Preprint of paper accepted to JCDL 2012.

Student Researchers, Citizen Scholars and the Trillion Word Library

Gregory Crane, Bridget Almas, Alison Babeu, Lisa Cerrato Matthew Harrington
Tufts University
Perseus Project

{gregory.crane,bridget.almas, alison.babeu,lisa.cerrato, matthew.harrington}@tufts.edu

David Bamman
Carnegie Mellon University
School of Computer Science
Language Technologies Inst.

dbamman@cs.cmu.edu

Harry Diakoff
Alpheios Project

Alpheios.net

Harry.Diakoff@gmail.com

ABSTRACT

The surviving corpora of Greek and Latin are relatively compact but the shift from books and written objects to digitized texts has already challenged students of these languages to move away from books as organizing metaphors and to ask, instead, what do you do with a billion, or even a trillion, words? We need a new culture of intellectual production in which student researchers and citizen scholars play a central role. And we need as a consequence to reorganize the education that we provide in the humanities, stressing participatory learning, and supporting a virtuous cycle where students contribute data as they learn and learn in order to contribute knowledge. We report on five strategies that we have implemented to further this virtuous cycle: (1) reading environments by which learners can work with languages that they have not studied, (2) feedback for those who choose to internalize knowledge about a particular language, (3) methods whereby those with knowledge of different languages can collaborate to develop interpretations and to produce new annotations, (4) dynamic reading lists that allow learners to assess and to document what they have mastered, and (5) general e-portfolios in which learners can track what they have accomplished and document what they have contributed and learned to the public or to particular groups.

Categories and Subject Descriptors

H.3.7. [**Information Systems**]: Information Storage and Retrieval: digital libraries.

General Terms

Documentation, Design, Human Factors,

Keywords

Automatic linking, collection development, document design, reading, browsing.

1. INTRODUCTION

Even deeply traditional disciplines such as the study of Greek and Latin language and literature have begun to shift the infrastructure upon which they depend from a focus primarily upon books and written artifacts to one that includes the words that these artifacts preserve. Every preserved word (or fragment) in every edition, manuscript, inscription, and papyrus,—in every surviving document produced in Greek or Latin—is now an object of interest that must possess a unique identifier. Each word then becomes the object for annotations that are multiple in the present (e.g., morphology, dictionary entry, link from transcription of region of written object) and that will increase in number as automated systems and human annotators identify an increasing range of phenomena. Even a few thousand books generate datasets with billions of objects. Traditional library systems, focused upon books and catalogue records, are inadequate, if not wholly irrelevant, to current, emerging research and instruction.

This shift to words and data demands not only a transformation in the infrastructure of earlier humanities work but has already begun to provoke a more general change in the culture of intellectual production and scholarship [13, 27]. Advanced researchers and library professionals must enlist student researchers and citizen scholars as key partners if they are to manage the quantity and range of historical sources already available under open licenses to the net public. Academic institutions must also develop correspondingly new collaborative and participatory models of education, where students contribute new knowledge as they learn and then learn so that they can contribute more new knowledge.

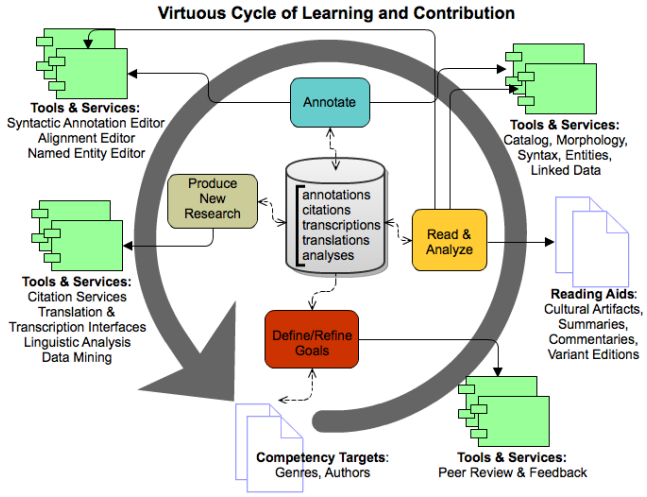


Figure 1: Virtuous cycle of contributing and learning

We focus upon how the production of annotations and internalization of linguistic knowledge can reinforce each other creating a virtuous cycle of contribution and learning (Figure 1).

JCDL'12, June 10–14, 2012, Washington, DC, USA.

Students produce data as they learn and they learn by producing useful data of increasing complexity. Figure 2 illustrates an implementation of the Son of Suda Online (SoSOL) editing environment [4], customized to support distributed editing for the TEI XML sources available from Perseus, with editorial workflows and mechanisms for contributors to document what they have contributed.

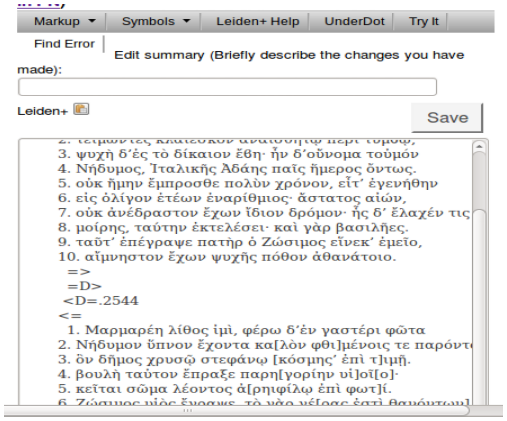


Figure 2: Editorial workflow for corrections, XML markup

Learners can thus know that they receive named credit for contributions to historical corpora that have been studied for many years—and in some cases for millennia—and will be the object of analysis for generations to come. At the same time, some categories of annotation—and often precisely those categories for which automated methods are least effective—not only depend upon, but also develop, linguistic mastery.

In the following section this paper provides background about the shift from books and written objects to words and associated annotations as the primary objects upon which infrastructure must focus. In section 3, we describe implementations of the five strategies of the aforementioned virtuous cycle. The final three sections cover related work, future development and the conclusions.

2. FROM BOOKS TO WORDS

Historical languages such as Classical Greek and Latin represent relatively finite corpora—language analysis determined that 22,000 of the first 1.2 million books extracted from the Internet Archive were primarily in Latin and contained 2.7 billion words [6]. The default OCR output for Classical Greek is too noisy to support such analysis but the Hathi Trust reports that only 8,600 of the 10 million books already in the collection are in Classical Greek.[1] For most students of these languages, however, the books are not the objects of primary interest. The words in the texts—if not the morpho-graphemic components—are the primary objects of interest. A collection with 30,000 books is small in conventional terms but even this modest collection represents a dataset of more than 3 billion words, each of which is the object of multiple annotations. Information infrastructures that focus on the books rather than upon the words that they contain cannot support either the research or pedagogy already taking shape.

As more of the human record becomes available in digital form,—representing not only Greek and Latin, but also Classical Arabic, Sanskrit, Chinese, Persian, the Cuneiform languages of the Middle East, Egyptian from hieroglyphics through Coptic, and the medieval dialects of Europe,—that figure will grow into the hundreds of billions. If we focus on the period before sound recording, for which only written sources preserve textual and linguistic data, then the figure may exceed one trillion words.

Table 1: Overview of open data for Classical Greek and Latin

	Classical Greek	*Latin*
MS pages, papyri, inscriptions	10,000	5,000
MSS metadata	720 MSS	340 MSS
Page Images	170,000	6,600,000
OCR output	50 million	2.7 billion
Transcribed Words	10 million	7 million
TEI XML tags	860,000	787,000
FRBR Authors	841	536
FRBR Works	2,186	1,042
Citations into	733,000	512,000
Citations within	265,000	185,000
Part of speech	39 million	2.1 billion
Morphology	40 million	2.2 billion
Dictionary words	125,000	62,000
Word senses	162,000	102,000
Treebank words	300,000	50,000
Grammatical categories	3,300	650
Named entities	580,000	352,000
Textual variants	24,000	51,000
Aligned English words	5 million	3 million

Table 1 provides some general figures about currently available open content Greek and Latin sources (primarily from Perseus, Papyri.info, and the Homer Multitext Project). The table contains a number of figures that may not be immediately clear (e.g., "*citations into*" describes citations from secondary sources to primary sources, "*citations within*" describes primary sources quoting each other, "*dictionary words"* and "*word senses"* enumerate categories while "*OCR output*" enumerates running words) but the point of the table is to demonstrate that the shift from books to words is not a future abstraction but a pressing reality with which students of historical languages are already grappling (and for which conventional library infrastructures are largely irrelevant). Collections such as those listed in Table 1 demand services that respond to at least six basic features.

(1) **Granularity**: Ultimately, we may need to focus upon the morpho-graphemic units out of which words are composed rather than on the words themselves, but a shift from cataloging books to annotating words provides an adequate—and dramatic—first step.

(2) **Scale**: once we move from cataloging physical artifacts to tracking the words within those artifacts, our objects of interest immediately increase by a factor of at least four orders of magnitude—from millions of books to tens of billions of words. At the same time, any arbitrary combination of these words can constitute an object of interest that must be recorded. Objects in the opening words of Vergil's *Aeneid* (arma virumque cano, "arms and the man I sing") include not only (1) *Arma,* (2) *virumque,* and (3) *cano*, but also (4) *virum* and (5) *-que* (a word that comprises two separable dictionary entries), and combinations of words such as (6) *arma virumque*, (7) *arma … cano*, (8) *virum … cano*. These combinations are fundamental to the way we talk about texts and thus the number of objects that we must be able to address is closer in scale to the set of sentences that we can generate from a language.

(3) **Inter-textuality**: The billions of words from the human record already available in digital form are not simply a stream of sources that were produced independently and in sequence over

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

time (where in fact we know the sequence, which we often do not). We have multiple versions of the same source—sometimes thousands of versions, when we consider every quotation of the Bible or Shakespeare as a separate version with its own context. We need to be able to track textual sources as they evolve, in some cases over thousands of years [31]. Large collections may have hundreds or thousands of versions of *Hamlet* and hundreds of thousands of passages that quote sections of the Bible.

(4) **Hyper-linguality**: Human beings can be multilingual and converse fluently in various languages. Traditional research problems are challenging enough. For instance, Aristotle's works circulated in Greek, Syriac translation, Arabic translation from the Syriac, Latin translation from the Arabic. Then, Aristotle (and Euclid) re-entered Europe and helped spark a process that led to the Renaissance and the formulation of the modern western world. With collections such as Google Books that contain more than 400 languages and with nearly 7,000 languages spoken today, we live in a hyper-lingual digital space where we cannot hope to study, much less master more than a tiny fraction of the languages that surround us. Our infrastructure must relentlessly provide increasingly accurate and increasingly sophisticated services whereby we can trace ideas across the full linguistic breadth of the human record [26].

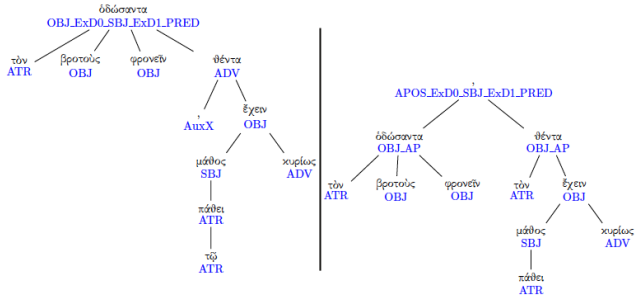
(5) **Dimensionality**: the words in our collections have multiple features, each of which can depend upon the particular context in which a particular word appears. Some of these features—such as dictionary entry and word sense, morphological features, syntactic function, name class (Washington as place vs. person) and identification (Washington State vs. Washington, DC)—are familiar from print but researchers in corpus linguistics and other fields have developed new and growing sets of features, each of which can affect the potential dimensionality of any word—and the relationship between subsets of words—in our collections. Each of the rows in Table 1 constitutes a separate class of annotation, each of which can have multiple components of data.

(6) **Materiality**: each written word preserved from the analog record reflects an interpretation, generated by human or machine, from a written object. Where those objects exist, they may be represented in multiple digital formats, including 2D and 3D and multiple spectra and resolutions (as is the case for carefully studied inscriptions and manuscripts) [3]. In each case our infrastructures need to map particular—and in some cases conflicting—transcriptions of the words or characters on subsets of preserved written objects.

Automated systems can provide many of the annotations to manage the six features outlined above but these systems in at least some cases produce too many errors for at least some activities. Advanced researchers and library professionals must engage a generation of student researchers and citizen scholars who can participate as critical partners, making credited contributions of tangible value that grow increasingly sophisticated, with advanced expertise representing space on a continuum that many can traverse given time and determination, whether or not they belong to formal academic programs.

3. ANNOTATION AS FOUNDATIONAL DATA STRUCTURE

Every element listed in Table 1 is an annotation, i.e., a labeled link from one object to another. Text generated by OCR or transcribed by humans constitutes a stream of annotations on regions of a written object. Translations consist of annotations upon subsets of the source texts, as do morphological and syntactic analyses, links from words to dictionary entries and word senses pertaining to a particular context, textual notes on variants and conjectures, the classification of named entities and the association of those named entities with external authority lists, and representations of inter-textual re-use/allusion. Nor are these annotations simple assertions of fact. Figure 3 shows two competing interpretations of the same sentence, originally produced decades ago but now represented as machine actionable annotations [8]. We need to manage not only multiple classes of annotation but multiple interpretations as well. Annotations thus are a foundational data structure for historical corpora and provide a common thread for the work described below.



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

Figure 3: Machine actionable interpretations

3.1 Annotations and Reading Environments

When the right kinds of annotations accumulate for particular texts, the realizable value of those texts can change radically. **First**, on the macro-level, automated systems can perform new operations and generate new services—much of the multilingual software now in use, for example, depends upon links between aligned source texts and translations. Corpus linguists have begun to transform our understanding of language by studying annotated corpora and these methods are—or must be—foundational to any modern philology: all students of historical languages are conducting corpus linguistics.

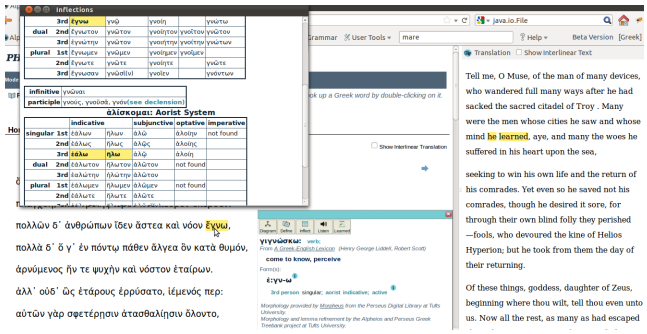


Figure 4: Alpheios.net annotation-based reading environment

Second, Figures 4 and 5 illustrate an environment developed by Alpheios.net, that integrates morphological analyses and syntactic analyses, automated dictionary lookups, and translations aligned to the source text to provide intensive reading support. In this environment, readers can quickly begin to make direct use of sources in languages that they have not studied—a substantive and tangible transformation in intellectual reach. Such intensive reading support has been studied in Chinese [2] and is established practice in languages such as Classical Greek and Latin for which collections, services, and annotations have evolved over decades.

Reading by looking up morphology, syntax, and translation equivalents for every word and corresponding section of a

translation is, of course, not the same as interacting directly with a language one has mastered. It is, however, a quantum leap beyond staring at inert symbols that convey no meaning at all. We can never again just shrug and say, "it's Greek to me," if we confront a source in an unknown language, certainly not if that source is in Classical Greek.

We can, however, only provide dense, curated annotations for subsets of the human record, with ever richer annotation covering progressively smaller subsets of available corpora. Thus, we have very large sets of text generated by OCR software (e.g.. 2.8 billion words of Greek and Latin), subsets that have been manually corrected or keyed in (e.g., 17 million words of Greek and Latin with TEI XML structural markup), and intensively analyzed corpora (e.g., the 350,000 words with curated morphological and syntactic analyses in the Perseus Greek and Latin Treebanks).

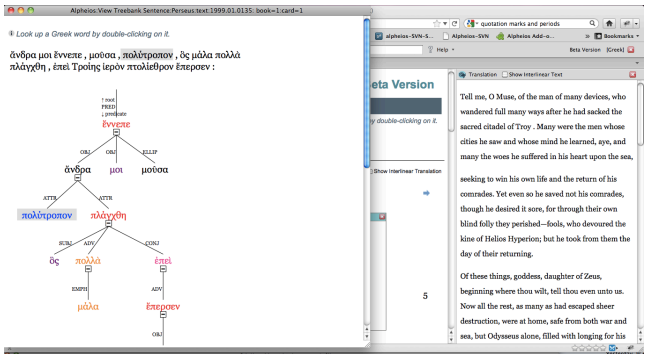


Figure 5: A treebank view in Alpheios.net

The interlinear visualization of aligned Greek and English versions of the Homeric *Odyssey* (figure 6), illustrates how readers can annotate English translations with structural markup or named entity identifiers (e.g., "Washington" in a given context is George Washington), then have these annotations applied to the original language source text [7]. Figure 2 illustrates an environment where introductory Greek students can not only fix typos but also help analyze and annotate the structure of complex manuscripts [12].

3.2 Annotation and Language Learning

Annotation is not only an essential chore as we organize historical sources. The act of annotation also provides new opportunities for learning the historical languages in which those sources are preserved, addressing in a fundamentally novel way the critical need for rapid and on-going feedback.

Unlike those learning modern languages, students of historical languages such as Classical Greek and Latin cannot receive immediate feedback by interacting with native speakers in a classroom or by visiting a foreign country. This presents a major barrier. Several experienced teachers of Greek and Latin in one recent departmental discussion, for example, went so far as to argue that independent reading of source texts was bad for students because they did not receive immediate feedback to call their mistakes and bad habits to their attention. While not all agree that the net result of independent reading was negative, there is virtually complete agreement that students reading source material require more and faster feedback.

For historical languages, access to pre-existing linguistic data thus opens up a major opportunity, because curated linguistic annotations provide training data not only for automated systems but for human learners as well. The same annotations by which we provide intensive reading support can, however, provide exhaustive questions by which learners can practice recognizing and generating morphology, syntax and (where parallel corpora are available) vocabulary in authentic source materials.

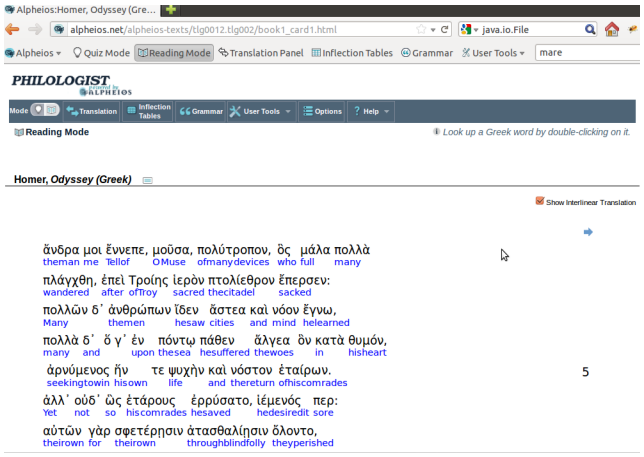


Figure 6: Interlinear view of aligned Greek/English texts

In first and second year language curricula students intensively study relatively small subsets of a language (e.g., corpora of c. 10,000 words and 2,000 vocabulary entries). Thus relatively small treebanks and/or aligned corpora can provide learners with enough curated training data to practice their ability to recognize and (by not showing selected forms) to generate the morphology and the syntax of a language. We can measure the fluency (which demands both the accuracy and the speed) with which learners can analyze and produce authentic language materials.

We then enable teams of learners to annotate new materials, comparing their independent analyses against each other and resolve differences by discussion or by a course instructor. Both processes provide far more accurate data than any morphological or syntactic analyzers.

Figure 7 illustrates differing interpretations of the same sentence from the *Res Gestae*[2] by two third semester students of Latin at Tufts University in the fall of 2011. Students in this class produced machine actionable interpretations of the morphological and syntactic functions of words from this text, adding to the machine actionable data by which researchers can fundamentally rebuild their understanding of these ancient languages.[3]

There are various methods whereby we can review contributions and reconcile differences where no canonical analyses are available. We can calculate the probability that students independently make the same errors and thus model the reliability of those instances where they agree. The two students can review their work together. A third student can annotate independently and then enter the discussion. The class can review the annotations. The instructor can review the annotations. Students then receive feedback on their performance. Or these methods can be combined, with students assigned the task of analyzing sentences and reviewing analyses by their peers before the

[2] The *Res Gestae* is official history published by the Emperor Augustus to celebrate his accomplishments and a major source for our understanding both of what happened and how Augustus wished posterity to understand what happened.

[3] Classes at Tufts and Holy Cross have already contributed to data for Homer and Petronius now published in the Greek and Latin Treebanks.

instructor then reviews the aggregate work. In this case, feedback is rapid, with students interacting with each other and the instructor entering to review overall results and to focus upon particularly difficult questions.

Princeps senatus fui usque ad eum diem quo scripseram haec per annos quadraginta .

index	word	New annotation (based on user 1)			user 2			
		head	relation	lemma + morph	head	relation	lemma	morph
0	Princeps	2	SBJ	noun sg masc nom		PNOM		
1	senatus	0	ATR	noun sg masc gen				
2	fui	-1	PRED	verb 1st sg perf ind act				
3	usque	2	ADV	adv				
4	ad	3	AuxP	prep	2			
5	eum	6	ATR	pron sg masc acc				
6	diem	4	OBJ	noun sg masc acc		ADV		
7	quo	8	SBJ	pron sg masc abl			qui1	
8	scripseram	6	ATR	verb 1st sg plup ind act				
9	haec	8	OBJ	pron pl neut acc				
10	per	8	AuxP	prep				
11	annos	10	OBJ	noun pl masc acc		ADV		

Figure 7: Two student interpretations of a source text

Results can then be presented in percentages. The list below illustrates the performance of selected annotators within the Greek Treebank on syntactical and morphological analyses.

```
<annotator n="10215" Syntax=".999" Morph=".889"/>
<annotator n="10090" Syntax=".959" Morph=".885"/>
<annotator n="10085" Syntax=".918" Morph=".885"/>
<annotator n="10079" Syntax=".918" Morph=".885"/>
<annotator n="10078" Syntax=".89"  Morph=".882"/>
```

As students contribute more data, their differing strengths and weaknesses begin to emerge. Figure 8, presented in two graphs, visualizes the overall performance of each student set against their ability to analyze particular phenomena. Most students are quite successful associating adjectives with nouns (the second graph: "attributive modifiers") but student ability to analyze the function of verbal participles (the first graph: "participial attachment") varies widely (user 12 has no idea how to perform this operation). Individual learners, their instructors, and automated tools can use such data to personalize subsequent questions and to reinforce phenomena that particularly challenge individual learners.

As students of the language become proficient, they can analyze new materials on their own, both to improve their understanding of the language and to contribute to human knowledge. We can solicit multiple analyses of the same sentence (or even the same word or phrase) from multiple contributors until shared agreements produce results of any realistic statistical reliability. If divergent answers suggest genuine ambiguity, we can flag that instance for manual review.

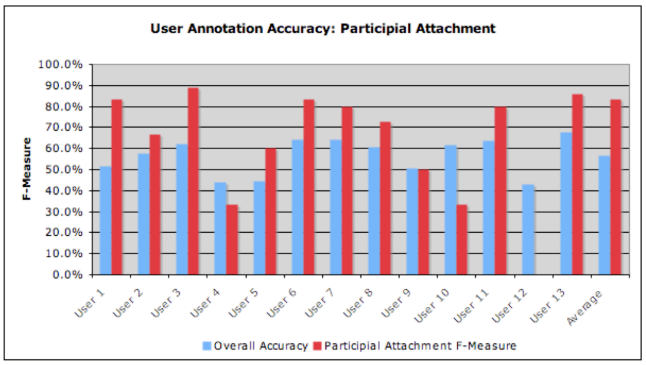


Figure 8a: Varying performance by students on two tasks

For Greek and Latin, the language communities remain substantial. Of the 1.2 million unique users who visited the Perseus Digital Library alone in Q4 of 2011, 120,000 worked directly with the Greek and Latin sources.

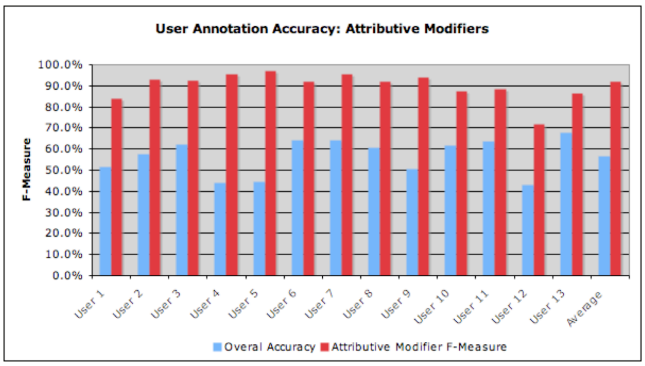


Figure 8b: Varying performance by students on two tasks

Contributions of this type are essential. Because no native speakers survive, annotating historical languages takes, in our experience, twice as long as the annotation of modern Chinese and with higher error rates—more than $1 per curated word when we pay for annotators and editors. Even if we can reduce these costs, we cannot expect to pay with external funding for a treebank representing 2000 years of Latin. The more learners practice, the more feedback they receive and the more proficient they become and the more they contribute. The act of learning, in this environment, expands and improves the scholarly data with which the learning began. Learners can thus receive the psychological benefits of contributing to something larger than themselves and of long term value—annotations to a historical source such as the *Res Gestae* will be used for years, if not generations. By combining two problems—the need to generate more linguistic data than we can fund and the need to provide better feedback for students of historical languages—we provided new methods by which to address both challenges at once.

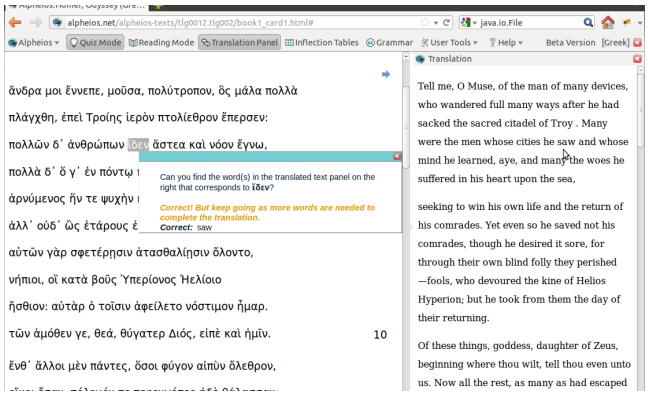


Figure 9: Aligned text as source for interactive quizzing

Access to structured data opens up a range of opportunities both to refine the results of automatic analyses and to use refined results as training data for human and machine learning alike. Figure 9 illustrates an exercise that builds upon aligned corpora. In this illustration, learners match words in the Greek to words in the English (the system is prompting the learner to match not only "saw" but "he saw"). Learners can go from English to Greek. We can also hide the corresponding Greek or English so that learners can practice generating language. Here we draw upon semantic analysis to allow for synonyms (e.g., student suggests "rock" and the translation contains "stone") to allow for multiple answers. The parallel text analysis not only allows learners new ways to reinforce their active and passive command of vocabulary, morphology and syntax but also enables new kinds of collaborative work, an example of which follows.

3.3 Annotations and Multilingual Learning

Various, emerging ways now permit students working together first, to explore problems that were intellectually inaccessible, not only to them but to most professional researchers, and then, to produce new data of lasting value as a by-product of their activities in a formal class.

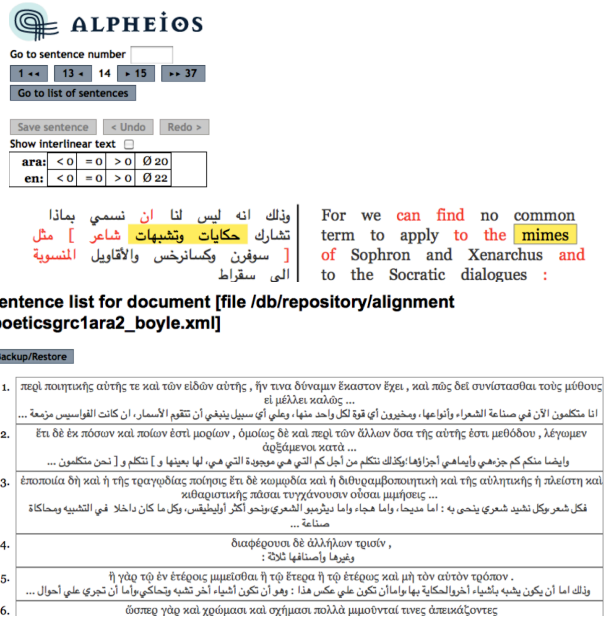


Figure 10: Collaborative alignment of Greek and Arabic

Students in a fall 2011 course at Tufts University, for example, studied the circulation of ideas from Greek sources, through Arabic translation and then back to Western Europe, via translation from Arabic into Latin, with a particular focus on Greek, Arabic, and Latin versions of Aristotle's *Poetics*. Three of the students had studied Greek and Latin, three were in third year Arabic, and one had studied both languages. In a take-home midterm exam, each group first used the Alpheios parallel text editor to align the words of the source text in the language with which they were most familiar (Greek or Arabic) with the words in an English translation. They then shifted from English to the language that, in most cases, they had not studied, using manual annotation and the morphological + dictionary lookup tools to explore the meaning of each word in the unfamiliar language.

There were two outcomes from this exercise. First, all the students were able to analyze, at a fine-grained level, the relationship between the Greek and the Arabic versions of Aristotle by using the aligned English (Figure 10) and the reading tools—an intellectual activity that would have been otherwise impractical and which allowed them to think about the circulation of ideas at a much greater level of precision than would have previously been possible. The Greek and Arabic sources were both immediately and directly accessible. Second, the students produced refined alignments from the Greek source to the Arabic translation, alignments that are stored as machine-actionable annotations that can support readers (who wish to see which words in the Greek correspond to the Arabic) and that contribute to a larger corpus of automatically aligned Greek and Arabic being developed by a project supported by the Mellon Foundation at Harvard and Tufts.

3.4 Annotations and Dynamic Reading Lists

We have developed a Dynamic Reading Lists service to expand the use and the value of reading lists at all levels. Dynamic Reading Lists support several functions that can generate annotations between citations and texts and between words and grammatical categories.

First, we need to measure the actual size of traditional reading lists. Such measurement is difficult when reading lists are static lists—we had to calculate the word counts for the Harvard undergraduate and Berkeley PhD reading lists. Another major PhD program estimated its new PhD reading lists in Greek and Latin at 1700 pages each—a very rough metric, since the content pages varies widely from publication series and genre. The reading list probably contains between 300,000 and 400,000 words for each language.

Support from the Google Digital Humanities program has allowed us to augment our holdings in Greek and Latin sources. We now can offer more than 95% of the sources covered in the published reading lists that we have analyzed.[4] These texts are encoded as TEI XML and, equally important, the structural markup represents the canonical citations—the chapter and verse style coordinate systems—whereby classicists have referenced and annotated their source texts for generations.

Second, we need to be able to convert citations that describe subsets of texts into machine actionable queries by which we can extract that section of a text that a citation describes. Because we have digitized versions of Classical Greek and Latin source texts in PhD reading lists, we can automatically calculate how many words are in books 1 and 6 of the *Iliad* or the *Medea* of Euripides. Identifying how many words are in particular subsets of a work (such as Herodotus' *Histories*, book 3, chapters 76-138), is also possible if we extract that particular chunk. The Dynamic Reading List service builds upon the Canonical Text Services protocol to resolve machine actionable citations such as "Herodotus 3.76-138" into a chunk of source text, extracting the precise citation.

Third, we need to be able to generate links between words and as many grammatical categories as possible so that learners understand what they need to know to comprehend a particular text or textual genre.

At present we can measure some general linguistic phenomena in corpora. Because we can generate links between inflected forms and their morphological features, we can, for example, detect that the optative mood—and with it, a long list of verbal conjugations—only occurs in a single form in the New Testament and thus help learners prioritize what they do and do not need to master for the corpora in which they are interested.

Where we have curated morpho-syntactic annotations in the Greek and Latin Treebanks we can provide far more precise analysis of the grammatical structures that learners will encounter. By using parallel texts (e.g., source texts in one language with translations into English), we can also begin to identify differing word senses (e.g., not only how often do forms of the Latin word *oratio* appear but how often do they correspond to a "speech" in a court room or political context and how often do they correspond to a "prayer" [5]). Learners can, as described above, refine automatic annotations of morphology, syntax, word sense and other features, providing (and receiving recognition for) more refined data for others to use.

[4] We have focused on the very large source texts that no individual or small group could digitize. The remaining sources are small texts, often fragmentary sources that survive only as scattered quotations in other authors and require special attention for proper digital publication [10].

Table 2: Cumulative words and vocabulary in a Dynamic Reading List, automatically calculated

Author	*Work*	*Cit.*	*Words*	*Vcb.*	*New Vcb.*
Homer	*Iliad*	bk 1	4519	**1172**	**1172** (25%)
Homer	*Iliad*	bk 9	5093	**1937**	**765** (15%)
Homer	*Odyssey*	bks 9-12	15973	**3585**	**1648** (10%)
Hesiod	*Theog.*	1-109	756	**3678**	**93** (12%)
Euripides	*Medea*	all	8105	**4642**	**964** (11%)
Thucydides	*Pel. War*	1.1-23	3522	**5045**	**403** (11%)
Thucydides	*Pel. War*	2.35-47	1933	**5200**	**155** (8%)
Plato	*Euthyph.*	all	5405	**5413**	**213** (3%)
Plato	*Apology*	all	8749	**5763**	**350** (4%)
Total			**54313**	**5763**	

We can also calculate one critical feature of a reading list, the number of distinct vocabulary entries in a given source text. Table 2 illustrates a Dynamic Reading List for a particular pathway through a Greek reading list. It calculates the number of words and vocabulary entries in a given reading and then calculates aggregate summaries for both. The total number of words covered increases linearly. But because readers encounter an increasing number of vocabulary items more than once, the total vocabulary increases more slowly. Readers who start by mastering the first book of the *Iliad* will actually have seen, on the average 75% (3,347) of the words they encounter as they work through the 4,519 words in this text, with a new vocabulary word appearing 25% of the time. As they move on to book 9 of the *Iliad* (the embassy to Achilles), new vocabulary drops to 15%, then to 10% when they move on to *Odyssey* books 9-12 (where Odysseus travels through the world of gods and monsters). Shifting to other authors introduces learners to different sets of vocabulary and the new vocabulary rate remains a bit higher (12%, 11%, 11%) until declining as the reader works through Greek prose—with only 4% of the words in Plato's *Apology* unseen.

We can recalculate these figures depending upon the paths that learners take. Thus, if a learner goes directly from Books 1 and 9 of the *Iliad* to Thucydides 1.1-23, 589 of the vocabulary items (16%) will be unseen as opposed to 403 (11%) for the learner who has worked through the path in Table 2.

We can also identify which vocabulary items will occur in a corpus. Table 3 shows what percentage of the vocabulary items in the Dynamic Reading List of Table 2 occur 5 or more times, 4 or more times etc. Learners (and, where formal course work is involved, their instructors) can prioritize vocabulary items based upon how often these words will occur in the particular sources that each learner chooses to explore—allowing the student of Greek philosophy to generate a list that differs from that of a student of Classical Greek historians or of the Homeric epics. Learners can review where they have encountered terms that they have forgotten.

We can also use this vocabulary data to retrieve unseen sentences that (1) reinforce new vocabulary items and (2) contain as many previously seen vocabulary items as possible (thus allowing the learning to reinforce both the new term and other known words), while automatically identifying, glossing and prioritizing any new words in the retrieved sentence. This service addresses a long-standing problem for assessment. We want to understand how well learners can recognize and use what they have encountered in the past in novel contexts, but it is impossible to determine with any precision what vocabulary items in a new passage are unseen and how often and when in the past learners have encountered those passages which they have seen. We now have the data so that learners can test themselves on new passages and so that we can generate smart exams that personalize themselves to the backgrounds of particular learners.

Table 3: Running words covered by vocabulary words

Frequency	*Cumulative*	*Words Covered*
5+	1,375	46,780 (86.4%)
4	1,694	48,056 (88.8%)
3	2,218	49,628 (91.7%)
2	3,180	51,552 (95.2%)
1	5,763	54,313 (100%)

Simple calculation of vocabulary items is not by itself a perfect measure of linguistic difficulty. The syntax of Thucydides is famous for its complexity—the Greek author Dionysius of Halicarnassus himself complained that he could not understand some passages from Thucydides—but the Thucydides' vocabulary is smaller than that of Homer or Plato. A student working through 75,000 words in each of these authors will encounter 3,966 vocabulary items in Thucydides, 4363 in Plato and 5,024 in Homer, even though Homeric Greek is commonly judged to be far easier to read than that of either Thucydides or Plato. We need additional metrics to capture other features of linguistic complexity when evaluate the difficulty of a corpus.

Nevertheless, even simple Dynamic Reading Lists such as that presented in Table 2 have immense potential. For the first time, we can immediately measure the absolute words and the number of vocabulary items in an arbitrary reading list, allowing us to compare the size and (by using other available metadata) chronological or genre coverage of various published reading lists.

More importantly, learners can establish their own dynamic reading lists, customized to help them learn the linguistic knowledge they need to pursue their own needs and interests. Learners can then publish these reading lists as a part of their e-portfolios. At present, a particularly ambitious student at a small program must depend upon their grades and upon the letters of their instructors to establish what he or she has accomplished. The independent learner has virtually no way to document what they have accomplished. Learners can, however, publish their dynamic reading lists as a part of their e-portfolios. They can also include any formal evaluation that they have received. Third parties can then use these dynamic reading lists to generate examination materials with which to evaluate the degree to which individual students have mastered their published lists.

The potential of published Dynamic Reading Lists is substantial. Learners are not so dependent upon the reputations of their home institutions and the grades that they have received if they can document what they have actually done. The Dynamic Reading Lists thus reward merit and provide concrete encouragement for ambitious learners, who have much better instruments whereby to establish their own accomplishments.

3.5 E-Portfolios, Linguistic Annotation, and the Shift from Authority to Achievement

Dynamic Reading Lists provide one particularly important feature for the e-portfolios of those who work with historical languages. To document what students have contributed and the linguistic

skills that they have acquired, we need to support the data structures by which we store this information. We have developed methods with which to manage both.

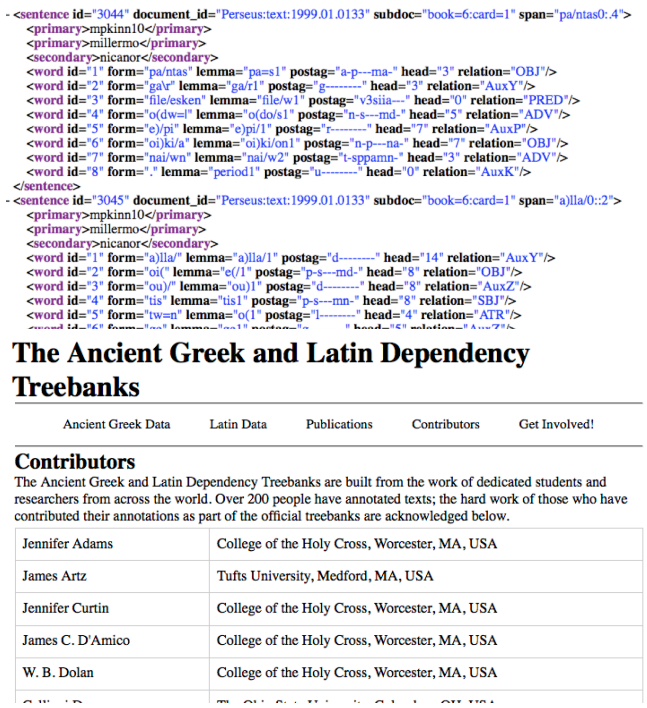


Figure 11: E-portfolios for Treebank contributions

We have created an environment with which users can develop and publish e-portfolios with a wide range of their *contributions*. These include not only morpho-syntactic analyses but varied curation and editorial tasks such as adding corrections to data entry errors, adding TEI XML structural markup, marking textual variants, disambiguating names or word senses, and creating parallel text alignments. Figure 11 illustrates how unique user IDs are stored as core elements to annotations within the Greek and Latin Treebanks in Perseus where two independent annotators analyzed each sentence *and* an expert reviewed instances where those annotators differed, producing a final shared analysis. Learners may choose to publish their own individual analyses for public review to demonstrate what they can accomplish on their own (or with less intensive supervision, such as statistical sampling of their work). This data allows us to provide e-portfolios either as stand-alone products or in a structured format for integration into Learning Management Systems (LMS) such as Sakai and Moodle, alongside more conventional materials such as papers or presentations.

E-portfolios also allow students of historical languages to re-invent one of the most traditional metrics of success: mastery of extensive reading lists that challenge learners to familiarize themselves with a cumulative network of sources building upon each other over time (e.g., as Vergil's *Aeneid* builds upon the Homeric *Iliad*). Such reading lists are demanding for independent learners—the Harvard Department of the Classics first reduced its undergraduate reading lists to c. 160,000 words of Greek and of Latin each and then finally abolished the reading lists altogether. Even at a competitive school such as Harvard, too few students had the linguistic training to cover a reading list of that size, much, if not most, of which students traditionally have had to tackle on their own. Nevertheless, extensive reading lists were, along with undergraduate theses, key elements for undergraduate education. The passage of language exams on reading lists for Greek and Latin remain an established rite of passage for PhD students of Classics, with demanding programs such as Berkeley posting reading lists of 500,000 for Classical Greek.

4. RELATED WORK

The study of annotations and their use in both digital libraries and online reading environments has an extensive amount of literature. One early and frequently cited web-based annotation system is Annotea [22] that made use of an RDF based model. A large body of work on the nature of scholarly annotation, hypertext and digital reading has also been presented by [24, 25]. Work by [1] has presented a formal model of the different types of annotations, and ongoing research by the Open Annotation Consortium (OAC) seeks to develop an "interoperable data model for scholarly annotation" based on the principles of Linked Data [11] and utilizing the technologies of the Semantic Web [16]. Recent work by [30] has experimented with the OAC model for providing persistent web annotations and also provides an excellent summary of the most significant recent work on the topic of annotation. [33] has also made use of the OAC model to develop the open source YUMA Media Annotation Framework.

Europeana and the **Europeana Data Model (EDM)** have addressed the larger shift from books to data to which the work presented here seeks to contribute. Europeana is best known for the millions of objects and images for which it has aggregated metadata but its long-term significance may lie in the fact that Europeana has moved beyond older metaphors of library and catalogue, and is currently experimenting with making all of its metadata available as Linked Open Data [20]. Thus, the EDM is designed to provide a single space in which words and physical objects of all kinds can be addressed [18]. For the Perseus Digital Library the CIDOC CRM, FRBR, and TEI XML have provided separate—and largely disconnected—spaces within which to organize information about art and archaeological objects, authors and their works, and textual data. The EDM provides the first space in which all of the Perseus collections can be fully represented. Every annotation to which this paper alludes can be represented in the EDM.

In terms of expanding contributions to digital cultural heritage, humanities projects have begun encouraging citizen scholars to contribute to digital history [32] and exploring crowd-sourcing tools such as Amazon's MechanicalTurk [23] and wikis [28] to support large scale user manuscript transcription (e.g., the Transcribe Bentham Project) as well as creating games to encourage users to assist in correcting OCR errors in historical text [15].

Additionally, a survey of potential technologies to be used in the assessment of student learning in intelligent language learning systems such as the one described in this paper has been offered by [29] and [14] has also provided a promising initial examination into the use of NLP in language learning assessment.

Another related field of work is that of e-portfolios and language learning. An e-portfolio augments traditional forms of assessment and it typically includes a variety of materials that document student capabilities in a given language. The research presented in [17] provides a through overview of two potential portfolio models, while [21] has offered an initial exploration in terms of the success of e-portfolios in helping students to successfully assess their own skills.

In terms of machine translation, the Chinese Room: Machine Translation Visualization Project has created [2] visualization for

diverse types of linguistic information to allow greater comprehension of machine translations by users with little or no fluency in the source language of the translation. They were able to develop an interface that exploits linguistic resources to allow these users to understand and to measurably improve broken machine translations.

The Homer Multitext Project [34], the first of its kind in Homeric studies, seeks to present the textual transmission of the *Iliad* and *Odyssey* in a historical framework [19]. It is particularly relevant to the work described here because undergraduate student researchers at Furman, Holy Cross, and the University of Houston collaborate to analyze, document, transcribe, markup, explicate, and translate not only the Homeric text but also the extensive Greek annotations on each manuscript. Students in first semester Greek can analyze page layout of manuscripts and then assume increasingly challenging tasks as their knowledge of the language and experience grows.

The Hestia Project [9] created for the Greek Historian Herodotus an Encoded Space-Text Archive, where each reference to a place in the text was geo-coded and aligned to an authority list with data such as longitude and latitude. The Hestia team was then able to use that data to explore ways in which Herodotus' *History* potentially represents a decentered or multi-centered understanding of the Mediterranean world based on relational flow and connectivity. Because an automatically geocoded English translation of Herodotus was available, project members used this as their data source, correcting and augmenting the annotations on the English rather than the Greek. They thus documented how annotation on an English translation could generate data that would advance expert knowledge about the underlying ancient source texts.

5. FURTHER WORK

A great deal remains to be done on each of the functions described above. Our larger goal is, however, to integrate these services from Perseus and Alpheios.net into Philologist (Figure 12), an open environment for annotating and learning historical languages.

The work presented above has been tested primarily with student researchers in formal academic programs. We recognize the need to generalize these results by working with independent learners who may emerge as citizen scholars. More generally, we need to globalize the tools that we have presented, reducing their dependence upon English. The annotations that we are collecting are largely independent of the annotator's native language—it does not matter if your first language is Arabic or English if you are publishing syntactic analyses for sentences in Plato. Such globalization is critical if we are to work with sources in multiple languages and to understand global cultural heritage as a network of cultures, each interacting with the other. The 2009 MLA "Enrollments in Languages other than English survey" cites 35,000 students of Modern Standard Arabic (p. 19) but only 306 students of the Classical and Quranic Arabic (p. 29), 202 students of Classical Chinese (p. 29), and 499 students of Sanskrit (p. 32). Every major cultural system from the Atlantic to the Pacific interacted over thousands of years but established North American and European centers of academic learning have produced enough experts in the relevant languages. We must develop new scholarly relationships in which colleagues from India, China, the Middle East and elsewhere play a fundamental, tangible and immediate role. Much critical work involves digital editing and annotation, the products of which do not necessarily depend upon the native language of the producer.

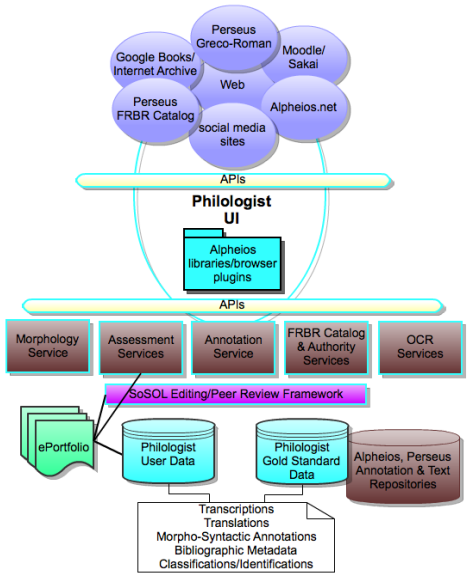


Figure 12: Philologist

6. CONCLUSIONS

We have shown that (1) billions of words with an open set of annotations, rather than tens of thousands of books, have already emerged as the objects which emerging digital infrastructure must manage; (2) that a participatory culture, combining features of laboratory and library, has begun to emerge in formal education, as student researchers (and ultimately citizen scholars) collaborate with library professionals and student researchers to manage this explosion of data; (3) that learners can track and document the annotations that they contribute, thus reaping the psychological rewards of making credited contributions to an intellectual enterprise that extends thousands of years into the past and forward into the future; (4) that the contribution of some annotations also advances core tasks of language learning, developing skills from introductory through advanced mastery of the language; (5) that learners can, by working with and adding to annotations, conduct more sophisticated research, working with a wider range of linguistic sources than was feasible with book-oriented print infrastructures; (6) that learners can document both what they have contributed and what they have mastered, publishing their own personalized reading lists and much more tangible, quantifiable and ultimately verifiable data about what they have accomplished than summary grades or institutional reputations alone can convey. We have thus created working elements of a larger infrastructure for the study, at all levels, of textual sources from the human record.

7. ACKNOWLEDGEMENTS

This work has been supported by grants from the NEH [Dynamic Lexicon (PR-50013-08), Digging into Data (HJ-50013-10), Hellespont Project (HG-50020-10)], NSF [Mining a Million Books (091065)], Mellon Foundation (Cybereditions, Bamboo, Greco-Arabic), the Cantus Foundation, and the Google Digital Humanities Program.

8. REFERENCES

[1] Agosti, M. and Ferro, N. 2007. A formal model of annotations of digital content. *ACM Trans. Inf. Syst.* 26 (Nov. 2007): 3+. DOI= http://dx.doi.org/10.1145/1292591.1292594

[2] Albrecht, J., Hwa, R., and Marai, G.E. 2009. The Chinese room: visualization and interaction to understand and correct

ambiguous machine translation. *Comput Graph Forum*, 28 (June 2009), 1047-1054.

[3] Babeu, A. 2011. "*Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classicists*. Technical Report. CLIR.

[4] Bagnall, R. 2010. Integrating digital papyrology. In *Online Humanities Scholarship: The Shape of Things to Come*. http://hdl.handle.net/2451/29592.

[5] Bamman, D, and Crane, G. 2011. Measuring historical word sense variation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (JCDL '11). ACM, New York, NY, USA, 1-10. DOI=http://doi.acm.org/10.1145/1998076.1998078.

[6] Bamman, D. and Smith, D. To appear. Extracting two thousand years of Latin from a million book library. *JOCCH*.

[7] Bamman, D., Babeu, A., Crane, G. 2010. Transferring structural markup across translations using multilingual alignment and projection. In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10)*. ACM, New York, NY, USA, 11-20. DOI= http://dx.doi.org/10.1145/1816123.1816126

[8] Bamman, D., Mambrini, F., and Crane, G. 2009. An ownership model of annotation: the Ancient Greek Dependency Treebank. In *TLT 2009: Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories Conference*.

[9] Barker, E., Bouzarovski, S., Pelling, C., Isaksen, L. 2010. Mapping an ancient historian in a digital age: the Herodotus Encoded Space-Text-Image Archive (HESTIA). *Leeds International Classical Studies*, 9 (March 2010). http://www.leeds.ac.uk/classics/lics/2010/201001.pdf

[10] Berti, M., Romanello, M., Babeu, A., and Crane, G. 2009. Collecting fragmentary authors in a digital library. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries (JCDL '09). ACM, New York, NY, USA,* 259-262. DOI=http://dx.doi.org/10.1145/1555400.1555442.

[11] Bizer, C. Cyganiak, R. and Heath, T. 2007. How to publish linked data on the Web. Available at: http://sites.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

[12] Blackwell, C., and Martin, T. 2009. Technology, collaboration, and undergraduate research. *Digital Humanities Quarterly*, 3, 1 (Jan. 2009).

[13] Bulger, M., Meyer, E.T., and de la Flor, G. 2011. *Reinventing Research? Information Practices in the Humanities*. Technical Report. Research Information Network.

[14] Chapelle, C.A., and Chung, Y.R. 2010. The promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27 (July 2010), 301-315.

[15] Chrons, O., and Sundell, S. 2011. Digitalkoot: Making old archives accessible using crowdsourcing. In *HCOMP 2011: 3rd Human Computation Workshop.*

[16] Cole, T., and Han, M. 2011. The Open Annotation Collaboration Phase I: Towards a shared, interoperable data model for scholarly annotation. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1 (3), July 2011.

[17] Cummins, P. W. and Davesne, C. 2009. Using electronic portfolios for second language assessment. *The Modern Language Journal*, 93 (2009), 848-867.

[18] Doerr, M., et al. 2010. The Europeana Data Model (EDM). In *World Library and Information Congress: 76th IFLA General Conference and Assembly*, 10-15 Aug. 2010, Gothenberg, Sweden.

[19] Dué, C., and Ebbott, M. 2009. Digital criticism: editorial standards for the Homer Multitext. *Digital Humanities Quarterly*, 3 (Jan. 2009).

[20] Haslhofer, B. and Isaac, A. 2011. data.europeana.eu - The Europeana Linked Open Data pilot. In *Proc. Int'l Conf. on Dublin Core and Metadata Applications 2011*, 94-104

[21] Hung, S. T. 2009. Promoting self-assessment strategies: an electronic portfolio approach. *The Asian EFL Journal Quarterly*, 11, 2 (June 2009), 129-146.

[22] Kahan, J. and Koivunen, M.R. 2002. Annotea: an open RDF infrastructure for shared Web annotations. In *Proceedings of the 10th international conference on World Wide Web* (WWW '01). ACM, New York, NY, USA, 623-632. DOI=http://dx.doi.org/10.1145/371920.372166

[23] Lang, A., and Rio-Ross, J. 2011. Using Amazon Mechanical Turk to transcribe historical handwritten documents. *The Code4Lib Journal* (Oct. 2011).

[24] Marshall, C. C. 1998. Toward an ecology of hypertext annotation. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia:* (HYPERTEXT '98). ACM, New York, NY, USA, 40-49. DOI=http://doi.acm.org/10.1145/276627.276632.

[25] Marshall, C.C. and Brush, A.J. 2004. Exploring the relationship between personal and public annotations. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '04). ACM, New York, NY, USA, 349-357. DOI=http://dx.doi.org/10.1145/996350.996432.

[26] Michel, J. B., Shen, Y.K., and Aiden, A.P. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331, 6014, (Jan. 2011), 176-182. DOI=http://dx.doi.org/10.1126/science.1199644

[27] Molin, C., Nyhan, J., Ciula, A, et al. 2011. *Research Infrastructures in the Digital Humanities*. Technical Report. European Science Foundation.

[28] Moyle, M., Tonra, J., and Wallace, V. 2011. Manuscript transcription by crowdsourcing: Transcribe Bentham. *Liber Quarterly - The Journal of European Research Libraries* 20 (Sept. 2011).

[29] Ockey, G. J. 2009. Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, 93 (2009): 836-847.

[30] Sanderson, R. and Van de Sompel, H. 2010. Making web annotations persistent over time. *Proceedings of the 10th annual joint conference on Digital libraries* (JCDL '10). ACM, New York, NY, USA, 1-10. DOI=http://dx.doi.org/10.1145/1816123.1816125

[31] Schmidt, D. and Colomb, R. 2009. A data structure for representing multi-version texts online. *Int. J. Hum.-Comput. Stud.*, 67, 6 (June 2009), 497-514.

[32] Sikarskie, A.G. 2011. Citizen scholars: Facebook and the co-creation of knowledge. In Jack Dougherty and Kristen Nawrotzki, eds. *Writing History in the Digital Age*. Under contract with the University of Michigan Press. Web-book edition, Trinity College (CT), Fall 2011.

[33] Simon, R., J. Jung, and B. Haslhofer. 2011. The YUMA media annotation framework. *Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries*. Berlin, Heidelberg: Springer-Verlag, 2011, 434-437.

[34] Smith, Neel. 2010. Digital infrastructure and the Homer Multitext Project. *Digital Research in the Study of Classical Antiquity*. Eds. Gabriel Bodard and Simon Mahony. Burlington, VT: Ashgate Publishing, 121-137.