Harvard Data Science Review • Issue 1.2, Fall 2019

Beyond Translation: Language Hacking and Philology

Gregory Crane[1]

[1]**Tufts University**

Published on: Nov 30, 2019

DOI: 10.1162/99608f92.282ad764

ABSTRACT

Among applications of machine learning, automatic language translation stands out for its spectacular gains from exploiting large data sets and its promise of cultural connection. Knowledge of a range of languages has been a marker of human erudition for centuries. The work described here outlines the rise of a third path, one situated between linguistic mastery and reliance upon translation. Manually annotated sources have made this third path possible for centuries, if not millennia, but a combination of increasingly powerful machine learning, growing bodies of linguistic annotation and distributed human contributions has the potential to generalize such 'language hacking.' In such an environment, translation (which may be the product of machines or humans) must be judged not only by traditional metrics (such as accuracy and readability) but also by the degree to which it enables readers to push beyond the translation and to analyze the original source text itself. The growth of such language hacking opens up a new intellectual space for the ancient discipline of philology—broadly defined as the sum of all available practices by which we use the human linguistic record to understand the past. This new space integrates fundamental goals from the humanities with emerging methods from computational and data science. I see a bright future for machine learning that acts as intelligence augmentation (Jordan, 2019), and for critical readers and explainable artificial intelligence to develop in tandem.

Keywords: digital humanities, digital classics, philology, translation studies, digital reading, smart texts, semantic annotations, language technologies

1. Motivation and Aims

My interest, and that of many of my collaborators, lies in how data science and/or machine learning, although designed at least in part to reduce the need for human intellectual labor, can foster greater intellectual activity within living human brains. My particular interest centers on reading and, even more precisely, upon how we interact with sources in languages that we have not mastered and from cultures that are not our own. This work seeks to address alarming consequences from reading on digital devices that researchers on reading in a digital age have reported. Maryanne Wolf, a leading expert in the effects upon reading, summarized many of these concerns in an August 2018 article published in the *Guardian*: "Skim reading is the new normal. The effect on society is profound" (Wolf, 2018). She reports that her own research "depicts how the present reading brain enables the development of some of our most important intellectual and affective processes: internalized knowledge, analogical reasoning, and inference; perspective-taking and empathy; critical analysis and the generation of insight. Research surfacing in many parts of the world now cautions that each of

these essential 'deep reading' processes may be under threat as we move into digital-based modes of reading."

I have no doubt that the unnerving results are well founded. But there are multiple modes of reading and the preferred mode of reading cited by Wolf has been linear reading of prose, primarily novels. While this can be an immersive experience and the negative effects of skim reading clearly highlight benefits of this immersive reading, there are other ways that readers engage with texts. As a student of historical sources that are produced in languages that are no longer spoken and reflect cultures that no longer exist, I focus on a different kind of reading, in which we read, reread, reflect, and analyze. We can never assume that we can rely upon the intuitions of a native speaker. We may have a general ability to understand historical sources, but we must always remember that we are engaging with an alien culture to which we will never have direct access. We never know when a new question or point of view will challenge us to rethink a work, a passage, or even a phrase that we may have read for decades. Those of us who study ancient sources are (for the most part at least) trained to exercise a particularly intensive mode of reading. We need to read works from beginning to end, and we may become immersed as we read texts such as the Homeric *Iliad* or the letters of Seneca. But that linear reading provides us with the foundation upon which all of our real work must build.

This article describes some of the methods by which digital media are already beginning to support something profoundly different from both skim reading and immersion in the linear flow of a novel. My interest is in developing reading environments that challenge audiences to pause and explore sources in languages that they may never have studied. My goal is to enable and to foster a form of deep and active reading that has its roots in traditional close reading but that, through the use of digital media, allows the emergence of new, deep, and open-ended forms of intensive reading. In my own work, I seek to foster a new and far more dynamic relationship between translations and the source texts that they represent.

When I turn to the broader world, most of the work on language technologies is indeed reductive: the goal is to make sources more comprehensible by allowing readers to grasp the substance of a source that they could have read if they had time (e.g., speakers of English working with publications in English). Deep learning and rapidly improving machine translation are unsurprisingly beginning to disrupt the localization industry. Localization involves translation, and translation has traditionally required human input. As my wife and I watch Americans decide where to live on *House Hunters International*, we don't think about the fact that the Discovery Channel itself aims for a global audience and, reportedly, translates more than 100,000 hours of content a year. The A&E network translates content into more than 20 different languages (Bond, 2019). Where HBO enjoyed great success in the early 21st century by translating anglophone content, such as *Game of Thrones*, into multiple languages, Netflix has, with both its own productions and with licensing, aggressively gathered content produced

in a range of languages. These include not only languages that are widely taught in the United States (such as Mandarin, French, Spanish, and German), but also Turkish, Malay, Korean, and Hindi. To reach the widest possible audience, content on such platforms needs to be localized for as many languages and cultural backgrounds as possible.

But if machine translation potentially automates a crucial function and even reduces (if it does not fully eliminate) the need for human translators, the range of language technologies opens up new ways in which we can interact with sources in languages that we do not know. In particular, we can view the translation not so much as an end, but as an initial approximation and a starting point for further exploration. This constitutes a third way of interacting with sources in languages that we do not know, one situated between traditional linguistic mastery, on the one hand, and dependence upon a translation, on the other.

2. Background: Some Traditional Efforts to Go Beyond Translation

This middle path has existed in limited forms for centuries, if not millennia. A parallel text such as the Rosetta Stone does not quite support readers who wish to explore an unknown language: its goal is simply to make the same message available to audiences familiar with different language systems (Ancient Egyptian hieroglyphs, demotic script, and Greek). This middle path would lead us to cross linguistic boundaries and to probe the language versions that we did not immediately understand and to determine whether, how, and where the different versions conveyed substantively different ideas.

Scholars have, however, sought to introduce materials in unfamiliar languages. When Franciszek Meninski published a poem by the Persian poet Hafez in his 1680 *Thesaurus Linguarum Orientalium* (Meninski, 1680), he included not only a literal translation into Latin (the language of scientific publication in Europe at the time), but a transliteration, information about the meter and pronunciation of the poem, general background on the literary genre, and detailed explanations (with transliterations) of each word in the poem. Meninski made it possible for readers to engage directly with a representative passage from a poet of immense cultural importance but virtually unknown in Europe. Other scholars, before and since, have made similar efforts. Linguists, in particular, have felt the need to develop elaborate standards (such as the Leipzig Glossing Rules [Committee of Editors of Linguistics Journals, 2015], so that they could work with precision on linguistic specimens from thousands of surviving languages. Scholars of earlier languages have applied these methods to Ancient Egyptian and other ancient languages [Topoi, 2019; Figure 1]).

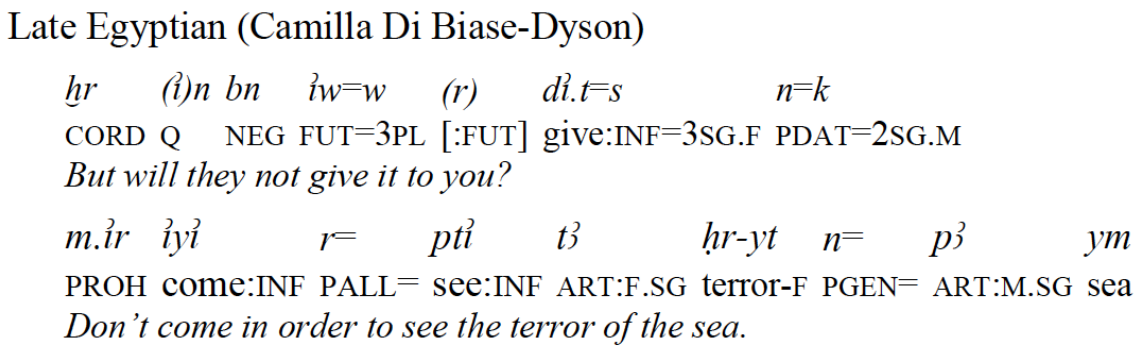

Figure 1. Leipzig Glossing Rules applied to Late Egyptian (DiBiase-Dyson, Kammerzell, & Werning, 2009, p. 363).

But useful as such extensive annotation may be, such annotations require a great deal of manual labor and it would be difficult to annotate even the most widely studied sources, much less the hundred million or so surviving words of Greek produced through 1453 or the billions of words of Latin that survive from antiquity through relatively recent times (such as in the Latin book on Arabic, Turkish, and Persian by Meninski cited above).

When I was a first-year student in college in the 1970s, I encountered a more algorithmic (if not automated) approach that scaled differently. I had spent a great deal of my time in secondary school learning Latin and especially Greek and, because of this unusual background, entered college with the ability to take more advanced courses than most of my fellow students. Early in the semester, though, I discovered, to my great surprise, that a young professor, Gregory Nagy, had developed a method by which students with no knowledge of Greek were able to begin analyzing the Greek text of Homer. Two resources made this possible. Each resource was profoundly rooted in print culture, but each represented a common structural principle that allowed each source to reinforce the other in ways that the authors had probably never imagined. The new combination created a system that was greater than the sum of its parts. Several of my fellow students took this course and it became clear to me from talking with them that the system worked: they really were able to work effectively with a text in a language that they had not learned. Of course, there were limits and I could find points to criticize but the main point seemed clear to me: my fellow students had broken through the linguistic barrier and I could talk to them about the Greek text itself. I never forgot what I saw in that first semester and those impressions have shaped my thinking—and much of the work that I have done in contributing to a digital infrastructure for Ancient Greek—in the decades that have followed.

The fundamental principle involved the acceptance of a shared and ancient coordinate system. Already in the ancient world, the *Iliad* and *Odyssey* had each been divided into 24 chunks, with each chunk containing an amount of Greek text that would fit comfortably in a scroll. (One of the advances of the

modern book form with turning pages is that it can ergonomically store larger texts.) Both epics are composed in a poetic form called dactylic hexameter. Each line follows the same metrical form and has a well-defined beginning and end. The poetic lines provided a natural coordinate system by which to identify chunks of the poem: Od. 1.6-9 designates lines 6 through 9 from the opening book of the *Odyssey*. While different editions may vary, a citation such as "Od. 1.6-9" designates the same basic text whether we are looking at a Greek edition of the *Odyssey* published in Germany in the 18th century or in the United States in the 20th. This common citation convention shaped two publications that were published thousands of miles apart and a century apart, allowing them to serve uses that the authors of neither publication could have anticipated.

In the second half of the 19th century, Guy Lushington Prendergast and Henry Dunbar published manually produced concordances to the *Iliad* (Prendergast, 1874) and *Odyssey* (Dunbar, 1880). Neither Prendergast nor Dunbar was a professional scholar but each played a role that was distinctive of their time: Prendergast had been a civil servant in India, while Dunbar was the medical advisor to a wealthy family in Scotland. Concordances provide what we would now call keyword-in-context indices: for each word in the Homeric Epics, they provide the full line of context. The concordances are not fully exhaustive (they do not, for example, cover common words such as *kai*, the most common word for 'and'), but they are daunting productions, requiring an immense amount of labor. The labor is, however, algorithmic: forms are listed on pieces of paper (one piece of paper for each indexed word in a line), the pieces of paper are sorted, and the resulting list is tabulated into a print index. Prendergast tells us in his preface to the *Iliad* concordance that he began work on it on June 5, 1847, and completed it more than 16 years later, on October 18, 1863. He had other things to do in his life, but 16 years is a long time by any account. Dunbar saw his *Odyssey* concordance as a complement to the *Iliad* concordance of Prendergast, adopting the same basic format and following the conventional book/line citation scheme.

A century later, the American poet and classicist Richard Lattimore produced English translations of the *Iliad* (1951) and *Odyssey* (1967). Unlike many translators, Lattimore decided to follow an algorithmic rule that would open up new possibilities for his translation. Most of those who translate Homer into English poetic form find it too unnatural to follow the line breaks in the Greek original and use their own lineation. Thus, line 348 of book 6 of the *Iliad* in one poetic translation does not correspond to line 348 of book 6 in the Greek. Readers can usually align Greek and English but this can be a laborious process if we are moving from one passage to another. Lattimore decided to translate the Homeric epics line by line so that the book/line numbers in his English version corresponded very closely to those of the Greek original.

Gregory Nagy saw that audiences who had not studied Greek could use the shared citation scheme to combine concordance and translation. He had students spend an hour or so getting reasonably comfortable with the Greek alphabet and then gave them the forms to look up in the concordance.

They did have to convert traditional citations that used Greek letters instead of numbers to designate the books: e.g., (α) 262 designates line 262 of book 1, (β) 44 designates line 44 of book 2, and so on. But once they had accomplished that, students could look up the book and line number in the Lattimore translation for the *Iliad* or *Odyssey*.

Figure 2 juxtaposes extracts from the print concordance on the left and corresponding English translation on the right. The concordance prints the single line of poetry in which the word appears. For the translation, I have selected minimal text so that readers can understand the basic context. In practice, readers in the 1970s would have looked the passages up in the translation and had the whole context for each instance of a word.

In the Figure 2, we look at instances of the word *mênis*, conventionally translated as 'anger' or 'wrath.' The term *mênis* is the first word of the *Iliad*, and the line is conventionally translated (as in the figure) "sing, goddess, the anger of Peleus' son Achilleus." Some translations use 'wrath,' 'rage,' or a similar term. If readers begin to examine instances where this word is used, however, a pattern begins to emerge.

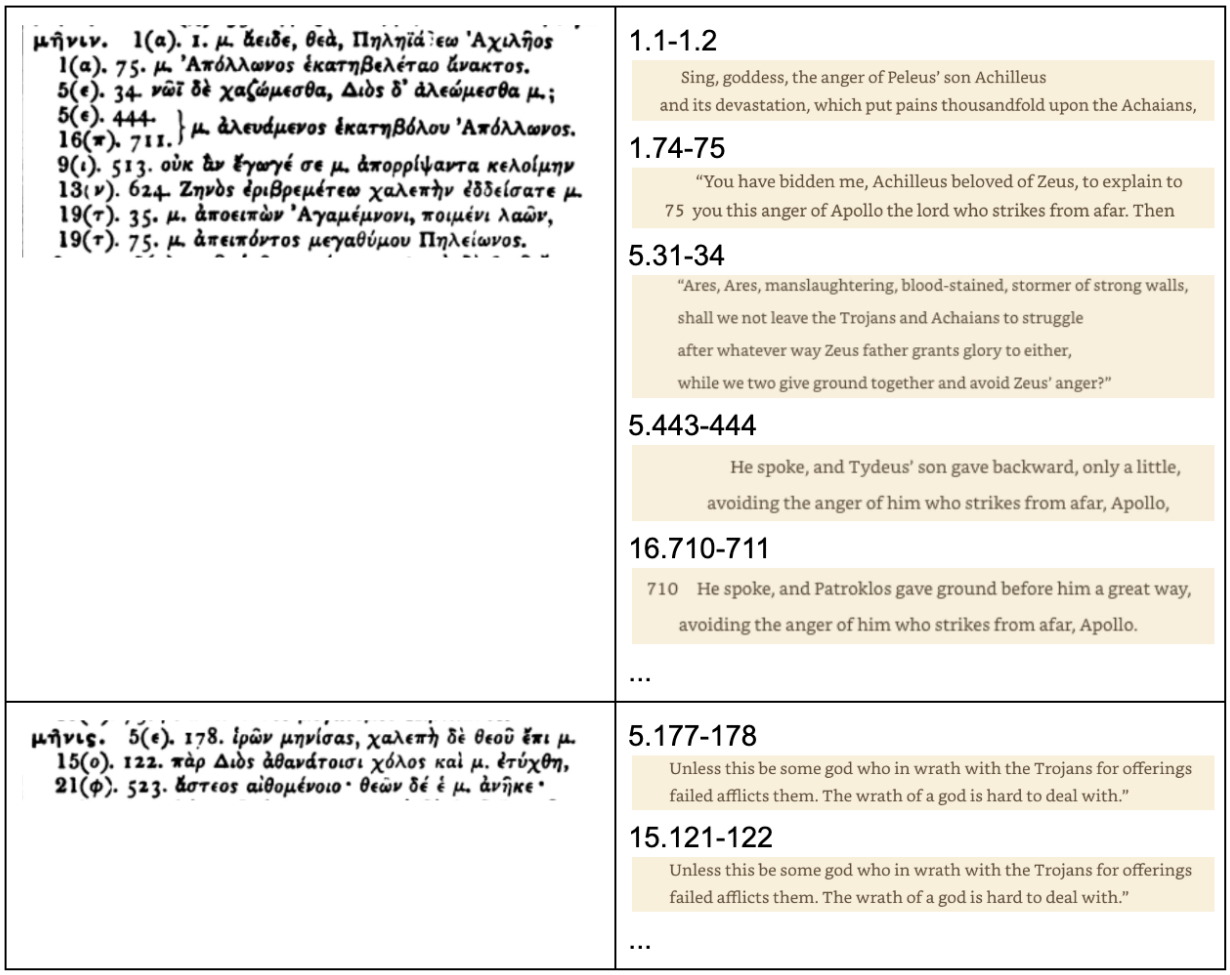

Figure 2. Left: Excerpts from Prendergast's 1874 concordance to the *Iliad*. Right: Excerpts from the current Kindle edition of Lattimore's 1951 translation of the *Iliad*.

Figure 2 includes some of the contexts.[1] The form *mênin* designates the noun when it is the object of a verb, while *mênis* is the form used when the word is the subject. While mênis designates the wrath of Achilles in the opening line of the *Iliad*, in the other six passages in the figure, the word mênis describes the wrath of divinities: Apollo (3x), Zeus (1x), and a general divinity (2x). In the first five passages, the word is translated as 'anger,' while 'wrath' is the translation in the final two passages selected.

My point is not to offer a definitive interpretation (my colleague Leonard Muellner has written an entire book about the term *mênis* [Muellner 1996]), nor do I argue that readers using the approach discussed here can always determine the relationship between source text and translation. Of course, they probably do not have specialized knowledge of the genre and cultural context. But readers can—and my own students invariably do—begin to see patterns that would never be visible if they relied

solely upon a modern language translation. These insights tend to affect not only how they understand the particular text with which they are working but how they think more generally about their relationship to translation as a whole. Thus, in the example discussed, readers examining the 12 instances of *mênis* in the *Iliad* would find that four described the anger of Achilles, while the other eight all designated that of divinities, and all three uses of this word in the *Odyssey* apply to divinities. The very first word of the *Iliad* thus attributes to Achilles a particular kind of anger that normally belongs to divinities. If readers were to explore this question today and searched the Kindle version for the English words 'anger' and 'wrath,' they would quickly find that other Greek words (e.g., *cholos*) more typically designate this emotion. However we interpret the meaning of this word (and I recommend Muellner's book to anyone interested in this term), readers can see that this word for anger introduces the tremendous status that the poem attributes to Achilles, a status that has elements of the divine but also of the excruciatingly mortal: for in asserting his status, Achilles also ensures his own death and emphasizes his mortality.

At the same time, more generally, once readers begin to see patterns that are invisible if they rely solely upon a translation, readers can (and my students normally do) undergo a fundamental shift in how they see translations and in how they understand language. This even happens with students who have expertise in one or more foreign languages. They may realize (especially if they are familiar with a non-Western language like Arabic or Mandarin) that many words simply don't translate. But we often ignore this fact when we rely upon translations—at some level, we realize that translation often hides and distorts the meanings of the original, but this knowledge has less impact if we can't do anything about it. More broadly, if students can cultivate this critical approach to reading translations produced by humans, they can generalize it to translations and other interpretations offered by machines.

3. Smart Texts as an Emergent Space for Reading

The following section outlines features that combine to form 'smart texts,' texts that combine a growing range of increasingly adaptive features to support readers from a growing range of linguistic and cultural backgrounds. The term smart texts implies that such texts will evolve and become more powerful. Although available in different environments and applied to different source texts, the services described in this section complement one another intellectually and provide tangible examples of services that could be combined into a coherent reading environment. These services are still at relatively early stages of development but would already, if combined smoothly with each other, provide an 'emergent reading environment,' that is, an environment with properties that could not be inferred by studying each component in isolation. While we can certainly build smart texts, such technology is only a means to a larger goal: to foster the development of smarter readers. These

readers are developing interpretive skills on the record of the human past and can thus, secondarily, serve as models for the next generation of explainable machine learning models

No one text has attracted each form of explanatory resource that I wish to emphasize. Thus, I jump from an edition of the Latin historian Livy to a map of places in Herodotus, rather than focusing on Livy or Herodotus (or creating an artificial storyboard of examples for this article). But the examples listed reflect work from different projects, each of which has evolved along its own independent trajectory. When we begin assembling these different threads and considering how they could be combined, an emergent experiential environment begins to take shape, one that can combine images of text with sound and images, still and moving, representing performance and/or cultural background. I draw from work on ancient and early modern texts because these areas have attracted substantial work, but I want to emphasize that the features reflect very general practices that readers face whenever they attempt to think deeply about sources, whether those sources are Latin histories, Persian poetry, or YouTube videos of a Latvian pagan metal band.

A researcher who is an intellectual historian of early modern thought but not a professional classicist may be interested in what kinds of notes intellectuals added in the margins of the books that they read. In Figure 3, we see an early modern edition of Livy's *History of Rome* (Livy, 1555), with annotations in the margins. The figure is drawn from the Archaeology of Reading, a project developed by faculty, library professionals, and students at Johns Hopkins, Princeton University, and University College London and funded by the Mellon Foundation. The Archaeology of Reading offers high-resolution scans of early modern printed books as well as transcriptions (and, where necessary, English translations) of annotations written in the margins by John Dee (1527–1608) and Gabriel Harvey (c. 1552/3–1631), two prominent early modern English intellectuals.

Figure 3 shows a page from a Latin edition of Livy's 1st century BCE *History of Rome* with annotations by Harvey.

Visit the web version of this article to view interactive content.

Figure 3. Early modern edition of Livy's *History of Rome* viewed via IIIF with manual transcriptions and translations of hand-written annotations in the margins (Archaeology of Reading, 2019).

A user might be an intellectual historian of early modern Europe who wishes to explore the reception of Livy and, in particular, which passages are most commonly cited in different periods and in different languages (e.g., English, French, and German). The intellectual historian interested in particular annotations by particular people in a particular book will also often want to see more

generally how often particular passages are quoted. To answer such a question, the reader can exploit data already available from the Proteus Project to align uncorrected optical character recognition (OCR)-generated text to Text Encoding Initiative (TEI) XML texts with Canonical Citation Schemes (CTS-encoded)[2] to determine that the text on this page begins in Livy, book 1, chapter 9, section 2 (Livy 1.9.2) and ends in book 1, chapter 10, section 7, which serves as the foundation for further exploration described in the following.

Once we have identified the canonical citations for the relevant text, we can bring in other available digital resources that are relevant to this passage of Livy.[3] These include:

- Maps. Where we have annotations linking place names in a text to authority lists such as Pleiades, we will create maps illustrating the geospatial coverage of a text as well as the context within the source text where the place name is mentioned. Figure 4 presents a map of places listed in the ancient Greek historian Herodotus.

Visit the web version of this article to view interactive content.

Figure 4. A map illustrating places cited by the Greek historian Herodotus in his History of the Persian War (Pelagios Project).

The map in Figure 4 is based on annotations to the English translation rather than the Greek texts. While annotating the English translation (rather than the Greek text itself) introduces some issues (e.g., the translation may represent 'the Athenians' as 'Athens,' converting an ethnic group to a place name), annotating a modern language translation greatly increases the pool of possible annotators and allows the use of far more advanced named entity analysis software than is available for ancient languages. The resulting annotations in the modern language translation can, to a very high degree of accuracy, be automatically aligned to the Greek or Latin source text for more precise analysis later. (The accuracy depends upon how closely the translation follows the source.)

- Translations, commentaries, textual notes, and other resources already available in the Perseus Digital Library, a digital library that has been under development since 1985 and that contains source texts in Greek, Latin, and other languages, as well as translations and explanatory information in English. Perseus has multiple editions and translations for Livy (figure 5), and Livy thus offers a very useful example of what will be possible.

Visit the web version of this article to view interactive content.

Figure 5. Chapter 9 from book 1 of Livy's *History of Rome*, with multiple translations and resources as it appears in the Perseus Digital Library.

The reader then returns to the annotation in the Archaeology of Reading. Many readers will turn to the English translation of the Latin annotations (Figure 6) and make relatively little, if any, direct use of the Latin.

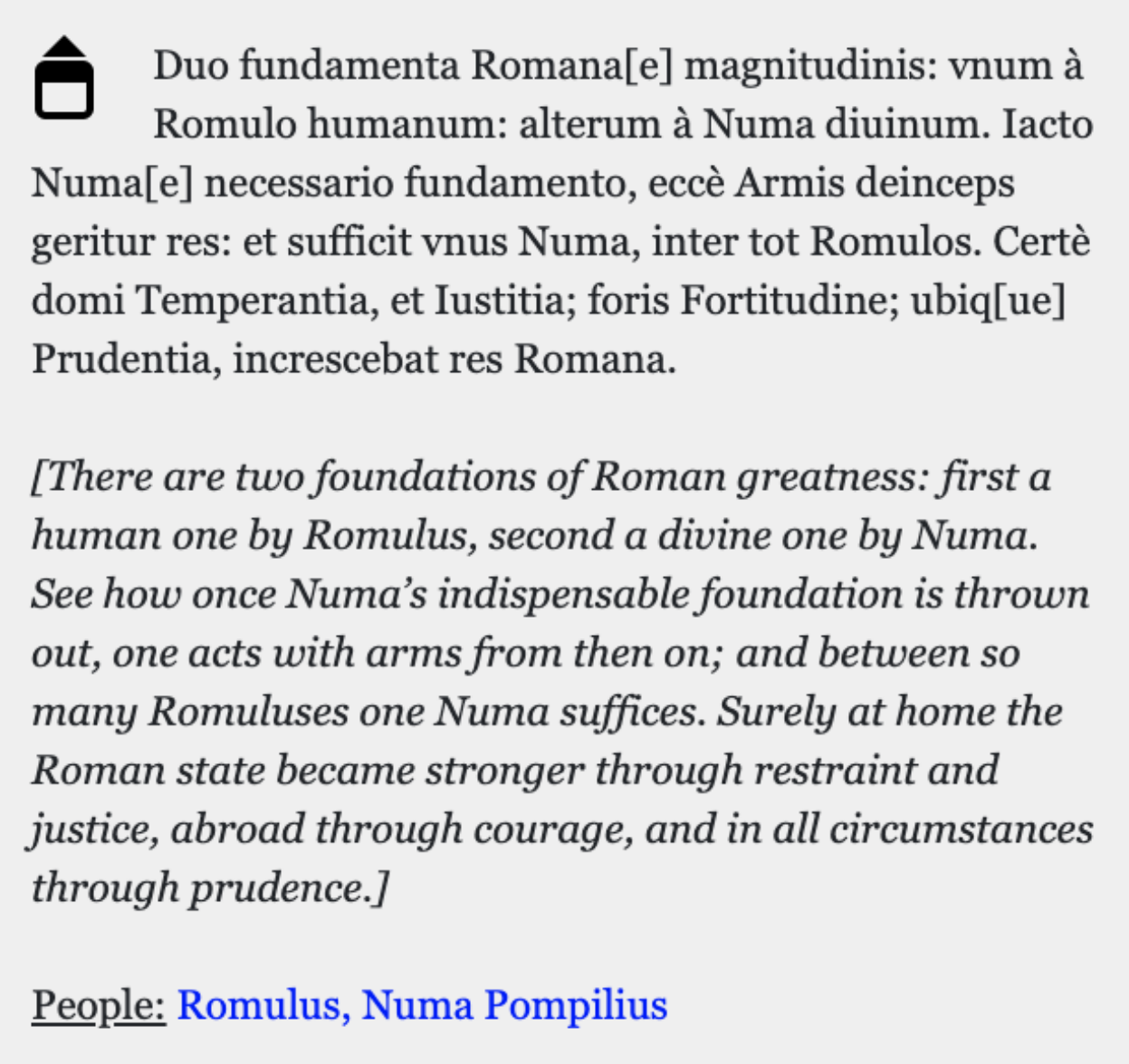

Figure 6. A manually transcribed and translated annotation from the early modern edition of Livy mentioned above.

A more powerful reading environment is beginning to take shape that can enable new forms of reading. Components of this reading environment include annotations that define precise spans of text in particular editions. Figure 7 illustrates how a particular reading environment, the Scaife Viewer[4] builds on the CTS data model to support annotations that describe precise spans of text in particular editions of a work.

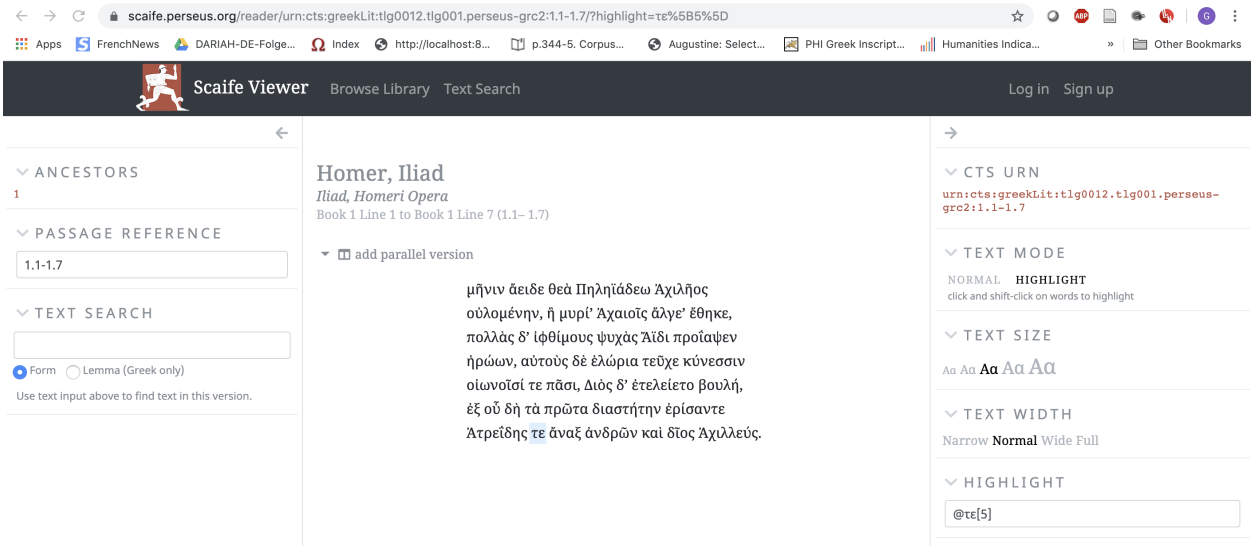

Figure 7. The Scaife Viewer, an evolving reading environment for the Perseus Digital Library and other resources.

In the example in Figure 7,[5] the Scaife Viewer the fifth instance of particular word (in this case, the Greek "τε", "and") within lines 1–17 of book 1 of a particular edition (Murray, 1924) of the *Iliad*—information that is encoded in the Uniform Resource Name: urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7.[6]

If we return now to the initial example, the 1555 edition of Livy's first century BCE *History of Rome*, we can use the annotation data model above to bring together annotations from what are now separate resources into a coherent reading environment. The set of possible annotations that can be aligned to the text of Livy is open-ended and will surely evolve over time.

To illustrate the possibilities of what could be done with convergent annotations about Livy, I list four classes of annotation that are already available for different texts: (1) morpho-syntactic analysis; (2) translation alignment; (3) synchronized performance and metrical analysis; and (4) text reuse detection. As the environment evolves and more data accumulates, all such annotations can begin to cluster around particular texts.

1. *Morpho-syntactic Analysis*. Figure 8 shows a sentence from a work by Lorenzo Valla with full morphological and syntactic analysis. This allows readers to see precisely how each word fits into the

sentence.[7]

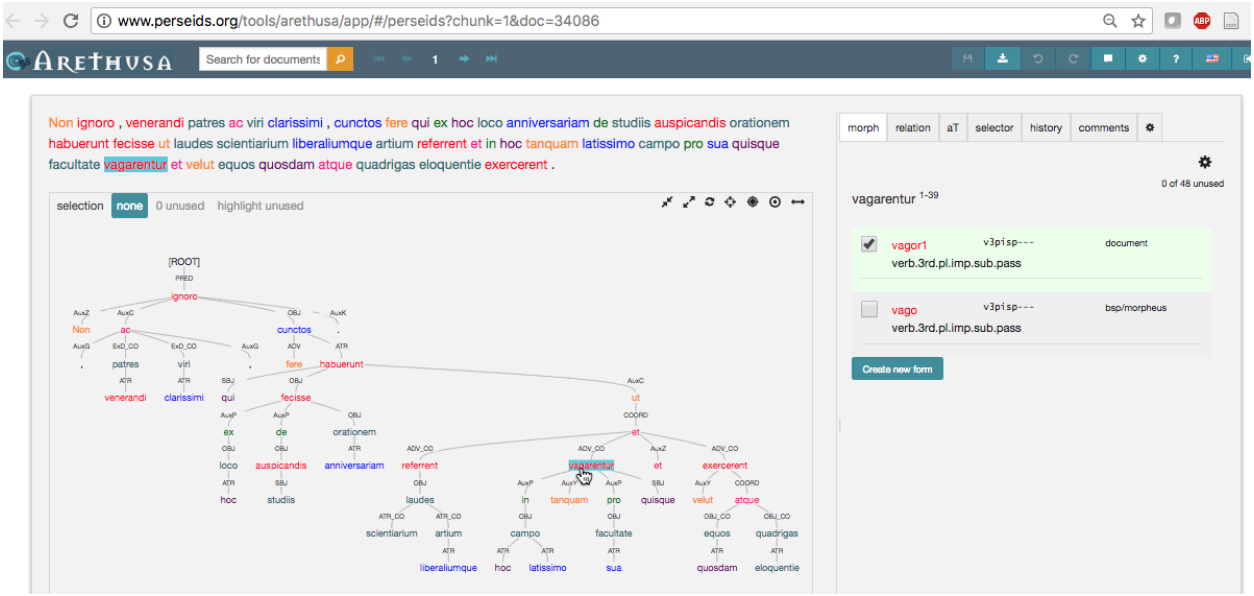

Figure 8. Morpho-syntactic analysis of a Latin text by Lorenzo Valla.

2. *Translation Alignment.* Here readers are able to see precisely how the English translation does (and does not) correspond to the Latin.

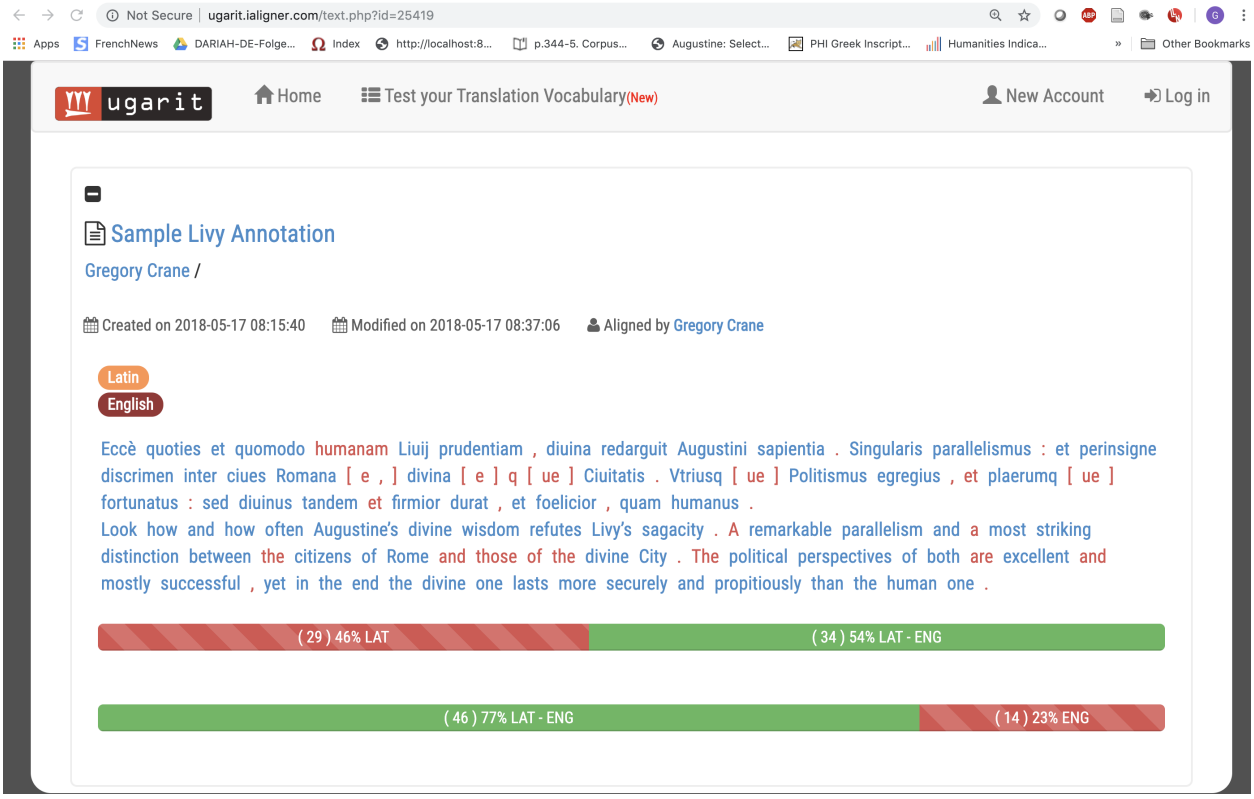

Figure 9. A translation aligned at the word and phrase level to the source text (Ugarit Text Aligner).

In Figure 9, the reader examines an aligned version of the Latin and this translation.[8] Whether or not the reader knows Latin, the visualization reveals at least two points: (1) The fact that the Latin word *humanam* remains highlighted in red reveals the fact that the English contains no direct equivalent to this Latin word (The amount of residual red provides a general overview of how close a text and a translation are.); (2) The reader can see that the phrase "political perspective" corresponds to the Latin *politismus*.

Additional functionality is visible in a reading environment specifically designed to show the word and phrase alignments between the works of the Persian poet Hafez and an English translation. Maryam Foradi, a PhD student at Leipzig University, carefully aligned the 70,000-word Divan of Hafez to a literal English translation as part of her Ph.D. dissertation and we thus have a database of curated translation alignments and an accompanying interface. Tariq Yousef, also a PhD student at Leipzig, developed an application to show how readers could exploit the alignments. This example illustrates that the methods we are developing already work for multiple languages—the key is to develop the textual data. In the example in Figure 10, the reader sees the Persian that corresponds to the "desire" in the English. The reader clicks on "desire" to learn more.

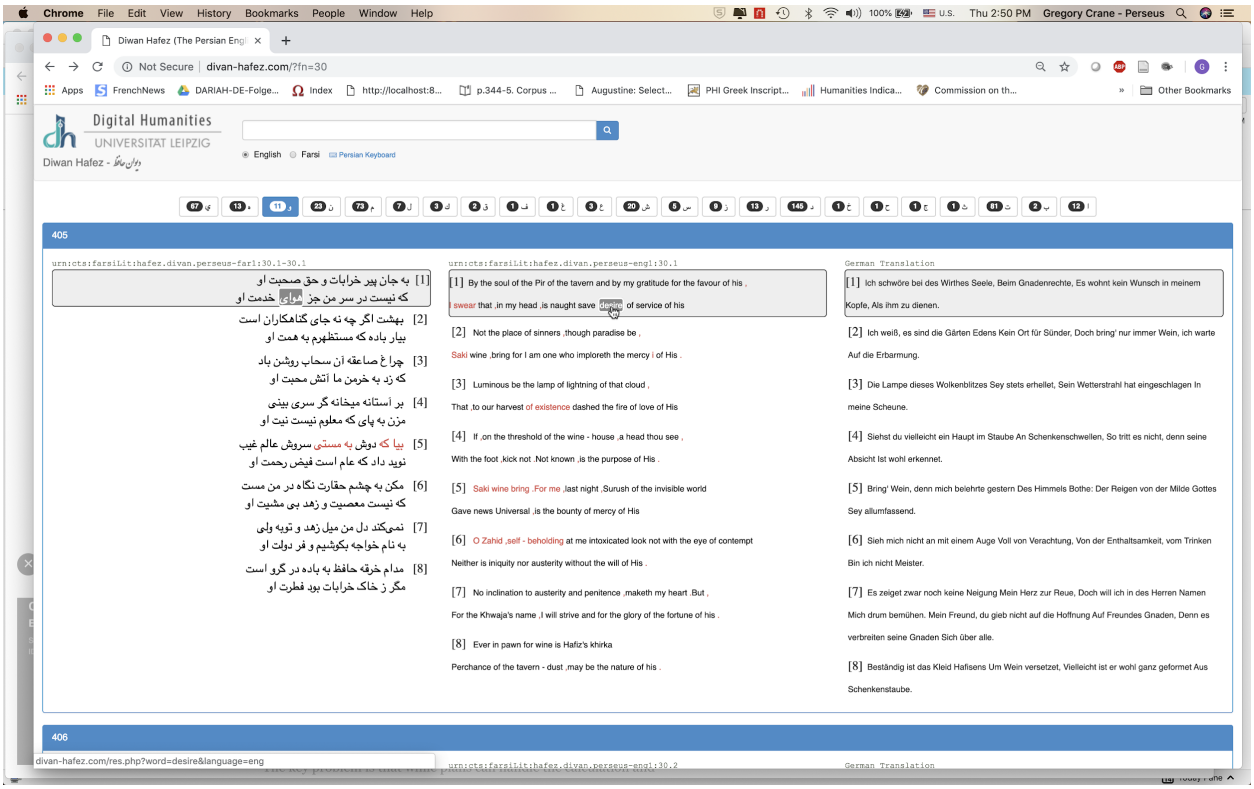

Figure 10. A reading environment aligning English and German translations to a Persian original (the poetry of Hafez, Yousef 2019b).

In Figure 11 the reader sees what Persian words have been translated as "desire." The same function will work for the reader who wanted to see how *politismus* had been translated elsewhere (e.g., Persian words translated into English as "desire").

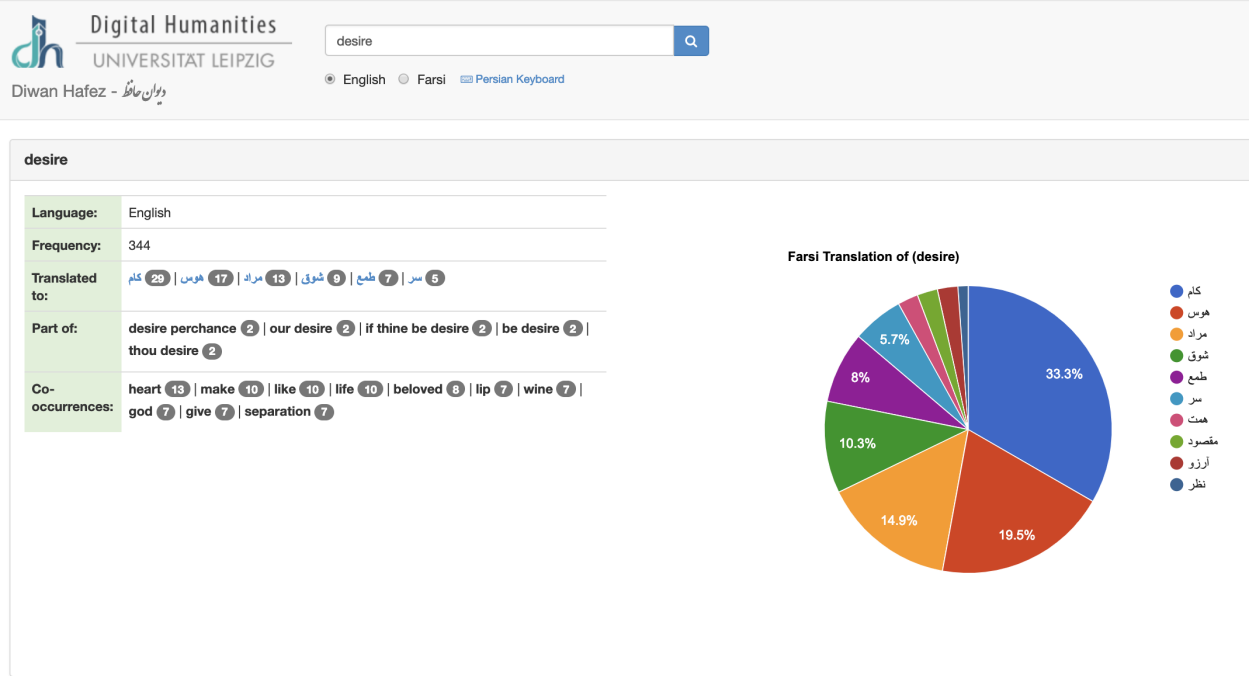

Figure 11. Statistics from the site in the previous figure about different Persian words and phrases aligned to the English term 'desire.'

At this point, the reader drills down into the text and begins to examine passages where "dark" appears and to compare the Persian source text in passages such as those in Figure 12.[9]

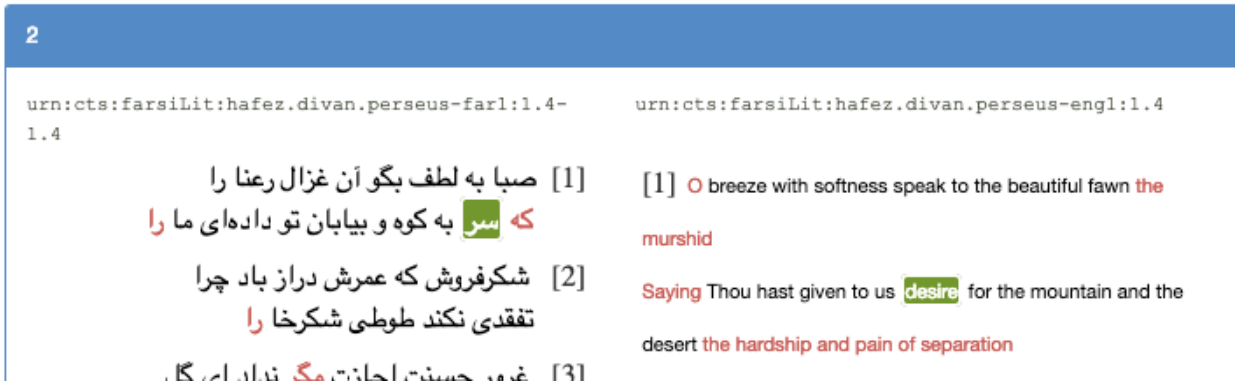

Figure 12. The reader explores the Persian equivalents to 'desire' in different contexts.

3. *Synchronized performance and metrical analysis.* While we have, by definition, no native speakers of historical languages and can only reconstruct models of how they were pronounced, integrating one or more reconstructions of the sound can (some would argue, must) be part of our understanding. This is certainly the case for poetry, where the metrical form and sound are crucial—in many traditions, poetry was always produced for performance and must be experienced in one or more reconstructions. Recordings have been available for generations and many older faculty experienced synchronization of words and sound in 'follow-the-bouncing-ball' cartoons when they were children. We can now weave such synchronization—along with visualizations of the underlying metrical patterns—into a reading environment. In Figure 13, an expert in Greco-Roman studies has created such an environment, including synchronized recordings of his performance for thousands of lines of Ancient Greek and Latin poetry.

The screenshot below provides a static view of an analysis of Greek metrical form. The accompanying video (IDENTIFIER) illustrates a reader viewing a standard text of Homer who shifts to this website to hear, as well as to see, the metrical form. A more advanced user experience would (and will ultimately) make the metrical form and performance available without context switching.

Visit the web version of this article to view interactive content.

Figure 13. David Chamberlain, metrical analysis and reading of the opening books of the *Iliad* (Chamberlain, 2019: http://hypotactic.com/homer/iliad1.htmlf).

Click to play:

4. *Text reuse detection.* Returning to the use case of Livy, the reader can exploit data already available from the Proteus Project that shows what passages in Livy have been quoted among more than 3 million scanned books. Proteus already provides a front end to visualize the data, as in Figure 14.

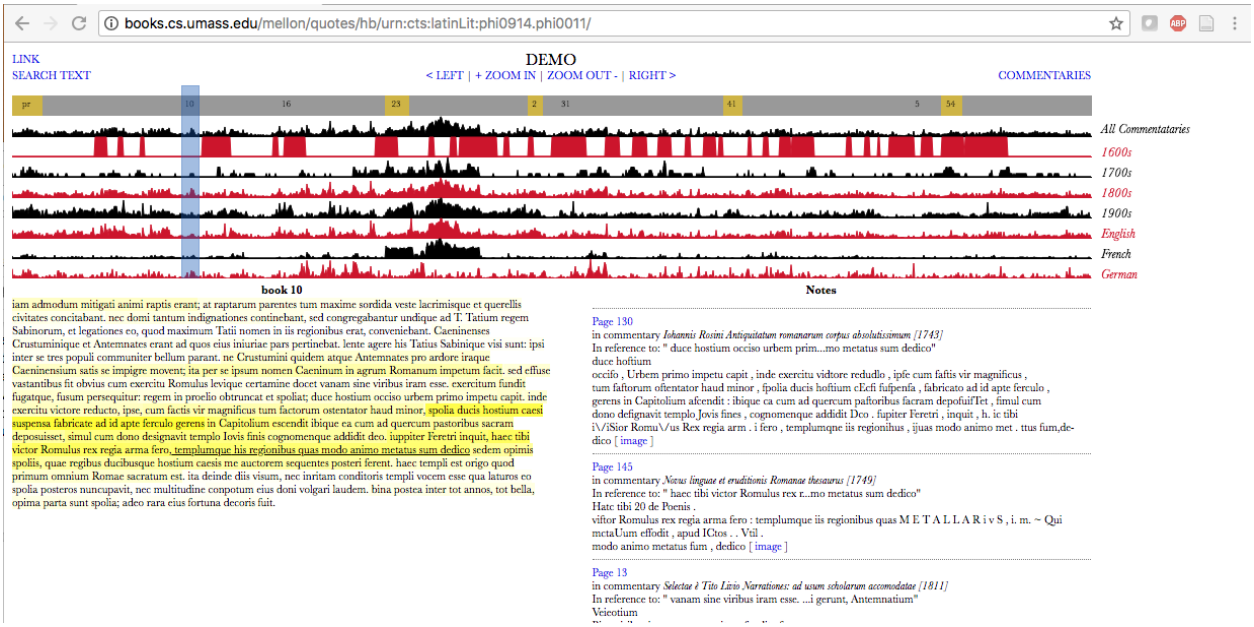

Figure 14. Automatically generated data about how frequently and where portions of a document have been quoted in uncorrected OCR-generated text from a collection of more than 1 million digitized books (Allan, Manmatha, Smith, & Zrozinski, 2019).

The reader drills down into the section of interest. The darker yellow illustrates which passages have been most quoted. Clicking on these passages reveals the uncorrected OCR text that was matched. Clicking on the page numbers calls up an image of the source text in the Internet Archive.

Visit the web version of this article to view interactive content.

Figure 15. Page image from one of the sources that a reader calls up to verify and more generally examine the context of the automatically detected quotation.

At this point, readers will (hopefully) find the illustration in Figure 15 both useful and frustrating. The illustration is useful because it shows how a reader can shift from the uncorrected OCR-generated text that supported automatic text reuse detection. But the image of a static printed page could (and ideally should) also leave readers of this article frustrated, wishing for the digital reading tools by which to understand and contextualize the text above. But if readers find it helpful to interpret the human texts from the past with maps, or syntactic analysis, or aligned translations, so much more will these tools be useful for making sense of the output of machine learning systems. When the users of data-driven systems shift from seeking *the answer* to considering evidence, they follow earlier readers down the path of scholarship.

4. Intensive Reading, Machine Learning, and Human Contributions

The examples provided show how readers have begun to interact with source texts in ways that have only become possible in a digital age. The following list suggests some of the ways in which machine-learning and human contributions reinforce one another to produce rich annotations and support new forms of intensive reading. The previous examples and this list are not in any way exhaustive, but they provide concrete examples of what has actually begun to change. While many research questions in a variety of fields emerge from this work, our biggest need is for software engineering: if we could simply integrate what is already available in fragmented systems, we would have something transformative. All of the data and services that we need for the first reading environment designed from the start for a new culture of digitally enabled intensive reading are available under open licenses.

The following describes six areas of work where hybrid human/machine contributions are needed and can enhance new forms of deep reading.

1. Examine far more detailed images of source documents than could be distributed in print.

- **Machine-learning**: Handwriting recognition and/or OCR to extract machine readable text from page images and then to generate a link from one or more transcriptions and regions of interest on the page.

- **Human Contributions**: Members of the community provide the training data. Where complex cultural documents appear in multiple different versions, readers correct OCR data or transcribe manuscripts as part of traditional, slow, and methodical close-reading. As readers are able to contribute training data and to improve the quality of automated generation of useful textual data, we cultivate a new synergy, making careful private reading produce publicly useful results.

2. View not only side-by-side source texts and translations but alignments between words and phrases in source text and translation that can be produced and viewed in the Ugarit alignment system developed at Leipzig (http://ugarit.ialigner.com/).

- **Machine-learning**: There are a range of methods to align words and phrases in source text and translation.
- **Human Contributions**: Readers can create fine-grained alignments of words and phrases between preexisting translations and source texts as part of a new generation of interactive reviews that reveal far more about the relationship between source text and translation than was possible in print. Readers can, however, also provide useful information by aligning syntactically meaningful units in the source text with chunks of preexisting translations. Automated methods can then perform much more precise analysis by comparing sentences and generate higher quality automatic alignment data. Translators can also produce born-digital translations designed to align as precisely as feasible with the source text. We may see the return of editions that include both literal and free translations to provide two different views of the source.

3. View commentaries about a particular passage, including not only modern commentaries (such as those available from Perseus), but ancient annotations preserved on medieval manuscripts[10] and early modern scholarly annotations (such as those published by the Archaeology of Reading project developed by Johns Hopkins and Princeton).

- **Machine-learning**: Methods such as topic modeling, text-reuse detection, and various forms of text mining (e.g., identifying language typically attributed to different genders or classes) can detect significant features in individual passages that can, in turn, highlight features upon which readers can focus.
- **Human Contributions**: The results of all of the automated methods can be refined to improve subsequent performance. More generally, we should expect to see a growing range of visualizations that increasingly reflect new ways to look substantively at particular passages of a text (and fewer visualizations that are produced because the technology supports them).

4. View linguistic annotations that play a core role in making sources intellectually accessible: for example, annotations that describe the morphological analysis of each word in the source text as are available in Perseus or the precise syntactic function of each word in a source text.

- **Machine Learning**: Data-driven part of speech taggers and automatic syntactic analyzers have been developed for many years. Part of speech taggers have produced about 90% accuracy with languages that have limited training data such as Greek and Latin. Syntactic analysis is much more challenging, both for human readers and machines, but can still produce useful output.
- **Human Contributions**: Many of the same linguistic features that are difficult for automated parsers are difficult for human annotators as well. Improving the results of automated analysis actually provides readers with an opportunity to focus on difficult features of the text and, at least in theory, improve their ability to understand what they read by focusing them on key points of difficulty.

5. Read the dictionary entries associated with particular words as are available in openly licensed dictionaries.

- **Machine Learning**: We can aggregate the results of automatic linguistic annotation and of translation alignment to detect shifting major syntactic patterns and different word senses associated with particular words over time. Word embeddings allow us to detect word senses within monolingual corpora. Where we have elaborate lexica for some heavily studied canonical texts, we will never have such manually curated resources for the vast majority of our sources—even with a language such as Latin, the billions of words of post-classical Latin already available in openly licensed digital form demand new, scalable automatically generated dictionary data.
- **Human Contributions**: We can expect to see a new generation of born-digital lexica, aimed at particular texts and corpora, building on the automatically generated results, with human lexicographers developing the skills to advance the power of the automated methods.

6. View maps of people and places mentioned in the particular passage along with the relationships between them.

- **Machine Learning**: We can recognize and classify proper names in a range of languages with reasonable accuracy (e.g., we can distinguish where Washington describes a person versus a place; the accuracy depends on the language, corpus, and training data). We can also, though with less success, attempt to associate a name in a text with an entity in the real world (e.g., distinguish references to Washington, DC, and the state of Washington from the other towns and cities named Washington). This is a much harder task (as anyone who has hunted through many identical names in Facebook has encountered).
- **Human Contributions**: Most references in well-understood source texts are easy to identify and to associate with a real-world entity. But there is always a subset of passages that are ambiguous because of the phrasing or (as in the case of many ancient source texts) because we lack information that an earlier audience took for granted. Determining to which Antigonus or Antonius a passage

refers is, in fact, an object of serious scholarly inquiry and part of a subfield of history called prosopography (literally 'writing about people').

5. Distributed Philology and an Interconnected Humanity

I would like to conclude by suggesting a possible way forward within the humanities that is both profoundly traditional and radically disruptive. It is radically disruptive insofar as it demands a model of intellectual production that redistributes authority and agency, engaging a far wider audience than was previously possible and producing results that are, at least with classification tasks with reasonable determinate answers, ultimately superior to what was possible in print.

First, where the sources for classification tasks are open and communities can offer new suggestions, the enthusiasm of the community and the quality of the data provide the limiting factors to the quality of the results. In some cases we can measure quality with a high degree of precision (as when we determine how accurately the individual characters from a modern printed book have been transcribed), but many, if not most, philological classification tasks involve fuzzy boundaries (as when we transcribe a damaged manuscript with idiosyncratic abbreviations or when expert lexicographers seek to distinguish different senses in a complex text). Transparent decisions may be improved over time. Reports by experts who were the only witnesses to the data, by contrast, are in some measure forever opaque (and, indeed, this is the challenge for all of us who work with historical sources).

Second, and (in my view) far more important, a new sociology of intellectual production, one that is fundamentally decentralized and potentially more egalitarian, is taking shape. This involves a shift from a bureaucracy of expertise (where specialists deliver judgments that non-experts cannot effectively assess) to a republic of discourse where each member of the community is a citizen with not just an opportunity, but an obligation, to contribute insofar as possible. In this model of intellectual production, the expansion of the intellectual franchise may constructively challenge the authority, but must also amplify the potential impact of, scarce specialist expertise. I have been helping my students without knowledge of Greek work directly with the Greek text. I cannot remember any students ever thinking that advanced knowledge of Greek would not have been a huge advantage. My hypothesis, based on personal observations, is that we could demonstrate that such immediate access only increases interest in, and respect for, understanding of the particular language.

I can already point to a first-generation citizen science project that demonstrated the reality of this new mode of intellectual production. Over the course of almost a decade, nearly 200 volunteers (mainly undergraduates in North America and Europe) produced the first full diplomatic edition of the 10th-century Byzantine Venetus A manuscript, the most important source for the text of the Homeric *Iliad*. This manuscript contains multiple categories of ancient commentary and is written in a script that is very different from the print of modern Greek books and contains a range of cryptic

abbreviations.[11] The same model has also inspired a number of other digital projects, some at least (if not more) challenging than the transcription of the Venetus A.[12]

One next step for such work is to broaden the element of international collaboration and, ideally, to move beyond the use of English—or any single lingua franca—as a requirement for participation. I offer as one example work attempting to open up Ancient Greek to an Iranian audience and Classical Persian poetry to those who do not know Persian. These two efforts depend upon the various tools outlined above, as well as upon new services such as free video-conferencing.[13] As a professor with a chair in Germany from April 2013 through September 2019, as well as in the United States, I was able to support Iranian and Syrian researchers who cannot get visas to visit, much less study in the United States. But my Leipzig seminar on Digital Reading includes participants from Tehran, as well as Leipzig and Medford, Massachusetts.

The two efforts are complementary and designed to reinforce one another. On the one hand, as a specialist on Ancient Greek, I feel that I have an obligation to disseminate understanding of Ancient Greek to the widest possible audience. Iran is particularly interesting because many of our most important sources on ancient Persian culture are Greek—from the 6th century BCE onwards, Greeks and Persians have interacted in a variety of ways, both peaceful and violent. Herodotus and Xenophon, for example, are fundamental sources for Cyrus the Great. Their accounts provoke vigorous debate, if not controversy, in 21st-century Iran where conservative Islamic thinking and Persian national identity can generate a very different view of Cyrus as brutal conqueror vs. founding figure for Iran. Neither Herodotus nor Xenophon have been translated directly from Greek into Persian and almost no Iranians have had a chance to master Ancient Greek. At the same time, Persian poetry constitutes a core element of modern Iranian identity. Iranian high school students do not read 19th- and 20th-century novels that often dominate secondary school education in much of Europe and North America but focus on their own poetry heritage. Persian poets such as Hafez and Rumi have attracted passionate and talented translators dedicated to making them accessible in other languages, but poetry, insofar as it is possible, cannot be translated. Our goal is to promote the flow of interest and of ideas in both directions. If I want to encourage people in Iran to learn Ancient Greek, I also feel that I should support the study of Persian poetry in Persian in the United States and beyond. I have my own views of Cyrus the Great but I also need to hear the varied perspectives on Cyrus from Iran. If we want to move beyond implicit cultural hegemony and toward an increasingly shared understanding of who we are as human beings, we need to support exchange rather than simple transmission. A bottom-up movement for shared cultural understanding may not immediately solve fraught relations between governments in Iran and the United States, but it is something.

Weaving new bonds of understanding across Classical Persian, Ancient Greek, modern Persian, English, and other languages is a daunting task by itself. Nevertheless, these four languages represent

only one small (if particularly strategic) cluster of flows that we must support across boundaries of culture and language. The intelligent infrastructure (II) and augmented intelligence (AI) that Michael Jordan (Jordan, 2019) stresses in his essay "Artificial Intelligence — The Revolution Hasn't Happened Yet " are essential to us as we struggle for a world where anyone, anywhere in the world, has the tools to engage immediately and then learn as much as they wish over time about any culture and language of humanity, whether it is the language in which their parents spoke to them as children or a language separated by thousands of miles and/or thousands of years.

Netflix and YouTube may have global audiences but the older, long-studied sources in languages for which there are no native speakers still can play a strategic role, as can the philologists who study them. First, there are dozens of translations of the *Iliad* and the *Odyssey* into English, specialized dictionaries and grammars explaining the language, high-resolution images of ancient scraps of papyrus from Egypt and manuscripts from Europe representing the epics in part and in full, born-digital morphosyntactic analyses for every word in each epic, and commentaries, developed over thousands of years, that seek to help readers understand these epics. When I listen to characters in a Netflix series speak Turkish and I long to understand how I might respond to what I am seeing if I had grown up in Istanbul, I turn back to sources such as Ancient Greek and Classical Persian, where we have the data to model for these languages the intellectual space that I would like to see for any song, book, poem, film, or TV series.

Second, works such as the *Iliad* and the *Odyssey* have accumulated a rich body of explanatory information—a process that began thousands of years ago in Greece and Alexandria and that only accelerated with the advent of print. The 2018 Netflix series *Troy: Fall of a City* may reach a wider audience in 2019 than the text of the *Iliad*, but the *Iliad* will continue to be read far into the future whether or not this particular television series maintains an audience over time.

I see a new beginning and vibrant potential future for the study of the past, one where we naturally integrate the science of data with our most traditional questions and our best values. For many of my colleagues, such a new beginning may seem implausible, if not unacceptable. But if we harbor concerns that the digital turn will lead to reductive thinking, where numbers alone dominate, our job is to explain, rigorously and clearly, why quantitative thinking provides at best a start—and a necessary start. Quantitative thinking may, to paraphrase a phrase attributed to Oscar Wilde, allow us to understand the price of everything and the value of nothing. It is our job as humanists—arguably our only job—to advance the public understanding of those values.

Acknowledgments

This article provides an overview of work available in fall 2019. This builds in part upon on-going funding from the Framework for Interoperability Project that has been funded by the Mellon

Foundation, that is led by the Johns Hopkins Library, and in which my colleagues at Tufts, Furman University, and Eldarion.com are collaborating: https://mellon.org/grants/grants-database/grants/johns-hopkins-university/1802-05569/. The work described here is supported by the Beyond Translation Project, which is led by Tufts University with funding from the National Endowment for the Humanities: https://www.neh.gov/sites/default/files/inline-files/NEH-Grant-Awards-August-2019.pdf; NEH grant #103048-00001.

References

Allan, J., Manmatha, R., Smith, D. A., & Zrozinski, M. (2019). *The Proteus Project*. Retrieved from http://books.cs.umass.edu/mellon/

The Archaeology of Reading. (2019). *Archaeology of Reading*. Retrieved from archaeologyofreading.org

Berti, M., & Bodard, G. (2019) *SunoikisisDC: An international consortium of digital classics programs*. Retrieved from https://sunoikisisdc.github.io/

Bond, E. (2019, May 31). Media localization buyers on outsourcing, automation, and last-minute schedule changes. *Slator*. Retrieved from https://slator.com/features/media-localization-buyers-on-outsourcing-automation-and-last-minute-schedule-changes/

Chamberlain, D. (2019). *Greek and Roman verse: Audio-visually enhanced*. Retrieved from http://hypotactic.com/

Committee of Editors of Linguistics Journals. (2015, May 31). *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.* Max Planck Institute for Evolutionary Anthropology. Retrieved from https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

DiBiase-Dyson, C., Kammerzell, F., & Werning, D. A. (2009). Glossing Ancient Egyptian: Suggestions for adapting the Leipzig Glossing Rules. *LingAeg, 17,* 343–366.

Dunbar, H. (1880). *A Concordance to the Odyssey and the Hymns of Homer* (Oxford, Clarendon Press).

Jordan, M. I. (2019). Artificial Intelligence—The Revolution Hasn't Happened Yet. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.f06c6e61

Meninski, F. (1680). *Thesaurus Linguarum Orientalium* (Vienna) (pp. 181–191). Retrieved from http://reader.digitale-sammlungen.de/de/fs1/object/goToPage/bsb10635924.html

Lattimore, R. (1951). *The Iliad of Homer* (University of Chicago Press).

Lattimore, R. (1967). *The Odyssey of Homer* (New York: Harper and Row).

Livy, T. (1555). *Romanae Historiae Principis* https://archaeologyofreading.org/viewer/#aor/PrincetonPA6452/1r/image.

Muellner, L. (1996). *The anger of Achilles: Mênis in Greek epic* (Cornell University Press). Retrieved from http://nrs.harvard.edu/urn-3:hul.ebook:CHS_MuellnerL.The_Anger_of_Achilles

Murray, A. T. (1924). *The Iliad with an English Translation in two volumes*. (Cambridge, MA., Harvard University Press).

Perseus Digital Library. (2019). Retrieved from http://www.perseus.tufts.edu/hopper/

Prendergast, G. L. (1874). *A Concordance to the Iliad of Homer* (London: Longman, Green and Company).

"Recogito: semantic annotation without the pointy brackets," a partner of the Pelagios Network. (2019). Retrieved from https://recogito.pelagios.org/

Yousef, Tariq (software design) and Foradi, Maryam (content preparation). (2019). *The Diwan of Hafez aligned to English and German translations*. Retrieved October 6, 2019, from http://divan-hafez.com/

Topoi. (2019, February 13). Glossing Rules. Humboldt University Berlin. Retrieved from https://wikis.hu-berlin.de/interlinear_glossing/Glossing_Rules

Troy: Fall of a City (2018). (Netflix): https://www.netflix.com/title/80175352.

Wolf, M. (2018, August 25). Skim reading is the new normal: The effect on society is profound. *The Guardian*. Retrieved from https://www.theguardian.com/commentisfree/2018/aug/25/skim-reading-new-normal-maryanne-wolf

Footnotes

1. The excerpts from the English are screen shots from the Kindle version. This translation only shows every fifth line number and thus the five excerpts that do not have a line ending in 0 or 5 have no line numbers. ↩
2. For a discussion of the Canonical Text Services, see http://cts3.sourceforge.net/; for a set of services built on CTS, see http://capitains.org/. ↩
3. The integration is manual and we have to move across different sites, each of which specializes in a particular form of annotation. Although each is composed of open data and open source software, they do not yet interact via stable application programming influences (APIs) and are effectively siloes. Nevertheless, because the particular work described in this paper is open and because there

is an active community carrying each category of work forward, the prospects of developing an integrated reading environment are very good—certainly, if we assume that open reading environments can confer a Darwinian advantage upon sources as they compete for attention among new generations of readers. ↩

4. On the Scaife Viewer, see http://sites.tufts.edu/perseusupdates/2018/03/15/first-version-of-the-scaife-digital-library-viewer-goes-live-building-the-future-while-remembering-a-friend/. The Scaife Viewer was named after the late Digital Classicist, Ross Scaife (1960–2008): https://en.wikipedia.org/wiki/Ross_Scaife. ↩

5. https://scaife.perseus.org/reader/urn:cts:greekLit:tlg0012.tlg001.perseus-grc2:1.1-1.7/?highlight=%CF%84%CE%B5%5B5%5D. Much of the Greek material in the Scaife Viewer comes from Open Greek and Latin, a collaborative effort in which Tufts, the University of Virginia, the Harvard Library, Harvard's Center for Hellenic Studies, Mount Allison University, and Leipzig University have collaborated to add more openly licensed Greek and Latin textual data: http://www.opengreekandlatin.org/. ↩

6. In this URN: greekLit defines a namespace set aside for Greek literature; tlg0012 designates Homer and tlg0012.tlg001 designates the Homeric Iliad; perseus-grc2 designates the particular print edition from which the digital version was derived; 1.1-1.7 specifies lines 1-7 of book 1. ↩

7. https://www.perseids.org/tools/arethusa/app/#/perseids?chunk=1&doc=34086; this analysis was done by Julia Lenzi as part of a digital edition of an oration by Lorenzo Valla on the importance of Latin to European culture: Johannes Vahlen, *Laurentii Vallae opuscula tria* (Vienna 1869): https://babel.hathitrust.org/cgi/pt?id=njp.32101064184920&view=1up&seq=161. ↩

8. In the aligned Latin and English derived from the Archaeology of Reading example, the reader can also note how *sapientia* had been translated into English in different passages: http://ugarit.ialigner.com/wordinfo.php?w=sapientia&lang=LAT. ↩

9. Note that the text in red has been judged by the human aligner (PhD student Maryam Foradi) not to have equivalents in the Persian. The amount of unaligned text provides readers who do not know Persian with information about how close a particular translation is to the original. ↩

10. Students, with some faculty support in the Homer Multitext Project, have, for example, collaborated to produce transcriptions of thousands of such ancient annotations from the medieval manuscripts in which they are preserved: http://www.homermultitext.org/about/. ↩

11. For an overview, see http://www.homermultitext.org/; for the openly licensed diplomatic edition data, see https://github.com/homermultitext/hmt-archive/tree/master/archive/iliad. ↩

12. See, for example, the projects at http://hcmid.github.io/ and https://github.com/ChiaraPalladino/TuftsDCC/wiki. ↩

13. Such interactions are already underway: Berti & Bodard, 2019. ↩