Extracting Two Thousand Years of Latin from a Million Book Library

David Bamman
The Perseus Project, Tufts University
and
David Smith
Department of Computer Science, University of Massachusetts-Amherst

With the rise of large open digitization projects such as the Internet Archive and Google Books, we are witnessing an explosive growth in the number of source texts becoming available to researchers in historical languages. The Internet Archive alone contains over 27,014 texts catalogued as Latin, including classical prose and poetry written under the Roman Empire, ecclesiastical treatises from the Middle Ages, and dissertations from 19th-century Germany written – in Latin – on the philosophy of Hegel. At one billion words, this collection eclipses the extant corpus of Classical Latin by several orders of magnitude. In addition, the much larger collection of books in English, German, French, and other languages already scanned contains unknown numbers of translations for many Latin books, or parts of books.

The sheer scale of this collection offers a broad vista of new research questions, and we focus here on both the opportunities and challenges of computing over such a large space of heterogeneous texts. The works in this massive collection do not constitute a finely curated (or much less balanced) corpus of Latin; it is, instead, simply *all* the Latin that can be extracted, and in its reach of twenty-one centuries (from ca. 200 BCE to 1922 CE) arguably spans the greatest historical distance of any major textual collection today. While we might hope that the size and historical reach of this collection can eventually offer insight into grand questions such as the evolution of a language over both time and space, we must contend as well with the noise inherent in a corpus that has been assembled with minimal human intervention.

Categories and Subject Descriptors: H.3.7 [**Information Systems: Information Storage and Retrieval**]: Digital Libraries

1. INTRODUCTION

In June 2010, Google released over 500 high-quality scans of major Greek and Latin works, curated with the help of Gregory Crane and Alison Babeu at the Perseus Project.[1] This collection included a carefully selected group of texts – largely authors from the Classical canon – drawn from the much deeper recesses of Google Books, which at the time had digitized a total of ca. 12 million works [21].

While this carefully selected set of texts with high-quality metadata stands on its own as a classical example of a curated corpus, those darker and more chaotic depths have the promise to yield up a far greater and potentially more valuable set of data. The Internet Archive contains a smaller set of digitized works (ca. 2 million), but all of them are publicly available for download, and 27,014 of these works have been catalogued as Latin from a range of authors, genres, and eras

[1] http://www.google.com/googlebooks/ancient-greek-and-latin.html

– the Classical Latin works of Vergil and Cicero, medieval religious authors such as Augustine and Thomas Aquinas, and later scientific writings by the likes of Newton, Copernicus and Kepler. These 27,014 works contain approximately one billion words of Latin, far more than the extant corpus of Classical Latin up to ca. 200 CE (around 10 million words[2]) and larger still than the largest existing Latin collection (J. Ramminger's *Neulateiniche Wortliste* [23] at 300 million words), which includes works up to 1700 CE. These 27,014 works also span a total of twenty-one centuries, capturing not only the written native Latin of a Roman elite but also its use as a second language of writers for the two millennia that follow.

As others have pointed out, however, problems plague these massive collections in their use for scholarly research, not only in the quality of the image scans and the resulting OCR but also in the metadata itself that describes the texts [18]. While a certain degree of error is to be expected among such large projects focusing on optimizing the usefulness of the average case [14], the first research question we must ask is what tasks such a huge collection is best suited for in the face of such noise. While not a curated or balanced collection in the same vein as the Brown Corpus [13] or a homogenous collection of canonical Classical authors, such a large textual collection has the potential to give us the ability to begin investigating the complex mechanics that govern Latin usage over those two thousand years – a period that involves not only significant linguistic change and a shift in use as a language of native speakers to second language learners, but two thousand years of historically trending ideas as well.

2. PROCESSING

To begin looking into these questions, we downloaded all of the 27,014 texts in the Internet Archive that had been classified by a librarian as being written in Latin. Here at once we are confronted with the metadata problems described elsewhere. While many of the works catalogued as Latin are in fact written in that language, 15.7% of a 9,203-work subset that we looked at (see below) were incorrectly marked as such, from Dutch botanical treatises to works in Ancient Greek with a Latin title. Additionally, since almost all of these digitized works are modern editions of historical texts, many contain large sections of text in modern languages such as English, German, French and Italian – not only a preface or introduction at the beginning of the work, but occasionally notes scattered throughout as well. In order to solve both problems – that of incorrect or incomplete metadata and of mixed works – we trained an n-gram language classifier [26] on the complete texts of Wikipedia in 24 different languages and the Ancient Greek and Latin collection of the Perseus Digital Library [6; 8; 7] and used that trained model to create a language fingerprint for each text, allowing us to pinpoint exactly where the Latin showed up in each volume.[3]

[2] 10 million words is an upper bound based on the archive of the Thesaurus Linguae Latinae (TLL), which houses a lexicographical slip for each occurrence of a word in texts from the Classical period [27]. The Packard Humanities Institute CD ROM of Latin, which is fairly comprehensive through 200CE and contains some later materials, holds c. 7.5 million words.

[3] In the future, we would like to apply this same language identification to the entire collection of books – not simply those manually classified as Latin – in order to detect the Latin in works

2.1 Enhancing the metadata

In addition to the major languages of the work, the library records that attend each text contain a wealth of other contextual information, including author, title, publisher, subject classifications, and the date of publication. One of our goals in this project, however, is to be able to track the spread of linguistic features within a language and ideas across languages over the two millennia that Latin was used as a *lingua franca* across Europe. While much of this research operates on the textual data itself, the ability to chart such movement in both space and time requires information on the place and date of a work's composition. The library records available to us, in contrast, report the place and date of publication for a specific edition – which, for historical texts, is often far removed from the time and place of original composition. For establishing the differences in usage between the Latin of Vergil's *Aeneid* and that of Jean Calvin's *Institutio Christianae Religionis*, it is far more important for us to know that the former was composed ca. 19 BCE and the latter in 1536 CE than the date of any later printed editions.

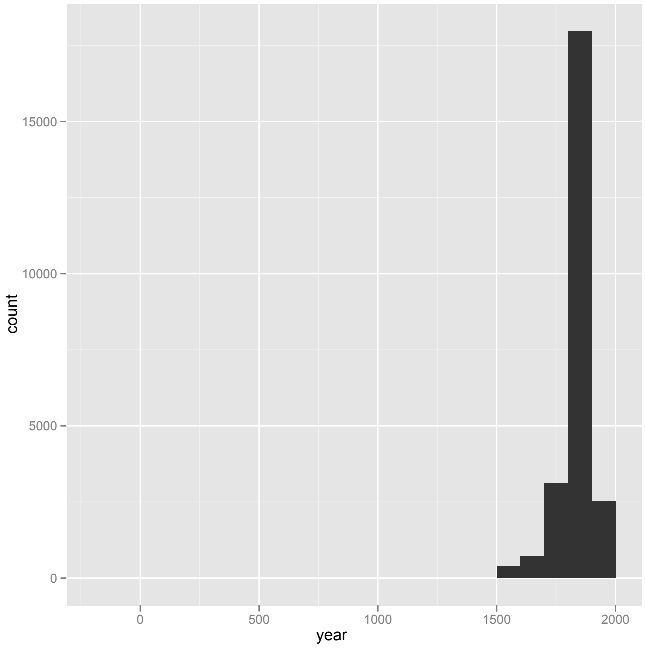


Fig. 1. 27,014 works catalogued as Latin in the Internet Archive, charted by date of publication.

Figure 1 charts the dates of publication for the 27,014 Latin works in this collection – all of these printed editions are of course dated post-Gutenberg, with the vast

where the language has either been misclassified (as a false negative) or the Latin appears too infrequently to be counted as a major language within it.

majority of them coming from the 19th century (i.e., after the Industrial Revolution but still in the public domain).

While this data is not without merit – it may be useful for researching questions on the changing publishing dynamics for Latin editions – what we are interested in is knowing when each of these works was originally composed. To address this, we are supplementing the existing metadata by researching the date of composition for each work. While the text of some authors (like Vergil and Calvin) have more-or-less established composition dates, others (such as those of more obscure medieval authors) do not. The task here – undertaken by three student researchers[4] – is to delimit the smallest time window possible given the state of current research on each author (e.g., if not a single year such as 1536, then a window of 1530-1540 etc.). At publication time, approximately 34% of the catalogue (9,203 works) has been completed – 2,818 works have been excluded as not being in Latin or not containing a significant amount of natural language (e.g., lists of manuscripts) while 6,385 have been dated.

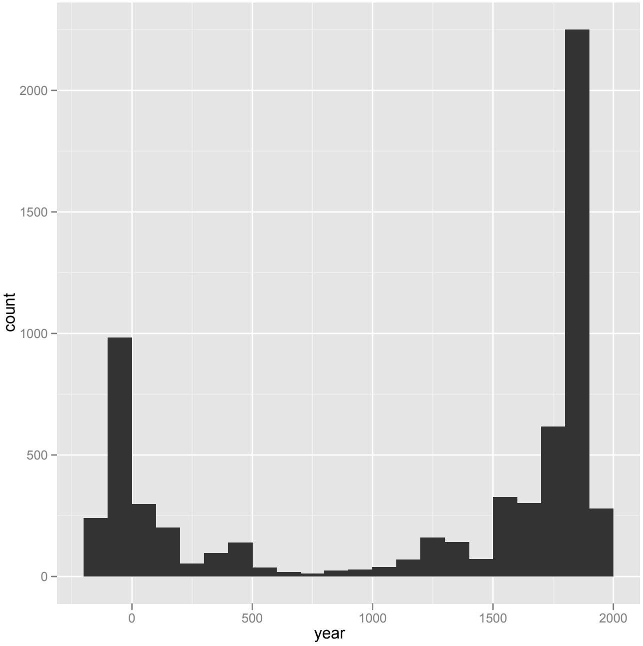


Fig. 2. 6,385 Latin works in the Internet Archive, charted by date of composition.

Figure 2 charts the resulting dates of composition for those 6,385 works. Major peaks immediately rise up around the Classical era of ca. 200 BCE–200 CE (including authors such as Cicero and Vergil), the works of church fathers such as Augustine (ca. 400 CE), and voluminous scholastic writers such as Thomas

[4] Thanks are due to Elise Goodwin-Tuchmayer, John Owen and Erin Shanahan and for their invaluable research assistance.

Aquinas (ca. 1200 CE), before yielding to an explosion in the number of printed works following the invention of the printing press.

From these 6,385 works we extracted all and only the Latin texts to create a historically dated corpus for subsequent analysis, comprising a total of 389 million words.

3. ANALYSIS

With a catalogue of actual dates of composition, we can begin to investigate the rise and fall of Latin topics and linguistic features over the course of its two-thousand-year lifetime. The result is not at all the traditional balanced corpus that has stood at the foundation of research in corpus linguistics – this historical collection contains a mix of authors, genres, and eras, and often several editions of core works. The library metadata gives us one possible method of organizing this disparate set of works, allowing us to assemble ad hoc corpora [9] – dynamic collections of texts that can be automatically compiled in response to a query. This would allow us to create and query a corpus restricted to, for example, 18th-century works on natural history or Classical Latin prose written between 100 BCE and 100 CE. Here the subject classifications for each work are crucial; we note, however, that in the data so far examined, only about 30% of the books in our 27,014-work collection have such explicit classifications. One method that may be worth pursuing is supplementing this metadata with the results of topic modeling [16] in order to automatically assign a set of subject classifications based simply on the words each work contains. This would allow us to take what would otherwise be an undifferentiated mass of text and still be aware, with a fine level of granularity, of its characteristics as a corpus.

3.1 Small sample: America

As a first initial exploration, we looked at mentions of “America” in this dated collection. In the 6,385 works in our corpus, “America” occurs a total of 3,085 times; however, the majority of these mentions occur in a work’s front matter, introduction, or notes in English and other modern languages. In the just the text that has been idenfified as Latin, “America” occurs 1,006 times. We see that its appearance in the historical record (figure 3) sees a sharp rise in the century after its discovery – appearing characteristically in texts ranging from jurisprudential dissertations (Huppé’s 1875 *De Civitate Discreta*) to insect catalogues (Friedrich Weber’s 1801 *Observationes entomologicae*). The first observed instance within this subset comes from Ulrich von Hutten’s poem *In Venetos Exhortatorium* to the Holy Roman Emperor Maximilian I (ca. 1516).

We can also see, however, how noise permeates this collection as well, especially when dealing with such small samples. Of the 1,006 total mentions of “America,” 20 appear before the 1500s (ca. 2.0%). Interestingly, none of these errors come from language misidentification or dating mistakes: all of them appear in notes or introductions written in Latin by modern editors, which suggests that we may need to use a finer level of granularity in our dating – instead of assigning a single date to a work, we may want to associate dates with specific sections instead. In a corpus that is subject to layers of automatic processing (not simply language identification but errorful OCR as well), such noise is to be expected, but this may be one approach to dealing with this particular weakness.

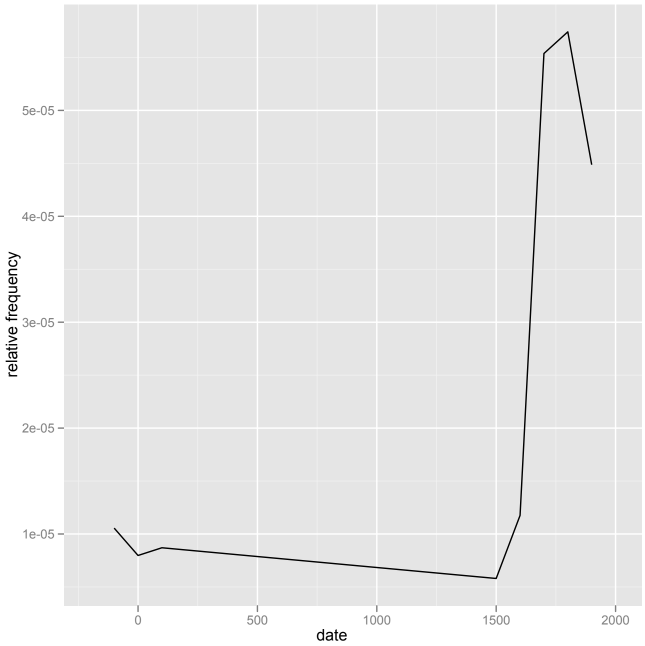


Fig. 3. Relative frequency of *America* (1,006 instances).

3.2 Large samples: prepositions and conjunctions

Once we begin to look at words with much larger frequencies, such as function words like prepositions and conjunctions, we can begin to the see the impact of investigating linguistic features over such a long period of time. A heterogeneous Latin corpus spanning two thousand years presents several complicators: not only is the corpus comprised of many different genres (erotic poetry, religious treatises, and insect catalogues, to name only three) and covers distinct eras of the language (Classical, Medieval/Ecclesiastical, Renaissance, Neo-Latin), each with their own linguistic characteristics, but it is used as a *lingua franca* by an increasingly diverse linguistic community, including speakers of both early German and Romance languages. When we note the rise and fall of words and features here over the course of two thousand years, we are looking at a combination of linguistic change in Latin, differing levels of influence of the writer's first language, and trends in style conditioned by language as well as genre – in short, what we are investigating is none of these variables in isolation, but the mechanics of their interaction.

The changing use of prepositions and conjunctions within Latin illustrates this complex phenomenon. Functional words like these are among the most frequent in any language [11], so we have little problem with noise in the collection – the preposition *de* ("from") occurs almost two orders of magnitude more frequently in this corpus than "America" above. Figures 4, 5 and 6 chart the changing frequencies of the top three prepositions – *de* ("from"), *ad* ("to"), and *in* ("in") – which all reveal the same pattern: a bursty but constant rise toward 1200-1300 CE, followed by a precipitous drop past 1500.

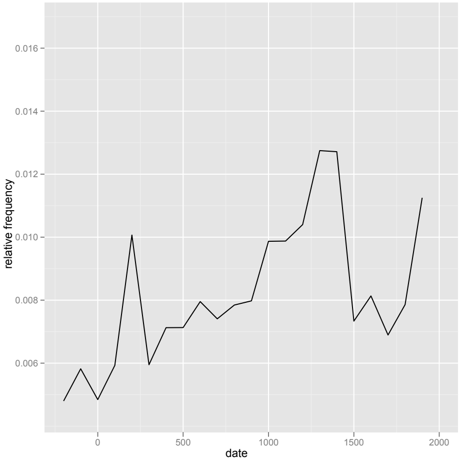


Fig. 4. Relative frequency of *de* (“from”) (2,955,462 instances).

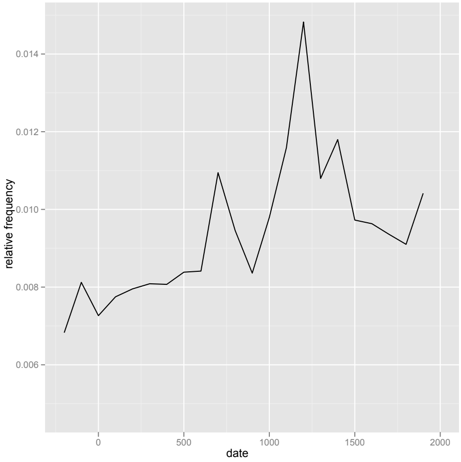


Fig. 5. Relative frequency of *ad* (“to”) (3,655,191 instances).

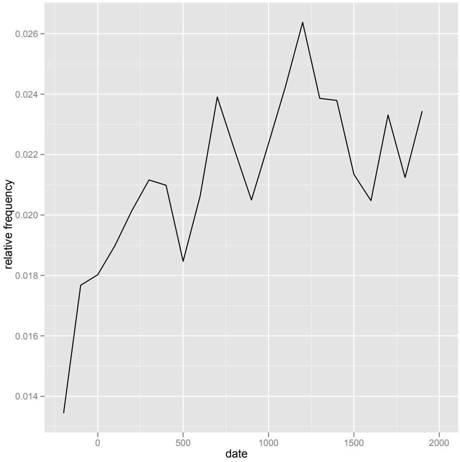


Fig. 6. Relative frequency of *in* (“in”) (8,126,487 instances).

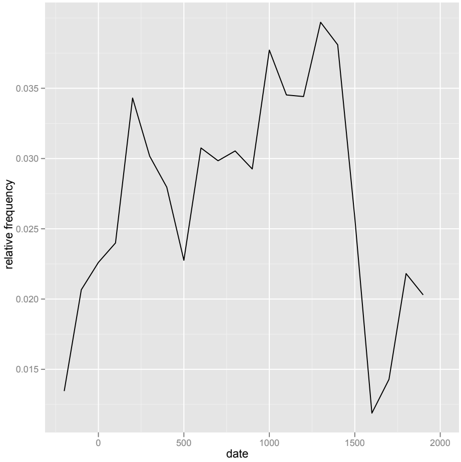


Fig. 7. Relative frequency of *et* (“and”) (9,317,773 instances).

From the end of the Classical period up through the the High Middle Ages, prepositions are used with increasing frequency (the slope of this ascent averages an increase of 30%-40% each year). One possible explanation for this rise could be the loss of case markings in the underlying vulgar languages; without the ablative case to denote separation from or the dative case to indicate direction towards, prepositional phrases could be used to compensate. As with the changing frequency of *accusativus cum infinitivo* constructions, this involves the influence of cross-linguistic transfer between the writer's native language and the Latin.

In figure 7, however, we note that the very common conjunction *et* ("and") manifests this same behavior. Beginning with the earliest Latin in the collection (Plautus), we see a sporadic but constant rise in the relative frequency of this word, all the way to the 1300s, when we see an even more dramatic drop to 1600. This rise in *et* is no doubt partly stylistic, a transition from the highly compact style of Classical Latin oratory and poetry to the more verbose style of ecclesiastical Latin. Jerome's *Vulgate* (ca. 400 CE), with its heavy use of polysyndeton, is perhaps the best exemplar of this trend, in which *et* constitutes a whopping 8.3% of the total words (this is ca. 40% more frequent than "the" in the contemporary English of the British National Corpus).

This may help explain the transition from a lean Classical Latin style to a more verbose ecclesiastical one, but what about the fall around 1500 CE? Style here may point toward an answer as well. The collection of the 1200s is dominated by scholastic authors such as Thomas Aquinas and Bonaventure. In the 1500s, the most frequent authors include German religious figures such as Martin Luther and his follower Ulrich von Hutten, but we also see a rise in scientific works such as those by the astronomer Tycho Brahe and the physicians Andrés Laguna de Segovia and Antonio Musa Brasavola.

The question that remains is whether the introduction of this new genre brings with it a tendency toward compact prose as well. Even in the Renaissance, there is an impetus among Neo-Latin authors to imitate Classical Latin style, especially that exemplified by Cicero [12]. One method of measuring this density is by calculating the standardized type/token ratio (TTR) of a text – i.e., the ratio of unique forms in a passage to the number of total words it contains. A low TTR means that the same words appear more than once; a high TTR of 1.0 means that every word in the passage is unique. This method is commonly used as a simple measure to evaluate vocabulary richness (e.g., among children who stutter [24]), but by assessing the TTR over all of our texts, we can use it to measure textual density and lexical variation as well.

Figure 8 presents the results of this evaluation. Since TTR has an inverse correlation with text length (longer texts have more repeated words), we normalize it by dividing each century into 10,000-word sections and calculating a standardized TTR as the average TTR of each 10,000-word section. The results match our anecdotal assumptions about Latin: Latin style has the highest density with Classical Latin ca. 100 CE and the Neo-Latin of ca. 1600 CE, and the lowest density in the Middle Ages, with low points ca. 600 and 1300 CE.

The fact that these valleys and peaks correspond to inverted peaks and valleys in the changing relative frequencies of prepositions and *et* points to a possible relationship between the two when dealing with datasets of this size: while relative frequency measures the frequency of use of a word compared to the use of all other words, there are two underlying causes for observing a change. One cause for a decrease, for example, may be the simple fact that a word is used less: "Prussia" has a lower relative frequency today than it did 100 years ago due to the simple fact that it's used less in conversation. Another underlying cause of a change is the

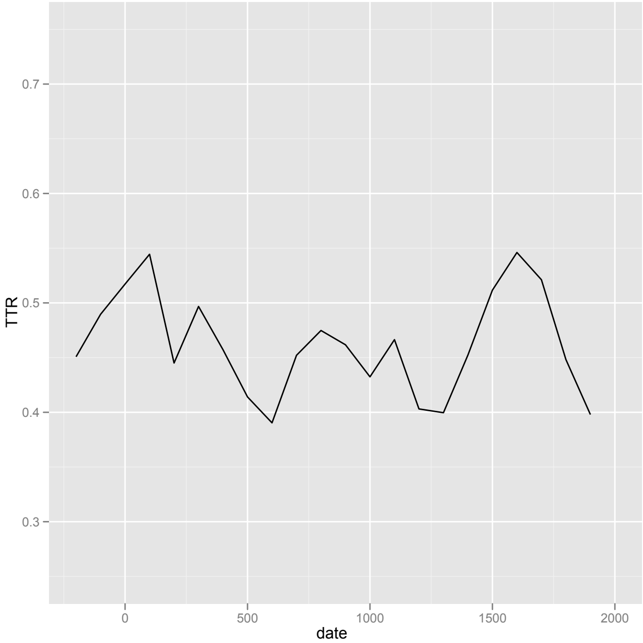


Fig. 8. Vocabulary density (Standardized Type-Token Ratio).

fact that everything else, so to speak, is used more. The decrease in prepositions and *et* may be an instance of this – it is not the case that either is going out of style, but rather that a stylistic increase in density (and variation) leads to an overall decrease in every word's relative frequency. It is only in looking at a such a long historical timeframe that we can see this effect unfold, as a result not only of linguistic change, but large-scale changes in stylistic diversity as well.

4. NEXT STEPS

With our current collection and metadata, we can investigate changing patterns in word use in Latin over a period of two thousand years, including variables such as individual relative frequency and vocabulary density, both token based. This is only the very beginning, however, and there are a number of research directions we'd like to pursue to extract more information out of this broad (and deep) collection.

4.1 Adding geographical information

One strong desideratum is being able to track how words and ideas spread across Europe through the vehicle of Latin. By supplementing the metadata records with each work's date of composition, we are making a first step in this direction. The next step in being able to track the flow of information is adding spatial data to this temporal dimension.

The metadata for each work includes the name of the publisher and the place of publication as well. As with the date of publication, however, this information does not pertain to the original composition but to the specific printed edition (e.g., a new edition of Newton's *Principia* was printed in 1822 by A. & J. M. Duncan in Glasgow, while the original *editio princeps* was published in 1687 in London). By researching the place of publication of the original printing, we can begin to track how these same features move not only across time, but across space as well.

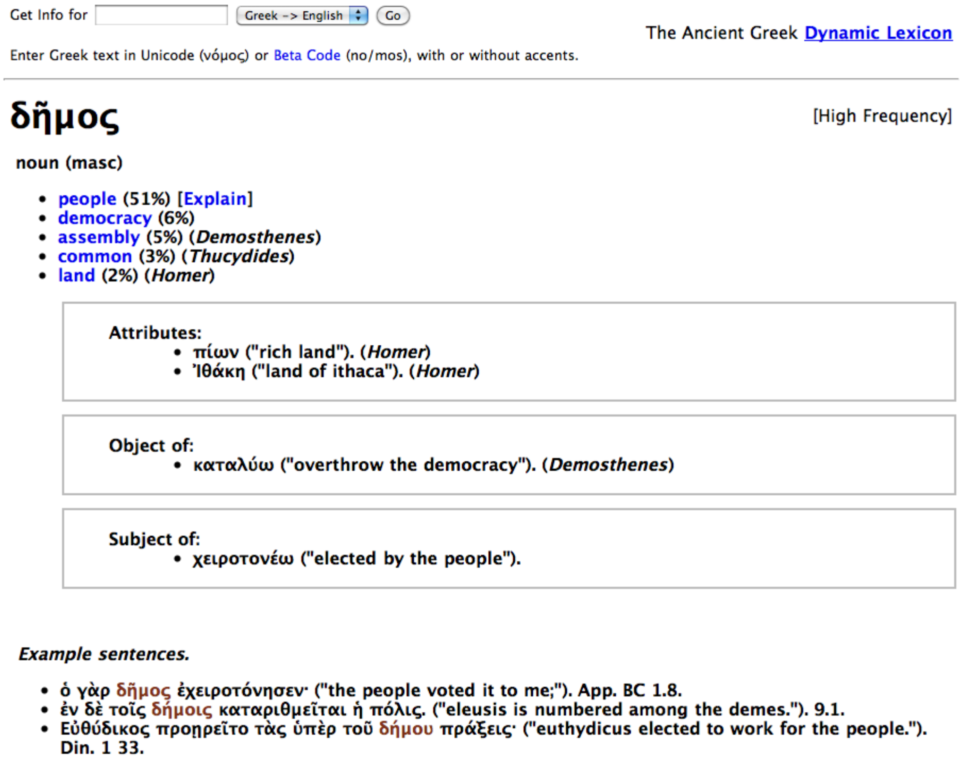
4.2 Duplicate detection

The older a text is, the more likely it is to have been printed in multiple editions. While a dissertation from 1871 may only have seen the light of day once, the Internet Archive contains over 300 different works in Latin attributed to the poet Vergil – several different printings of his complete *opera*, individual editions of each work, commentaries on specific books, and so on. While multiple editions of a single work can provide a sense of the influence that work has had throughout history, for many linguistic tasks we will want to only count each distinct usage once, rather than once for each edition. One possibility for including this kind of information is to encode each work with its place in a FRBR hierarchy [20; 5], an ontology that specifies the relationship between a given abstract work (such as Vergil's *Aeneid*) and its expression and manifestation in different editions. Including this kind of information will give us more flexibility in creating ad hoc corpora, where we can either choose to include or exclude multiple versions of the same work.

4.3 Translation identification

In addition to multiple editions of the same work, massive textual collections also include multiple translations for many core texts. The Internet Archive contains editions of Horace's *Odes* in at least 8 different languages – not only in the Latin original, but also in English, Spanish, Italian, French, Early Modern French, Portuguese and German. Identifying these translations and then subsequently encoding their own place in a FRBR hierarchy (as expressions of the same abstract work) will allow us to perform a range of parallel text analysis [4; 19] on the ensuing corpus. One such possibility includes using parallel translations as a pivot point around which to chart how lexical information in Latin changes over time (and place). We have already undertaken some of this work in our efforts to create dynamic lexica for Greek and Latin [2; 3]. Figure 9 includes one such example, an entry for the Greek word δῆμος from the Ancient Greek Dynamic Lexicon.

The sense inventory here (that δῆμος can mean "people," "democracy," etc.) has been automatically induced from a collection of classical Greek source texts and parallel English translations. We can see how this word has both major and minor nuances of meaning within these different authors by analyzing how it has been translated in each them – for Homer, the sense is predominantly a fundamental one of "land," while for Demosthenes it also has a dominant sense of "assembly" in addition to its universal meaning of "people." The work that we have done so far has focused on how individual words can have subtly different meanings in different authors from different eras, but the same techniques can be used to see how a given word changes in meaning within common use over time. By identifying parallel texts in English, we will be able to extend this lexicographical work to create a much more comprehensive dictionary of Latin that can chart the lexical shift in words from Classical Latin to the 19th century.



The Ancient Greek **Dynamic Lexicon**

Enter Greek text in Unicode (νόμος) or Beta Code (no/mos), with or without accents.

δῆμος [High Frequency]

noun (masc)

- **people (51%) [Explain]**
- **democracy (6%)**
- **assembly (5%) (*Demosthenes*)**
- **common (3%) (*Thucydides*)**
- **land (2%) (*Homer*)**

Attributes:

- **πίων ("rich land"). (*Homer*)**
- **Ἰθάκη ("land of ithaca"). (*Homer*)**

Object of:

- **καταλύω ("overthrow the democracy"). (*Demosthenes*)**

Subject of:

- **χειροτονέω ("elected by the people").**

Example sentences.

- **ὁ γὰρ δῆμος ἐχειροτόνησεν· ("the people voted it to me;"). App. BC 1.8.**
- **ἐν δὲ τοῖς δήμοις καταριθμεῖται ἡ πόλις. ("eleusis is numbered among the demes."). 9.1.**
- **Εὐθύδικος προῃρεῖτο τὰς ὑπὲρ τοῦ δήμου πράξεις· ("euthydicus elected to work for the people."). Din. 1 33.**

Fig. 9. Dynamic lexicon entry for δῆμος.

4.4 Morphological/syntactic features

The syntactic information shown in the middle of figure 9 above is made possible by extracting information from a morphologically tagged and syntactically parsed corpus. By training a statistical dependency parser [15; 17; 25] on a treebank – a large collection of sentences that have been manually parsed by hand – we can assign an automatic syntactic parse to a much larger body of work than could be parsed manually; while this automatically parsed corpus contains errors, the volume of data is such that the patterns are strong enough to emerge through the noise. Here we have induced that πίων is a common attribute of δῆμος (and from our parallel texts we can figure out that together they mean "rich land") and that δῆμος is a common attribute of καταλύω (which together means to "overthrow the democracy").

Creating a lexical entry that is supplemented with this kind of morphosyntactic behavior is one downstream application of having a corpus that has been syntactically analyzed (either manually or automatically). A direct benefit of this level of annotation is the ability to research linguistic phenomena that reach beyond simple word forms, such as measuring the changing distribution of Latin word order over those two thousand years, or the differing ratios of poetical hyberbaton in Classical Latin poets compared to their Neo-Latin imitators.

One complicating factor that prevents us from easily parsing the entire corpus is the range of different authors, genres and eras included in it – we would expect strong out-of-genre effects on parsing accuracy if we attempted to parse a 12th-century scholastic author using training data from Classical Latin. Several projects, however, are undertaking the creation of syntactically parsed data for different stages of Latin, including not only our own Latin Dependency Treebank for texts of the Classical period [1], but also the PROIEL corpus of the New Testament (which includes the Latin Vulgate) [10], and the Index Thomisticus treebank [22] on the works of Thomas Aquinas. Our hope is to be able to leverage these different treebanks as a diverse set of training materials for parsing the range of genres included in the Internet Archive's collection. By adding morphosyntactic information to this corpus, we open it up to a much broader range of inquiry.

5. CONCLUSION

The nascent collections that are rising out of large digitization projects promise to hold the largest historical corpora in existence. In beginning to work with these collections, our first task must be to characterize them – to locate their strengths and weaknesses and identify the crucial points where additional effort can have the biggest impact. In our work to date, two of these critical points have been accurate language identification and supplementing the metadata with dates of composition, but these are only the first steps. While these variables allow us to dynamically assemble a historical Latin corpus and start investigating the impact of several linguistic and historical variables on the mechanics of its usage across two thousand years, there still remains much more to be done to transform this massive but errorful set of texts into a structured and dynamic corpus useful for linguistic processing.

6. ACKNOWLEDGMENTS

This work was supposed by grants from the National Science Foundation (IIS-910884, "Mining a Million Scanned Books: Linguistic and Structure Analysis, Fast Expanded Search, and Improved OCR"), the National Endowment for the Humanities (PR-50013-08, "The Dynamic Lexicon: Cyberinfrastructure and the Automated Analysis of Historical Languages") and the Digging into Data Challenge sponsored by the National Endowment for the Humanities, the National Science Foundation, the Social Sciences and Humanities Research Council and the Joint Information Systems Committee ("Towards Dynamic Variorum Editions").

REFERENCES

David Bamman and Gregory Crane. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78, Prague, 2006. ÚFAL MFF UK.

David Bamman and Gregory Crane. Building a dynamic lexicon from a digital library. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 11–20, New York, NY, USA, 2008. ACM.

David Bamman and Gregory Crane. Computational linguistics and classical lexicography. *Digital Humanities Quarterly*, 3(1), 2009.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer.

The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311, 1993.

George Buchanan. Frbr: enriching and integrating digital libraries. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 260–269, New York, NY, USA, 2006. ACM.

Gregory Crane. From the old to the new: Integrating hypertext into traditional scholarship. In *Hypertext '87: Proceedings of the 1st ACM conference on Hypertext*, pages 51–56. ACM Press, 1987.

Gregory Crane. New technologies for reading: The lexicon and the digital library. *Classical World*, pages 471–501, 1998.

Gregory Crane, David Bamman, Lisa Cerrato, Alison Jones, David M. Mimno, Adrian Packel, David Sculley, and Gabriel Weaver. Beyond digital incunabula: Modeling the next generation of digital libraries. In Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, and Rafael C. Carrasco, editors, *ECDL*, volume 4172 of *Lecture Notes in Computer Science*, pages 353–366. Springer, 2006.

William H. Fletcher. Facilitating compilation and dissemination of ad-hoc web corpora. In *Corpora and Language Learners*, page 271, 2004.

D.T.T. Haug and M.L. Jøhndal. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakesh, 2008.

Jean-Rémy Hochmann, Ansgar D. Endress, and Jacques Mehler. Word frequency as a cue for identifying function words in infancy. *Cognition*, 115(3):444–457, 2010.

Paul Oskar Kristeller. *Renaissance Thought and Its Sources.* Columbia University Press, New York, 1979.

H. Kucera and W. N. Francis. *Computational analysis of present-day American English.* Brown University Press, Providence, RI, 1967.

Kalev Leetaru. Mass book digitization: The deeper story of Google Books and the Open Content Alliance. *First Monday*, 13(10), 2008.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, 2005.

David Mimno and Andrew McCallum. Organizing the oca: learning faceted subjects from a library of digital books. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 376–385, New York, NY, USA, 2007. ACM.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135, 2007.

Geoffrey Nunberg. Google's book search: A disaster for scholars. *The Chronicle of Higher Education*, August 31, 2009.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records: final report. 2009.

Jon Orwant. Our commitment to the digital humanities. `http://googleblog.blogspot.com/2010/07/our-commitment-to-digital-humanities.html`, July 14, 2010.

Marco Passarotti. Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In Petrilli Raffaella and Femia Diego, editors, *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, Settembre 2006*, pages 187–205. Roma, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio, 2007.

Johann Ramminger. Neulateinische Wortliste. Ein Wörterbuch des Lateinischen von Petrarca bis 1700. `http://www.neulatein.de`, 2003ff.

Stacy Silverman and Nan Bernstein Ratner. Measuring lexical diversity in children who stutter: application of vocd. *J Fluency Disord*, 27(4):289–303, 2002.

David A. Smith and Jason Eisner. Dependency parsing by belief propagation. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 145–156, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

William John Teahan. Text classification and segmentation using minimum cross-entropy. In Joseph-Jean Mariani and Donna Harman, editors, *RIAO*, pages 943–961. CID, 2000.

Thesaurus Linguae Latinae, fourth electronic edition. K. G. Saur. `http://www.thesaurus.badw.de/`, 2006.