing aptitude, ability, obligation, etc. 335: with participles 346, 347: particles 379 ff: adverbs 380 ff: accumulation of negatives 381: do not destroy each other 382: in questions 387 ff.

Nominative case with imperative 57.

O

Optative mood, with ἄν as weak future 224: in wishes 230: re-

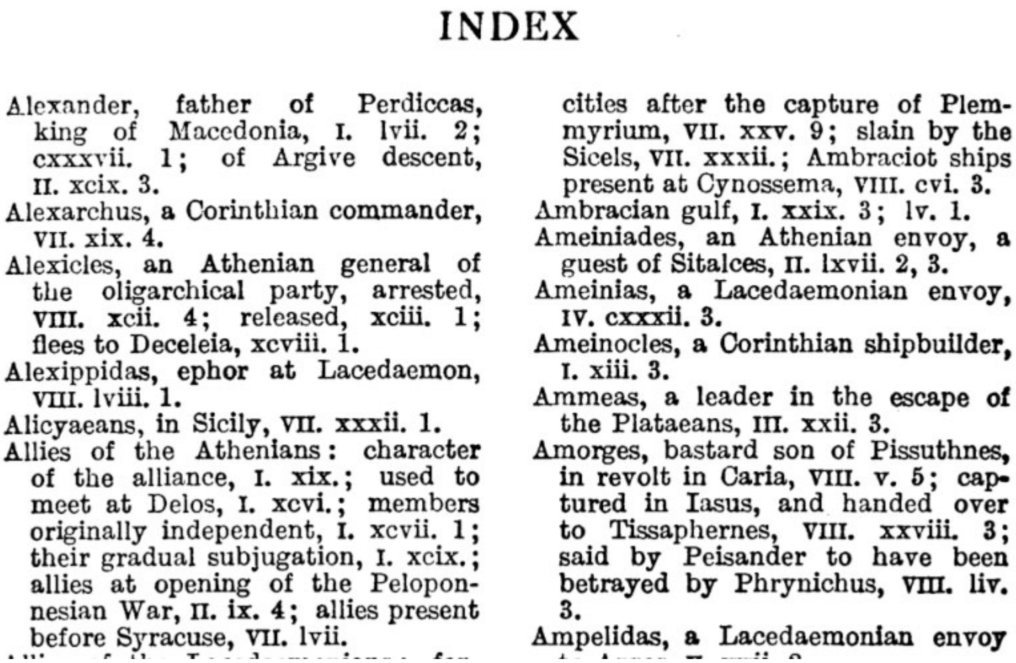
accumulation of 361: in the accusative following ὡς and ὥσπερ 367: with the article 348 ff: with the article, marking time 350: completing the sense of verbs 368, 371 ff: as equivalent for causal propositions 355: as equivalent for temporal propositions 354: as equivalent for concessive propositions 358: as equivalent for conditional propositions 357: as equivalent for dependent propositions introduced by ὅτι 372: as equivalent for causal propositions introduced by ὅτι 375: expressing the circumstances of an action 359: following τυγχάνειν 368: future with ὡς 356: meaning of the tenses of 218 ff: present 219: aorist 220: perfect 221: future, after verb of motion 356.

Particles, negative 379 ff.

Passive voice 197 ff: replaced by intransitive active 179: periphrasis for 198: of verbs governing a dative etc. 200: impersonal use 202.

Perfect tense 213: often to be rendered by English present

Figure 19. Section from the index to the Loeb Edition of Thucydides. In this case, the index uses the canonical book/chapter/section citation scheme, using upper case Roman numerals for books, lower case Roman numerals for chapters and Arabic numbers for sections.



INDEX

Alexander, father of Perdiccas, king of Macedonia, I. lvii. 2; cxxxvii. 1; of Argive descent, II. xcix. 3.

Alexarchus, a Corinthian commander, VII. xix. 4.

Alexicles, an Athenian general of the oligarchical party, arrested, VIII. xcii. 4; released, xciii. 1; flees to Deceleia, xcviii. 1.

Alexippidas, ephor at Lacedaemon, VIII. lviii. 1.

Alicyaeans, in Sicily, VII. xxxii. 1.

Allies of the Athenians: character of the alliance, I. xix.; used to meet at Delos, I. xcvi.; members originally independent, I. xcvii. 1; their gradual subjugation, I. xcix.; allies at opening of the Peloponnesian War, II. ix. 4; allies present before Syracuse, VII. lvii.

cities after the capture of Plemmyrium, VII. xxv. 9; slain by the Sicels, VII. xxxii.; Ambraciot ships present at Cynossema, VIII. cvi. 3.

Ambracian gulf, I. xxix. 3; lv. 1.

Ameiniades, an Athenian envoy, a guest of Sitalces, II. lxvii. 2, 3.

Ameinias, a Lacedaemonian envoy, IV. cxxxii. 3.

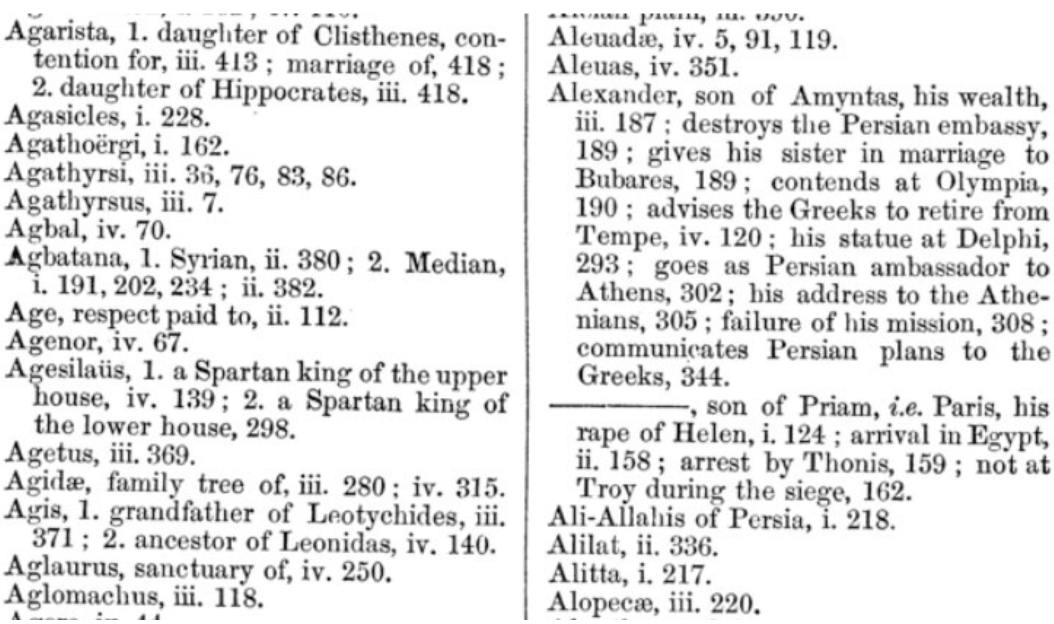
Ameinocles, a Corinthian shipbuilder, I. xiii. 3.

Ammeas, a leader in the escape of the Plataeans, III. xxii. 3.

Amorges, bastard son of Pissuthnes, in revolt in Caria, VIII. v. 5; captured in Iasus, and handed over to Tissaphernes, VIII. xxviii. 3; said by Peisander to have been betrayed by Phrynichus, VIII. liv. 3.

Ampelidas, a Lacedaemonian envoy

Figure 20. Index to Rawlinson’s Herodotus. In this case, the citations point to the particular volumes and page numbers of this translation rather than to the conventional book and chapter references. These references are, however, in the original pages and we could convert the idiosyncratic citations above to a more standard format by checking vol. 3, page 187, for example, to determine that Alexander appears in Herodotus, book 5, chapter 17.



Agarista, 1. daughter of Clisthenes, contention for, iii. 413; marriage of, 418; 2. daughter of Hippocrates, iii. 418.

Agasicles, i. 228.

Agathoërgi, i. 162.

Agathyrsi, iii. 36, 76, 83, 86.

Agathyrsus, iii. 7.

Agbal, iv. 70.

Agbatana, 1. Syrian, ii. 380; 2. Median, i. 191, 202, 234; ii. 382.

Age, respect paid to, ii. 112.

Agenor, iv. 67.

Agesilaüs, 1. a Spartan king of the upper house, iv. 139; 2. a Spartan king of the lower house, 298.

Agetus, iii. 369.

Agidæ, family tree of, iii. 280; iv. 315.

Agis, 1. grandfather of Leotychides, iii. 371; 2. ancestor of Leonidas, iv. 140.

Aglaurus, sanctuary of, iv. 250.

Aglomachus, iii. 118.

Aleuadæ, iv. 5, 91, 119.

Aleuas, iv. 351.

Alexander, son of Amyntas, his wealth, iii. 187; destroys the Persian embassy, 189; gives his sister in marriage to Bubares, 189; contends at Olympia, 190; advises the Greeks to retire from Tempe, iv. 120; his statue at Delphi, 293; goes as Persian ambassador to Athens, 302; his address to the Athenians, 305; failure of his mission, 308; communicates Persian plans to the Greeks, 344.

———, son of Priam, *i.e.* Paris, his rape of Helen, i. 124; arrival in Egypt, ii. 158; arrest by Thonis, 159; not at Troy during the siege, 162.

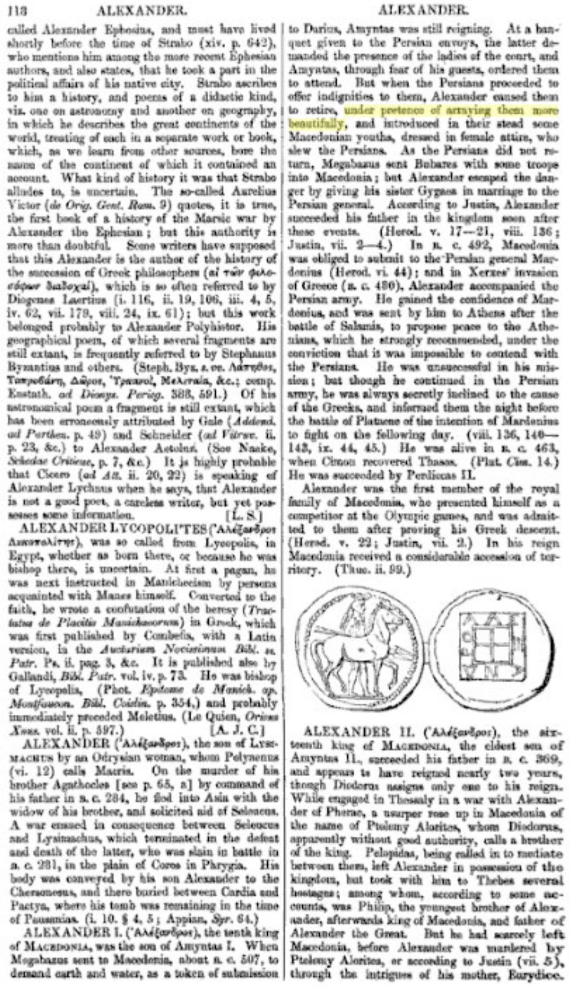
Ali-Allahis of Persia, i. 218.

Alilat, ii. 336.

Alitta, i. 217.

Alopecæ, iii. 220.

Figure 21. Page from vol. 1 of Smith’s *Dictionary of Greek and Roman Biography* (1848).



113 ALEXANDER. ALEXANDER.

called Alexander Ephesius, and must have lived shortly before the time of Strabo (xiv. p. 642), who mentions him among the more recent Ephesian authors, and also states, that he took a part in the political affairs of his native city. Strabo ascribes to him a history, and poems of a didactic kind, viz. one on astronomy and another on geography, in which he describes the great continents of the world, treating of each in a separate work or book, which, as we learn from other sources, bore the name of the continent of which it contained an account. What kind of history it was that Strabo alludes to, is uncertain. The so-called Aurelius Victor (*de Orig. Gent. Rom.* 9) quotes, it is true, the first book of a history of the Marsic war by Alexander the Ephesian; but this authority is more than doubtful. Some writers have supposed that this Alexander is the author of the history of the succession of Greek philosophers (*αἱ τῶν φιλοσόφων διαδοχαί*), which is so often referred to by Diogenes Laertius (i. 116, ii. 19, 106, iii. 4, 5, iv. 62, vii. 179, viii. 24, ix. 61); but this work belonged probably to Alexander Polyhistor. His geographical poem, of which several fragments are still extant, is frequently referred to by Stephanus Byzantius and others. (Steph. Byz. *s. vv.* Λάπηθος, Ταπροβάνη, Δῶρος, Ὑρκανοί, Μελιταία, &c.; comp. Eustath. *ad Dionys. Perieg.* 388, 591.) Of his astronomical poem a fragment is still extant, which has been erroneously attributed by Gale (*Addend. ad Parthen.* p. 49) and Schneider (*ad Vitruv.* ii. p. 23, &c.) to Alexander Aetolus. (See Naeke, *Schedae Criticae*, p. 7, &c.) It is highly probable that Cicero (*ad Att.* ii. 20, 22) is speaking of Alexander Lychnus when he says, that Alexander is not a good poet, a careless writer, but yet possesses some information. [L. S.]

ALEXANDER LYCOPOLITES ('Αλέξανδρος Λυκοπολίτης), was so called from Lycopolis, in Egypt, whether as born there, or because he was bishop there, is uncertain. At first a pagan, he was next instructed in Manicheeism by persons acquainted with Manes himself. Converted to the faith, he wrote a confutation of the heresy (*Tractatus de Placitis Manichaeorum*) in Greek, which was first published by Combefis, with a Latin version, in the *Auctarium Novissimum Bibl. ss. Patr.* Ps. ii. pag. 3, &c. It is published also by Gallandi, *Bibl. Patr.* vol. iv. p. 73. He was bishop of Lycopolis, (Phot. *Epitome de Manich. ap. Montfaucon. Bibl. Coislin.* p. 354,) and probably immediately preceded Meletius. (Le Quien, *Oriens Xnus.* vol. ii. p. 597.) [A. J. C.]

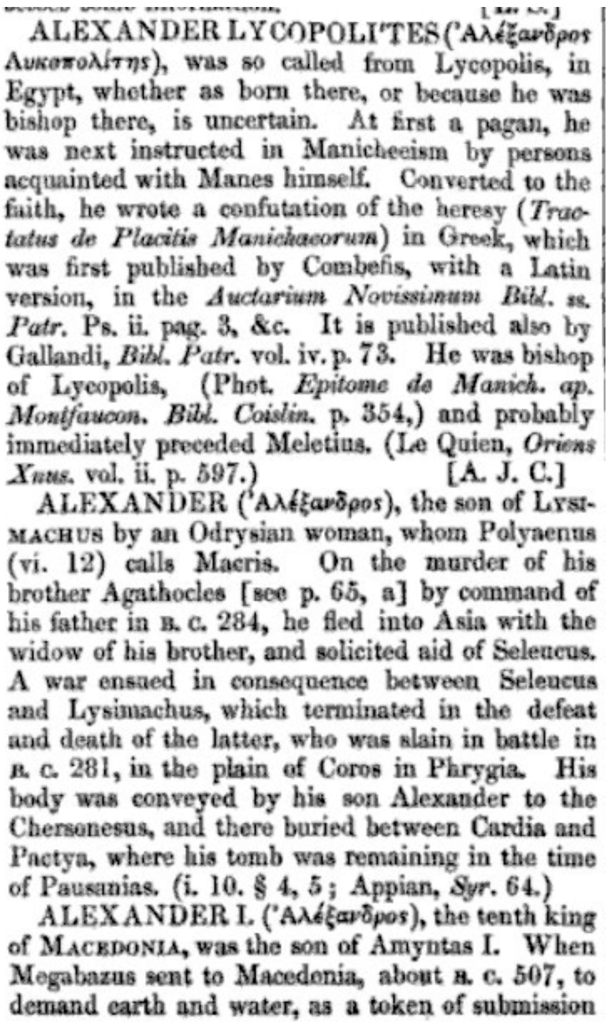
ALEXANDER ('Αλέξανδρος), the son of LYSIMACHUS by an Odrysian woman, whom Polyaenus (vi. 12) calls Macris. On the murder of his brother Agathocles [see p. 65, a] by command of his father in B. C. 284, he fled into Asia with the widow of his brother, and solicited aid of Seleucus. A war ensued in consequence between Seleucus and Lysimachus, which terminated in the defeat and death of the latter, who was slain in battle in B. C. 281, in the plain of Corus in Phrygia. His body was conveyed by his son Alexander to the Chersonesus, and there buried between Cardia and Pactya, where his tomb was remaining in the time of Pausanias. (i. 10. § 4, 5; Appian, *Syr.* 64.)

ALEXANDER I. ('Αλέξανδρος), the tenth king of MACEDONIA, was the son of Amyntas I. When Megabazus sent to Macedonia, about B. C. 507, to demand earth and water, as a token of submission to Darius, Amyntas was still reigning. At a banquet given to the Persian envoys, the latter demanded the presence of the ladies of the court, and Amyntas, through fear of his guests, ordered them to attend. But when the Persians proceeded to offer indignities to them, Alexander caused them to retire, under pretence of arraying them more beautifully, and introduced in their stead some Macedonian youths, dressed in female attire, who slew the Persians. As the Persians did not return, Megabazus sent Bubares with some troops into Macedonia; but Alexander escaped the danger by giving his sister Gygaea in marriage to the Persian general. According to Justin, Alexander succeeded his father in the kingdom soon after these events. (Herod. v. 17—21, viii. 136; Justin, vii. 2—4.) In B. C. 492, Macedonia was obliged to submit to the Persian general Mardonius (Herod. vi. 44); and in Xerxes' invasion of Greece (B. C. 480), Alexander accompanied the Persian army. He gained the confidence of Mardonius, and was sent by him to Athens after the battle of Salamis, to propose peace to the Athenians, which he strongly recommended, under the conviction that it was impossible to contend with the Persians. He was unsuccessful in his mission; but though he continued in the Persian army, he was always secretly inclined to the cause of the Greeks, and informed them the night before the battle of Plataeae of the intention of Mardonius to fight on the following day. (viii. 136, 140—143, ix. 44, 45.) He was alive in B. C. 463, when Cimon recovered Thasos. (Plut. *Cim.* 14.) He was succeeded by Perdiccas II.

Alexander was the first member of the royal family of Macedonia, who presented himself as a competitor at the Olympic games, and was admitted to them after proving his Greek descent. (Herod. v. 22; Justin, vii. 2.) In his reign Macedonia received a considerable accession of territory. (Thuc. ii. 99.)

ALEXANDER II. ('Αλέξανδρος), the sixteenth king of MACEDONIA, the eldest son of Amyntas II., succeeded his father in B. C. 369, and appears to have reigned nearly two years, though Diodorus assigns only one to his reign. While engaged in Thessaly in a war with Alexander of Pherae, a usurper rose up in Macedonia of the name of Ptolemy Alorites, whom Diodorus, apparently without good authority, calls a brother of the king. Pelopidas, being called in to mediate between them, left Alexander in possession of the kingdom, but took with him to Thebes several hostages; among whom, according to some accounts, was Philip, the youngest brother of Alexander, afterwards king of Macedonia, and father of Alexander the Great. But he had scarcely left Macedonia, before Alexander was murdered by Ptolemy Alorites, or according to Justin (vii. 5), through the intrigues of his mother, Eurydice.

Figure 22. Detail from Smith's



ALEXANDER LYCOPOLITES (Ἀλέξανδρος Λυκοπολίτης), was so called from Lycopolis, in Egypt, whether as born there, or because he was bishop there, is uncertain. At first a pagan, he was next instructed in Manicheeism by persons acquainted with Manes himself. Converted to the faith, he wrote a confutation of the heresy (*Tractatus de Placitis Manichaeorum*) in Greek, which was first published by Combefis, with a Latin version, in the *Auctarium Novissimum Bibl. ss. Patr.* Ps. ii. pag. 3, &c. It is published also by Gallandi, *Bibl. Patr.* vol. iv. p. 73. He was bishop of Lycopolis, (Phot. *Epitome de Manich. ap. Montfaucon. Bibl. Coislin.* p. 354,) and probably immediately preceded Meletius. (Le Quien, *Oriens Xnus.* vol. ii. p. 597.) [A. J. C.]

ALEXANDER (Ἀλέξανδρος), the son of LYSIMACHUS by an Odrysian woman, whom Polyaenus (vi. 12) calls Macris. On the murder of his brother Agathocles [see p. 65, a] by command of his father in B. C. 284, he fled into Asia with the widow of his brother, and solicited aid of Seleucus. A war ensued in consequence between Seleucus and Lysimachus, which terminated in the defeat and death of the latter, who was slain in battle in B. C. 281, in the plain of Coros in Phrygia. His body was conveyed by his son Alexander to the Chersonesus, and there buried between Cardia and Pactya, where his tomb was remaining in the time of Pausanias. (i. 10. § 4, 5; Appian, *Syr.* 64.)

ALEXANDER I. (Ἀλέξανδρος), the tenth king of MACEDONIA, was the son of Amyntas I. When Megabazus sent to Macedonia, about B. C. 507, to demand earth and water, as a token of submission

Figure 23. Detail from the article on Alexander I from Smith's Dictionary above.

Notes

[1] This number can be found in [Lavoie 2005].

[2] See http://googleblog.blogspot.com/2008/10/new-chapter-for-google-book-search.html.

[3] Workshops took place at the University of Chicago (November 2006), Tufts University (May 2007), the Council for Library and Information Resources (Washington, DC, November 2007), Imperial College London (March 2008) and Humboldt University in Berlin (March 2008).

[4] [Crane 2006f]; [Crane 2006]; [Crane 2008]

[5] http://www.tlg.uci.edu/, accessed October 19, 2008.

[6] [Crane 2006b]

[7] [Smith 2001]; [Crane 2006a]

[8] For a list of ARL members, see http://www.arl.org/arl/membership/members.shtml, accessed August 25, 2008. For statistics from 2005-06, see [Kyrillidou 2008].

[9] http://publicaccess.nih.gov, accessed August 25, 2008.

[10] http://nih.gov/about/index.html, accessed August 25, 2008, listed an NIH budget of $27.8 billion dollars and stated that "NIH distributes 80% of its funding to research grants."

[11] This work with classical Greek and OCR has been reported in [Stewart 2007].

[12] http://www.annee-philologique.com/aph/, accessed August 25, 2008.

[13] The number of publications indexed was taken from http://citeseer.ist.psu.edu/, accessed August 25, 2008. For basic information on CiteSeer, see http://citeseer.ist.psu.edu/citeseer.html, accessed August 25, 2008. For the initial description of the CiteSeer system, see [Bollacker 1998].

[14] The challenges of corpus design and representativeness have been explored by many authors, including [Biber 1993] and [Douglas 2003].

[15] Research into document image analysis and retrieval within historical digital libraries is a growing area of research, for example see [Marinai 2007].

[16] For some promising work in this area please see [Faure 2007].

[17] For some recent state-of-the-art work in OCR for historical text collections, please see [Reynaert 2008].

[18] For more on this research area, see for example, [Ramel 2007].

[19] Recent successful efforts in extracting structural markup on a large scale from volumes within the OCA have been reported by [Lu 2007].

[20] See [Schmid 2008] and [Fitschen 2008].

[21] For a recent exploration of text mining in humanities documents, please see [Don 2007].

[22] The automatic markup of humanities texts with relevant ontologies (such as CIDOC-CRM) has a large and varied body of research, for some recent discussions please see [Doerr 2008] and [Lin 2008].

[23] Making of America: http://quod.lib.umich.edu/m/moagrp/; http://cdl.library.cornell.edu/moa/, accessed October 20, 2008. JSTOR: http://www.jstor.org/, accessed October 20, 2008.

[24] [Hall 1913].

[25] Most surviving classical Latin was composed after antiquity. Johannes Ramminger had, as of 2008, assembled more than 200 million words of Latin in digital form (http://www.neulatein.de/, accessed October 19, 2008). The Thesaurus Linguae Latinae (TLL) is based on an archive of 10 million slips, which contain, for the older texts, a slip for each occurrence of a word (http://www.thesaurus.badw.de/english/index.htm, accessed October 19, 2008). The Packard Humanities Institute CD ROM of Latin, which is fairly comprehensive through 200CE and contains some later materials contains c. 7.5 million words.

[26] [Cheng 2001]; [Cheng 2002]

[27] Google Books offers a “popular passages” feature that seeks to identify and link quotations, work that was recently reported in [Schilit 2008].

[28] http://www.eaqua.net/, accessed October 20, 2008

Works Cited

Biber 1993 Biber, Douglas. “Representativeness in Corpus Design”. *Literary and Linguistic Computing* 8: 4 (1993), pp. 243-257.

Bollacker 1998 Bollacker, Kurt D., Steve Lawrence and C. Lee Giles. “CiteSeer: an Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications”. Presented at *AGENTS 1998*. *Proceedings of the Second International Conference on Autonomous Agents* (1998), pp. 116-123.

Cheng 2001 Cheng, Jiun Yuan, and W. Brent Seales. “Guided Linking: Efficiently Making Image-to-Transcript Correspondence”. Presented at *JCDL 2001*. *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (2001).

Cheng 2002 ChengJ. Y. *Extensible Tools for Building and Using Digital Library Collections*. Thesis, University of Kentucky Department of Computer Science: 2002.

Crane 2006 CraneGregory. *Comments on the APA Task For on Electronic Publications: Issues and Recommendations for Discussion (draft of October 20, 2006).* http://www.stoa.org/varia/apacomments.pdf.

Crane 2006a Crane, Gregory, and Alison Jones. *The Perseus American Collection 1.0*. 2006. http://dl.tufts.edu//view_pdf.jsp?urn=tufts:facpubs:gcrane-2006.00001.

Crane 2006b Crane, Gregory. “The Perseus Digital Library: New content and services for 19th century American documents”. *D-Lib Magazine* 12: 3 (March 2006).

Crane 2006f Crane, Gregory. “What Do You Do with A Million Books?”. *D-Lib Magazine* 12: 3 (2006). http://www.dlib.org/dlib/march06/crane/03crane.html.

Crane 2008 Warning: Biblio formatting not applied. CraneGregory. AmyFriedlander. *Many More than a Million: Building the Digital Environment for the Age of Abundance . Report of a One-Day Seminar on Promoting Digital Scholarship Sponsored by the Council on Library and Information Resources. November 28, 2007 Final Report.* March 1, 2008. http://www.clir.org/activities/digitalscholar/Nov28final.pdf.

Doerr 2008 Doerr, Martin, and Dolores Iorizzo. “The Dream of a Global Knowledge Network -- A New Approach”. *Journal of Computing and Cultural Heritage* 1: 1 (2008), pp. 1-23.

Don 2007 Don, Anthony, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman and Catherine Plaisant. “Discovering Usage Patterns in Text Collections: Integrating Text Mining with Visualization”. Presented at *CIKM 2007*. *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management* (2007), pp. 213-222. http://hcil.cs.umd.edu/trs/2007-08/2007-08.pdf.

Douglas 2003 Douglas, Fiona M. “The Scottish Corpus of Texts and Speech: Problems of Corpus Design”. *Literary and Linguistic Computing* 18: 1 (2003), pp. 23-37.

Faure 2007 Faure, Claudie, and Nicole Vincent. “Document Image Analysis for Active Reading”. Presented at *SADPI 2007*. *Proceedings of the 2007 International Workshop on Semantically Aware Document Processing and Indexing* (2007), pp. 7-14.

Fitschen 2008 Fitschen, Arne, and Piklu Gupta. “Lemmatising and Morphological Tagging”. In Anke Lüdeling and Merja Kytö, eds., *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, 2008.

Hall 1913 Hall, F.W. *A Companion to Classical Texts*. Oxford: Clarendon Press, 1913.

Kyrillidou 2008 KyrillidouM., and M. Young. *ARL Statistics 2005-06: a compilation of statistics from the one hundred and twenty-three members of the Association of Research Libraries*. http://www.arl.org/bm~doc/arlstats06.pdf.

Lavoie 2005 Lavoie, Brian, Lynn Silipigni Connaway and Lorcan Dempsey. “Anatomy of Aggregate Collections: The Example of Google for Print Libraries”. *D-Lib Magazine* 11: 9 (2005). http://www.dlib.org/dlib/september05/lavoie/09lavoie.html.

Lin 2008 Lin, Chia-Hung, Jen-Shin Hong and Martin Doerr. “Issues in an Inference Platform for Generating Deductive Knowledge: a Case Study in Cultural Heritage Digital Libraries using the CIDOC-CRM”. *International Journal on Digital Libraries* 8: 2 (2008), pp. 115-132.

Lu 2007 Lu, Xiaonan, James Z. Wang and C. Lee Giles. “Intelligent Parsing of Scanned Volumes for Web Based Archives”. Presented at *ICSC 2007*. *Proceedings of the International Conference on Semantic Computing* (2007), pp. 559-568.

Marinai 2007 Marinai, Simone, Emanuele Marino and Giovanni Soda. “Exploring Digital Libraries with Document Image Retrieval”. Presented at *ECDL 2007*. *Proceedings of the Conference on Research and Advanced Technology for Digital Libraries* (2007), pp. 368-379.

Mimno 2007 Mimno, David, and Andrew McCallum. “Mining a Digital Library for Influential Authors”. Presented at *JCDL 2007*. *Proceedings of the 2007 Joint ACM-IEEE Conference on Digital Libraries* (2007), pp. 105-106. http://www.cs.umass.edu/~mccallum/papers/authors-jcdl07.pdf.

Ramel 2007 Ramel, Jean-Yves, S. Leriche, M.L. Demonet and S. Busson. “User-Driven Page Layout Analysis of Historical Printed Books”. *International Journal on Document Analysis and Recognition* 9: 2-4 (2007), pp. 243-261.

Reynaert 2008 Reynaert, Martin. “Non-Interactive OCR Post-Correction for Giga-Scale Digitization Projects”. Presented at *COLING 2008*. *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing* (2008), pp. 617-630.

Schilit 2008 Schilit, Bill N., and Okan Kolak. “Exploring a Digital Library through Key Ideas”. Presented at *JCDL 2008*. *Proceedings of the Eighth ACM/IEEE-CS Joint Conference on Digital Libraries* (2008), pp. 177-186.

Schmid 2008 Schmid, Helmut. “Tokenizing and Part-of-Speech Tagging”. In Anke Lüdeling and Merja Kytö, eds., *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, 2008.

Smith 2001 Smith, David A., and Gregory Crane. “Disambiguating Geographic Names in a Historical Digital Library”. Presented at *ECDL 2001*. *Proceedings of the Fifth European Conference on Research and Advanced Technology for Digital Libraries* (2001), pp. 127-136. http://perseus.mpiwg-berlin.mpg.de/Articles/geodl01.pdf.

Stewart 2007 Stewart, Gordon, Gregory Crane and Alison Babeu. “A New Generation of Textual Corpora: Mining Corpora from Very Large Collections”. Presented at *JCDL*. *Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital Libraries* (2007), pp. 356-365. http://www.cs.princeton.edu/~jsseven/papers/corpora/corpora.pdf.