NATIONAL ENDOWMENT FOR THE HUMANITIES

SAMPLE APPLICATION NARRATIVE

Preservation and Access Grants
Institution: Tufts University

The Dynamic Lexicon: Cyberinfrastructure and the Automatic Analysis of Historical Languages

Abstract

With the rise of Google Books, the Open Content Alliance and other projects aimed at mass digitization, we are beginning to see the birth of an emerging cyberinfrastructure that has the potential to transform the way scholarly research is conducted. These large "million book" projects are more than simple stocks of incunabular page images – they are large collections of texts that have the potential to communicate with one another, but only through the technologies that one applies to them.

For the past twenty years, the Perseus Project has been concerned with the varieties of knowledge that emerge when texts are given this opportunity to speak. Our digital library of Greek and Latin texts has been a testbed for research in morphology, syntax and named entity analysis. As new texts are added to this library, they are subjected to a variety of automatic processes – a morphological analyzer inspects each source word and presents a list of possible parses, while a tagger selects the most probable one based on the other texts in our collection; a named entity analyzer that has been trained on these texts does the same for all proper names found therein. All Greek or Latin source words are linked to their respective dictionary entries, and all canonical citations are linked to their source text. Every time a new text is added, it is analyzed by systems that have been trained on the texts that are already there, and it becomes part of the cyberinfrastructure itself.

The million book projects that are now emerging have the potential to significantly transform these processes by their sheer volume alone. We have been able to make great progress with a Classical collection of nine million words – over 400,000 distinct users rely on our services each month – but we stand to go much further with a collection several hundred times that size.

On a large scale, we propose to research core functions for the automatic analysis of historical languages within this emerging cyberinfrastructure; specifically, we will research three technologies for building a dynamic lexicon, as well as the processes required to automatically create such a reference work for any collection of texts. Our efforts will focus on parallel text analysis – word sense induction and disambiguation – as well as automatic syntactic parsing. These technologies will enable us to create a reference work that enumerates the possible senses for a headword while also providing detailed syntactic information and statistical data about its use in a corpus. The methods we use to create such a reference work will also let us create an intelligent search index into our nine-million-word corpus – enabling users to search a text not only by word form, but also by word sense, syntactic subcategorization and selectional preference.

While other projects have developed reference works based on large collections of texts, our contribution to this line of research is the role that automated methods can play within an emerging cyberinfrastructure. While a tangible product of our research will be a sophisticated lexicon, our main contribution will be the steps that any digital library can take to create a reference work of their own and interface it with the texts in their collection. It is the nature of cyberinfrastructure to be in a constant state of change; the processes we develop will allow large collections not only to create reference works but also to let them adapt and grow.

The Dynamic Lexicon: Cyberinfrastructure and the Automatic Analysis of Historical Languages

Table of Contents

The Dynamic Lexicon: Cyberinfrastructure and the Automatic Analysis of Historical Languages

I. Significance

With the rise of Google Books, the Open Content Alliance, and other projects aimed at mass digitization, we now have the opportunity to exploit a new generation of digital texts. We propose to research core functions for the automatic analysis of historical languages within this emerging cyberinfrastructure, applying methods available in computational linguistics to the growing body of materials relevant to Greek and Latin. Scholars and students alike need lexicographic tools that combine traditional and emerging services. They need machine actionable lexica that combine scarce human labor with automated methods to provide richer services than have ever been possible before and to make historical languages intellectually more accessible, both for professional scholarship and for more general intellectual inquiry. Audiences include experts looking for new knowledge, novices with no linguistic background using cross language information retrieval and machine translation, and students of the language augmenting their existing knowledge with queries about the meanings and functions of words in context.

We focus on two processes: identifying significant patterns (e.g., word X has meanings Y and Z; verb A takes the dative or accusative), then finding additional instances of these patterns in the text (e.g., search for places in Herodotus where the Greek word *archê* more closely resembles English "beginning" than "empire"; locate instances of the Latin word *libero* where it takes construction A vs. B). Wholly manual lexicography has produced extraordinary resources for Greek and Latin (such as the massively informative *Thesaurus Linguae Latinae*[1]) but manual methods cannot in the immediate future provide for all texts the same level of coverage available for the most heavily studied materials. Moreover, purely manual reference materials cannot grow more effective automatically as new data becomes available nor can they be customized for particular users looking at particular texts, authors or genres. For instance, while the *Oxford Latin Dictionary* focused on a canon of classical authors that ends around the second century CE, Latin continued to be a productive language for the ensuing two millennia, with prolific writers in the Middle Ages, Renaissance and beyond. The Index Thomisticus (Busa 1974-1980) alone contains 10.6 million words attributed to Thomas Aquinas and related authors, which is by itself larger than the entire corpus of extant classical Latin.[2] Many handcrafted lexica exist for this period, from the scale of individual authors (cf. Ludwig Schütz' 1895 *Thomas-Lexikon*) to entire periods (e.g., J. F. Niermeyer's 1976 *Mediae Latinitatis Lexikon Minus*), but we can still do more: we can create a dynamic lexicon that can change and grow when fed with new texts, and that can present much more information about a word than reference works bound by the conventions of the printed page.

Manual labor remains, however, crucial and, while lexicographic practice may evolve, the value of lexicographic work can only increase in a digital environment – one only needs to witness the impact of the COBUILD project (Sinclair 1987) on corpus lexicography to see the added value that digital methods can provide. If we can increase the impact of lexicographic work, it

[1] http://www.thesaurus.badw.de/

[2] The Biblioteca Teubneriana BTL-1 collection, for instance, contains 6.6 million words, covering Latin literature up to the second century CE. For a recent overview of the Index Thomisticus, including the corpus size and composition, see Busa (2004).

becomes possible to bring more professional labor to bear: if we can link smaller lexicographic contributions (e.g., studies of particular words in genres, authors, or passages) to relevant passages and thus render them more visible, we increase their potential impact and thus raise the potential rewards for such contribution. At the same time, if we can create lexicographic resources that are more powerful than those modeled on static print, we can make a more compelling case to attract resources for purposeful, long-term lexicographic projects in every country where classical and, indeed, historical languages are studied.

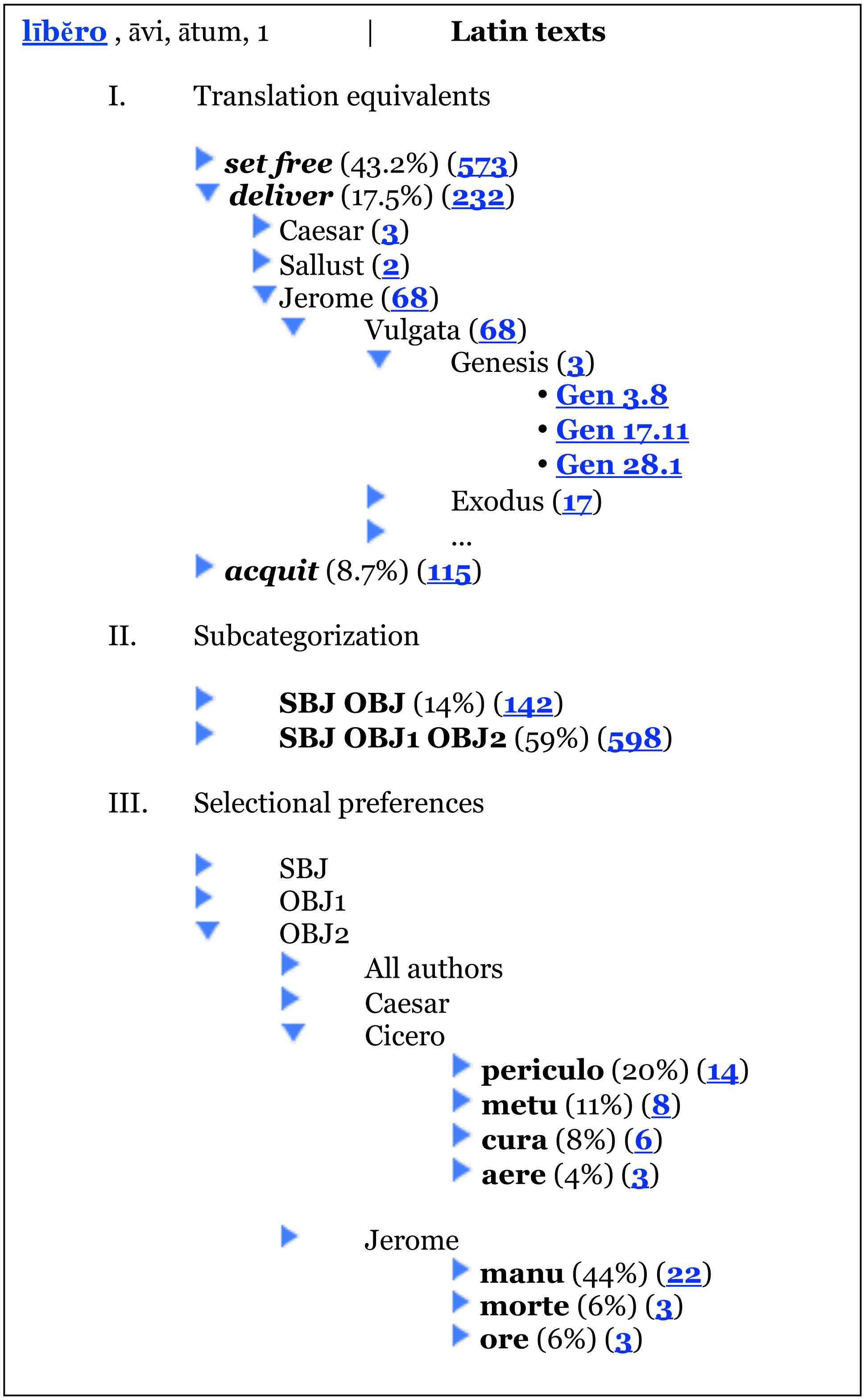
We therefore concentrate on two problems. First, how much can we automatically learn from a large textual collection using machine learning techniques that thrive on large corpora? And second, how can the vast labor already invested in handcrafted lexica help those techniques to learn?

We propose that what we can learn from such a corpus is actually quite significant. With a large bilingual corpus, we can induce a word sense inventory to establish a baseline for how frequently certain definitions of a word are manifested in actual use; we can also use the context surrounding each word to establish which particular definition is meant in any given instance. With the help of a treebank (a handcrafted collection of syntactically parsed sentences), we can train an automatic parser to parse the sentences in a monolingual corpus and extract information about a word's subcategorization frames (the common syntactic arguments it appears with – for instance, that the verb *dono* requires a subject, direct object and indirect object), and selectional preferences (e.g., that the subject of the verb *amo* is typically animate). With clustering techniques, we can establish the semantic similarity between two words based on their appearance in similar contexts.

We propose to leverage all of these techniques to create dynamic lexica for Latin and Greek. In each of these reference works, the headwords will include the following:

1. a list of possible senses, weighted according to their probability
2. a list of instances of each sense in the source texts
3. a list of common subcategorization frames, weighted according to their probability
4. a list of selectional preferences, weighted according to their probability

In creating a lexicon with these features, we are exploring two strengths of automated methods: they can analyze not only very large bodies of data but also provide customized analysis for particular texts or collections. We can thus not only identify patterns in one hundred and fifty million words of later Latin but compare which senses of which words appear in the one hundred and fifty thousand words of Thucydides. Figure 1 presents a mock-up of what a dictionary entry could look like in such a dynamic reference work. The first section ("Translation equivalents") presents items 1 and 2 from the list, and is reminiscent of traditional lexica for classical languages: a list of possible definitions is provided along with examples of use. The main difference between a dynamic lexicon and those print lexica, however, lies in the scope of the examples: while print lexica select one or several highly illustrative examples of usage from a source text, we are in a position to present far more.



līběro , āvi, ātum, 1 | **Latin texts**

I. Translation equivalents

- ▶ ***set free*** (43.2%) (573)
- ▼ ***deliver*** (17.5%) (232)
  - ▶ Caesar (3)
  - ▶ Sallust (2)
  - ▼ Jerome (68)
    - ▼ Vulgata (68)
      - ▼ Genesis (3)
        - **Gen 3.8**
        - **Gen 17.11**
        - **Gen 28.1**
      - ▶ Exodus (17)
      - ▶ …
- ▶ ***acquit*** (8.7%) (115)

II. Subcategorization

- ▶ **SBJ OBJ** (14%) (142)
- ▶ **SBJ OBJ1 OBJ2** (59%) (598)

III. Selectional preferences

- ▶ SBJ
- ▶ OBJ1
- ▼ OBJ2
  - ▶ All authors
  - ▶ Caesar
  - ▼ Cicero
    - ▶ **periculo** (20%) (14)
    - ▶ **metu** (11%) (8)
    - ▶ **cura** (8%) (6)
    - ▶ **aere** (4%) (3)
  - ▶ Jerome
    - ▶ **manu** (44%) (22)
    - ▶ **morte** (6%) (3)
    - ▶ **ore** (6%) (3)

Figure 1: Mock-up of sample dynamic lexicon entry

The resources we will use to build this work break down into two groups, and are resources that any large historical collection would have: first, the existing structured corpora that are the product of years of labor by the field; and second, those taking shape as part of "million book" collections.

On the one hand, substantial labor has gone into the creation of carefully curated resources, often with SGML/XML markup in the Text Encoding Initiative. In classics, these resources include not only source texts but also lexica and grammars, often with very extensive markup capturing hierarchical models of word senses, encyclopedias about people, places, technical terms and other specialized topics, commentaries on particular texts, catalogues of buildings and objects from museums and archaeological sites and other more or less structurally homogeneous resources. We have at our disposal a representative library of such materials for Greek and Latin, developed over twenty years and covering every major genre of reference work.

At the same time, we also have very large collections of books scanned as raw page images and with text automatically extracted by OCR software. While recognizing page layouts of complex reference works (e.g., the separate entries and the hierarchically encoded sense definitions) remains challenging, we have already been able to extract the basic text of primary sources from critical editions with a high degree of accuracy.[3] Google has as of late spring 2007 made it possible within the US to search c. 2,000,000 digitized books. Within this corpus, classics is surprisingly well represented: a substantial portion of public domain books within university libraries cover classical subjects and include many classic editions, commentaries, and reference works. At the same time, the Open Content Alliance has established mechanisms whereby third parties can have public domain books from research libraries scanned and made publicly available. The University of Toronto has begun scanning for us an initial set of Greek and Latin editions. We can use these editions as a testbed for the work proposed here, and they will be available without restriction from the Internet Archive Open Content Alliance server.

While we will apply methods from the field of text mining, our focus will be on the implications of these methods for the data structures already available for humanists and for the reference works designed from the start to be both comprehensible to human readers and to provide data on which automated methods can depend. Other groups in the past have leveraged a large digital corpus to facilitate the construction of new reference works and also to provide structured access to the corpus itself, from the COBUILD project to more modern endeavors such as Kilgarriff's Sketch Engine (Kilgarriff et al. 2004) or the German *elexiko* project (Klosa et al. 2006). These groups largely work with high-resource modern languages (the *elexiko* project, for instance, is built on a corpus of 1.3 billion words) with powerful existing tools for automatic analysis. Our contribution to this line of research is the role that such automatic methods can play within an emerging cyberinfrastructure. We distinguish cyberinfrastructure from vast corpora not only in the structure imposed upon the texts that comprise it, but also in the very composition of those texts: while modern reference corpora are typically of little interest in themselves (e.g., newspaper articles), our texts have been the focus of scholars' attention for millennia. The meaning of the word *child* in a single sentence from the *Wall Street Journal* is hardly a research question worth asking, except for the newspaper's significance in being representative of the language at large; but this same question when asked of Vergil's fourth *Eclogue* has been at the center of scholarly debate since the time of the emperor Constantine (see Bourne 1916 for an

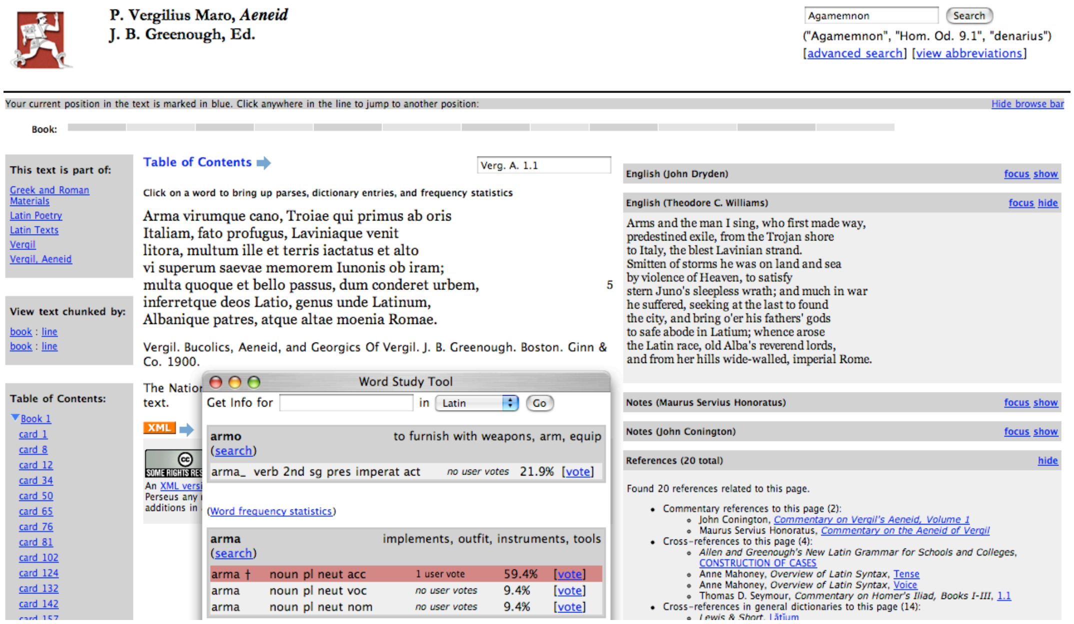
[3] In a study of OCR for classical Greek, we found that we could achieve accuracy levels of transcribed characters that approached the levels demanded in professional data entry (99.94% for automated methods vs. 99.95% for professional data entry) (Stewart et al. 2007).

overview of *puer* in *Ec*. IV). We need to provide traditional scholars with the apparatus necessary to facilitate their own textual research. This will be true of a cyberinfrastructure for any historical culture, and for any future structure that develops for modern scholarly corpora as well.

II. Background of applicant

The Perseus Digital Library Project (Crane 1987, Crane et al. 2006) has been under continuous development since 1987 and actively serves structured texts, advanced morphological services and image data produced in that first year of work. CD-ROM publications in the early 1990s gave way to a website which has been in continuous and expanding operation since 1995 and currently serves up to 15,000,000 pages per month.

Perseus has accumulated over this period the collections and services that, in part, make possible the work proposed here. These resources include well-structured XML collections for classical studies, including source texts and translations. These collections are large enough to model key services for the much larger and less structured collections that are taking shape. Perseus also has space and long term support for core staff, thus providing us with continuity for the future.



P. Vergilius Maro, *Aeneid*
J. B. Greenough, Ed.

Agamemnon — Search
("Agamemnon", "Hom. Od. 9.1", "denarius")
[advanced search] [view abbreviations]

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position:

This text is part of:
Greek and Roman Materials
Latin Poetry
Latin Texts
Vergil
Vergil, Aeneid

View text chunked by:
book : line
book : line

Table of Contents:
Book 1
card 1
card 8
card 12
card 34
card 50
card 65
card 76
card 81
card 102
card 124
card 132
card 142
card 157

Table of Contents — Verg. A. 1.1

Click on a word to bring up parses, dictionary entries, and frequency statistics

Arma virumque cano, Troiae qui primus ab oris
Italiam, fato profugus, Laviniaque venit
litora, multum ille et terris iactatus et alto
vi superum saevae memorem Iunonis ob iram;
multa quoque et bello passus, dum conderet urbem,
inferretque deos Latio, genus unde Latinum,
Albanique patres, atque altae moenia Romae.

Vergil. Bucolics, Aeneid, and Georgics Of Vergil. J. B. Greenough. Boston. Ginn & Co. 1900.

The Natio
text.

An XML versi
Perseus any
additions in

Word Study Tool

Get Info for — in Latin — Go

armo (search) — to furnish with weapons, arm, equip

arma_	verb 2nd sg pres imperat act	*no user votes*	21.9%	[vote]

(Word frequency statistics)

arma (search) — implements, outfit, instruments, tools

arma †	noun pl neut acc	1 user vote	59.4%	[vote]
arma	noun pl neut voc	*no user votes*	9.4%	[vote]
arma	noun pl neut nom	*no user votes*	9.4%	[vote]

English (John Dryden) — focus show

English (Theodore C. Williams) — focus hide

Arms and the man I sing, who first made way,
predestined exile, from the Trojan shore
to Italy, the blest Lavinian strand.
Smitten of storms he was on land and sea
by violence of Heaven, to satisfy
stern Juno's sleepless wrath; and much in war
he suffered, seeking at the last to found
the city, and bring o'er his fathers' gods
to safe abode in Latium; whence arose
the Latin race, old Alba's reverend lords,
and from her hills wide-walled, imperial Rome.

Notes (Maurus Servius Honoratus) — focus show

Notes (John Conington) — focus show

References (20 total) — hide

Found 20 references related to this page.

- Commentary references to this page (2):
  - John Conington, *Commentary on Vergil's Aeneid, Volume 1*
  - Maurus Servius Honoratus, *Commentary on the Aeneid of Vergil*
- Cross-references to this page (4):
  - *Allen and Greenough's New Latin Grammar for Schools and Colleges*, CONSTRUCTION OF CASES
  - Anne Mahoney, *Overview of Latin Syntax*, Tense
  - Anne Mahoney, *Overview of Latin Syntax*, Voice
  - Thomas D. Seymour, *Commentary on Homer's Iliad, Books I-III*, 1.1
- Cross-references in general dictionaries to this page (14):
  - *Lewis & Short*, Lătium

Figure 2: The Perseus Digital Library

Currently, our fourth digital library system (Figure 2) is in operation. Written in Java as a modular, extensible framework, this system provides a stable software base with which to bring services before a broader audience. This digital library system, the fourth and last that we will develop, leads to another foundation for the work proposed here.

The Digital Collections and Archives group at Tufts (DCA) has established an institutional repository and digital library system, based on the Cornell/Virginia Fedora system. As part of the university library system, the DCA provides the long term home for all digital objects within the Perseus collections. Equally important, the Fedora repository system allows us to contribute not only data but services as well: users of Perseus do not work with inert texts but expect dynamically generated links from inflected Greek and Latin words to lexicon entries, automatic citation collection, and other services. The services underlying the new Perseus Digital Library system were designed to become disseminators within the Tufts Fedora system. The core services on which Perseus users depend will thus also become part of the Tufts DCA system. The DCA provides us with a mechanism to preserve objects and services alike.

Finally, Tufts Academic Technology maintains a cluster of, as of 2007, 32 Linux machines. This provides us with enough computing power to pursue more complex tasks than were feasible with dual or quadruple processor machines. Our named entity analysis system is, for example, demanding but we can process 1 billion words a day by using the research cluster.

III. History, scope and duration of the project

For almost twenty years Perseus has focused on creating well-structured content and associated services. We have digitized scores of texts from the classical canon, and now offer 4.9 million words of Greek source texts along with 3.4 million words of Latin, all marked up in TEI-compliant XML and many paired with a coarsely aligned English translation. We have successfully built a number of useful services on top of these collections, including a morphological analyzer, morphological recommender system (selecting the correct morphological form in context), and a named entity analyzer. All of these services have been incorporated into our online digital library and are used by thousands of distinct users every day.

Our project also has a history in digital lexicography. In 1994, the NEH funded a proposal to digitize the unabridged Liddell, Scott and Jones *Greek-English Lexicon* (LSJ9). Including this reference work into our digital library had a dramatic impact on our audience – while the unabridged Greek lexicon is so bulky and hard to read that most students traditionally work with the intermediate version, we found that by 1998 the unabridged version was being used five times as often as the shorter edition. Additionally, since the digitized LSJ contains 500,000 canonical citations to passages in Greek literature, we were able not only to link each citation to its actual source text within our digital library, but also to use the reference work as a commentary by a process of reverse citation: if a given dictionary entry (such as *mênis*) cites, for instance, *Iliad* 1.1, we can create a link in the digital library presentation of the passage to that specific entry in the lexicon. This research has subsequently gone far beyond the scope of our original funding since it has allowed us to transform *any* reference work with canonical citations into a virtual commentary. Our work in collocation analysis also suggests the many ways it can be useful to lexicographers, from simply confirming what we already know to suggesting important subsenses not found in existing dictionaries (Rydberg-Cox 2002).

The lexica, source texts and commentaries in the Greco-Roman collections of Perseus reflect generous support from the NEH in previous grants. For a full list of the project's publications, please see our website.[4] The proposal itself builds on NEH's contribution to the DLI-2 as well as support from the NSF and other agencies.

[4] http://www.perseus.tufts.edu/hopper/publications

IV. Methodology and standards

Our approach in this project is to transfer already established NLP methods to the classical language community. As such, our methodology involves the application of three core technologies to cyberinfrastructure. The result will be two deliverables: a dynamic lexicon that can grow when fed with new texts, and an intelligent search index into the texts contained within it.

A. Technologies

Our first process, that of automatically identifying significant patterns in a text, is based on three core technologies:

1. identifying word senses from parallel texts;
2. locating the correct sense for a word using contextual information; and
3. parsing a text to extract important syntactic information.

Each of these technologies has a long history of development both within our group and in the natural language processing community at large. In the following section we will delineate how we will leverage each of them in this project to uncover large-scale usage patterns in a text.

A.1 Word Sense Induction

We have already begun work on building a Latin sense inventory from a small collection of parallel texts in our digital library. Our work is based on that of Brown et al. (1991) and Gale et al. (1992), who suggest that one way of objectively detecting the real senses of any given word is to analyze its translations: if a word is translated as two semantically distinct terms in another language, we have *prima facie* evidence that there is a real sense distinction. So, for example, the Greek word *archê* may be translated in one context as *beginning* and in another as *empire*, corresponding respectively to Liddell and Scott definitions I.1 and II.2.

Finding all of the translation equivalents for any given word then becomes a task of aligning the source text with its translations, at the level of individual words. The Perseus Digital Library contains at least one English translation for most of its Latin and Greek prose and poetry source texts. Many of these translations are encoded under the same canonical citation scheme as their source, but must further be aligned at the sentence and word level before individual word translation probabilities can be calculated. The workflow for this process is shown in Figure 3.

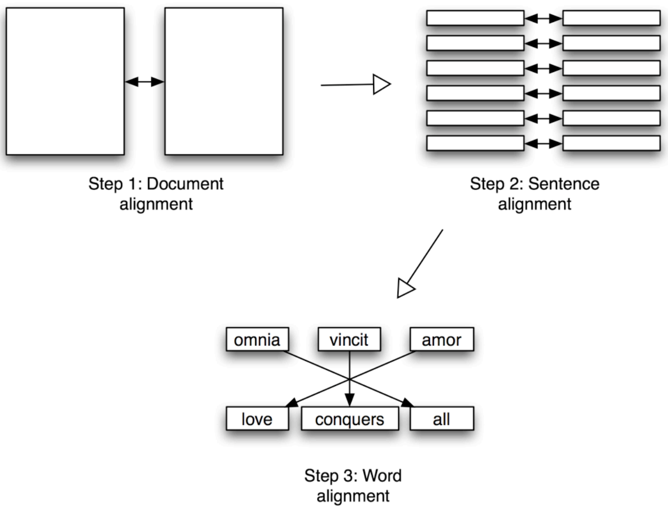


Figure 3: Alignment workflow

Since the XML files of both the source text and its translations are marked up with the same reference points, "chapter 1, section 1" of Tacitus' *Annales* is automatically aligned with its English translation (step 1). This results (for Latin at least) in aligned chunks of text that are 217 words long. These chunks are then aligned on a sentence level in step 2 using Moore's Bilingual Sentence Aligner (Moore 2002), which aligns sentences that are 1-1 translations of each other with a very high precision (98.5% for a corpus of 10,000 English-Hindi sentence pairs (Singh and Husain 2005)).

In step 3, we then align these 1-1 sentences using GIZA++ (Och and Ney 2003). Prior to alignment, all of the tokens in the source text and translation are lemmatized, where each word is replaced with all of the lemmas from which it can be inflected (for example, the Latin word *est* is replaced with *sum1 edo1* and the English word *is* is replaced with *be*). This word alignment is performed in both directions in order to discover multi-word expressions (MWE's) in the source language.

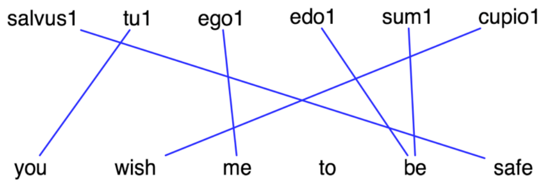
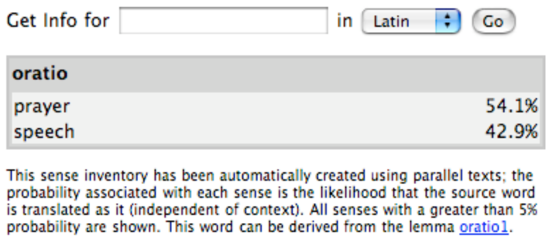


Figure 4: Sample word alignment from GIZA++.

Figure 4 shows the result of this word alignment (here with English as the source language). The original, pre-lemmatized Latin is *salvum tu me esse cupisti* (Cicero, *Pro Plancio*, chapter 33). The original English is *you wished me to be safe*. As a result of the lemmatization process, many source words are mapped to multiple words in the target – most often to lemmas which share a common inflection. For instance, during lemmatization, the Latin word *esse* is replaced with the two lemmas from which it can be derived – *sum1 (to be*) and *edo1* (*to eat*). If the word alignment process maps the source word *be* to both of these lemmas in a given sentence (as in Figure 4), the translation probability is divided evenly between them.

From these alignments we can calculate overall translation probabilities, which we currently present as an ordered list, as in Figure 5.



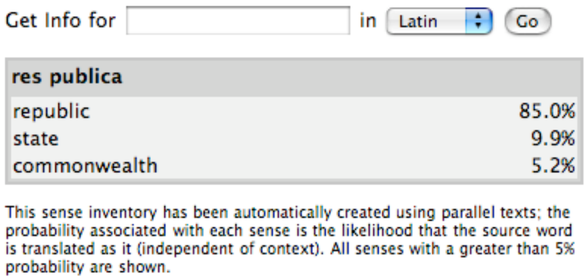
oratio	
prayer	54.1%
speech	42.9%

This sense inventory has been automatically created using parallel texts; the probability associated with each sense is the likelihood that the source word is translated as it (independent of context). All senses with a greater than 5% probability are shown. This word can be derived from the lemma oratio1.

Figure 5: Sense inventory for *oratio* induced from parallel texts.

The weighted list of translation equivalents we identify using this technique will provide the foundation for our lexical work. In the example above, we have induced from our collection of parallel texts that the headword *oratio* is primarily used with two senses: *speech* and *prayer*.

The granularity of the definitions in such a dynamic lexicon cannot approach that of human labor: the Lewis and Short *Latin Dictionary*, for instance, enumerates fourteen subsenses in varying degrees of granularity, from “speech” to “formal language” to the “power of oratory” and beyond. Our approach, however, does have two clear advantages which complement those of traditional lexica: first, this method allows us to include statistics about actual word usage in the corpus we derive it from. The use of *oratio* to signify *prayer* is not common in classical Latin, but since the corpus we induced this inventory from is largely composed of the *Vulgate* of Jerome, we are also able to mine this use of the word and include it in this list as well. Since the lexicon is dynamic, we can generate a sense inventory for an entire corpus or any part of it – so that if we were interested, for instance, in the use of *oratio* only until the second century CE, we can exclude the texts of Jerome from our analysis. And since we can run our word alignment at any time, we are always in a position to update the lexicon with the addition of new texts.



res publica	
republic	85.0%
state	9.9%
commonwealth	5.2%

This sense inventory has been automatically created using parallel texts; the probability associated with each sense is the likelihood that the source word is translated as it (independent of context). All senses with a greater than 5% probability are shown.

Figure 6: Sense inventory for the multi-word expression *res publica* induced from parallel texts.

Second, our word alignment also maps multi-word expressions, so we can include significant collocations in our lexicon as well. This allows us to provide translation equivalents for idioms and common phrases such as *res publica* (republic) or *gratias ago* (to give thanks), which are often difficult to find in dictionaries, buried within the definition of one of the component headwords.

A.2 Word Sense Disambiguation

Approaches to word sense disambiguation generally come in three varieties:

- knowledge-based methods (Lesk 1986, Banerjee and Pedersen 2002), which rely on existing reference works with a clear structure such as dictionaries and Wordnets (Miller 1995);
- supervised corpus methods (Grozea 2004), which train a classifier on a human-annotated sense corpus such as Semcor (Miller et al. 1993) or any of the SENSEVAL competition corpora (Mihalcea and Edmonds 2004); and
- unsupervised corpus methods, which train classifiers on “raw,” unannotated text, either a monolingual corpus (McCarthy et al. 2004) or parallel texts (Brown et al. 1991, Tufis et al. 2004).

Corpus methods (especially supervised methods) generally perform best in the SENSEVAL competitions – at SENSEVAL-3, the best system achieved an accuracy of 72.9% in the English lexical sample task and 65.1% in the English all-words task.[5] Manually annotated corpora, however, are generally cost-prohibitive to create, and this is especially exacerbated with sense-tagged corpora, for which the human inter-annotator agreement is often low.

Since the Perseus Digital Library contains two large monolingual corpora (the canon of Greek and Latin classical texts) and sizable parallel corpora as well, we have investigated two different techniques using parallel texts for word sense disambiguation: one using stochastically word-aligned texts with naive Bayesian classification and one using coarsely aligned parallel texts with Kullback-Leibler distance for term weighting.

The first method uses the same techniques we used to create a sense inventory to disambiguate words in context. After we have a list of possible translation equivalents for a word, we can use the surrounding Latin or Greek context as an indicator for which sense is meant in texts where we have no corresponding translation. There are several techniques available for deciding which sense is most appropriate given the context, and several different measures for what definition of “context” is most appropriate itself. One technique that we have experimented with is a naive Bayesian classifier (following Gale et al. 1992), with context defined as a sentence-level bag of words (all of the words in the sentence containing the word to be disambiguated contribute equally to its disambiguation).

Bayesian classification is most commonly found in spam filtering. A filtering program can decide whether or not any given email message is spam by looking at the words that comprise it and comparing it to other messages that are already known to be spam – some words generally only appear in spam messages (e.g., *viagra*, *refinance*, *opt-out*, *shocking*), while others only appear in non-spam messages (*archê*, *subcategorization*), and some appear equally in both (*and*, *your*). By counting each word and the class (spam/not spam) it appears in, we can assign it a probability that it falls into one class or the other.

We can also use this principle to disambiguate word senses by building a classifier for every sense and training it on sentences where we do know the correct sense for a word. Just as a spam

[5] At the time of writing, the SEMEVAL-1/SENSEVAL-4 (2007) competition is currently underway.

filter is trained by a user explicitly labeling a message as spam, this classifier can be trained simply by presence of an aligned translation.

For instance, the Latin word *spiritus* has several senses, including *spirit* and *wind*. In our texts, when *spiritus* is translated as *wind*, it is accompanied by words like *mons* (mountain), *ala* (wing) or *ventus* (wind). When it is translated as *spirit*, its context has (more naturally) a religious tone, including words such as *sanctus* (holy) and *omnipotens* (all-powerful). If we are confronted with an instance of *spiritus* in a sentence for which we have no translation, we can disambiguate it as either *spirit* or *wind* by looking at its context in the original Latin.

Latin context word	English translation	Probability of accompanying *spiritus = wind*
Mons	Mountain	98.3%
Commotio	Commotion	98.3%
Ventus	Wind	95.2%
Ala	Wing	95.2%

Table 1: Latin contextual probabilities where *spiritus = wind*.

Latin context word	English translation	Probability of accompanying *spiritus = spirit*
Sanctus	Holy	99.9%
Testis	Witness	99.9%
Vivifico	Make alive	99.9%
Omnipotens	All-powerful	99.9%

Table 2: Latin contextual probabilities where *spiritus = spirit*.

Word sense disambiguation will be most helpful for the construction of a lexicon when we are attempting to determine the sense for words in context for the large body of later Latin literature for which there exists no English translation. By training a classifier on texts for which we do have translations, we will be able to determine the sense in texts for which we don't: if the context of *spiritus* in a late Latin text includes words such as *mons* and *ala*, we can use the probabilities we induced from parallel texts to know with some degree of certainty that it refers to *wind* rather than *spirit*. This will enable us to include these later texts in our statistics on a word's usage, and link these passages to the definition as well.

The second method allows the texts to simply be coarsely aligned (i.e., at the level of a canonical citation, such as Thuc 1.38 rather than at the sentence or word level). In order to find the most probable sense in the corresponding translation, we use a term weighting scheme based on information theory. The Greek word *agathos*, for example, has several senses, including *good*, *brave*, and *aristocratic*. If we are attempting to determine which of these senses corresponds to a given instance of *agathos* in a section of text, finding the word *good* in multiple English translations of that section may be indicative of the first sense. However, *good* has a relatively high document frequency in most reference corpora, so its inverse document frequency score is low. Thus, the importance of finding *good* in several translations of the same section may be overpowered by a single rare term that does not actually hold disambiguating information. Applying the Kullback-Leibler distance allows us to see that although a term may occur with relatively high frequency in general, it occurs with even higher frequency in sections containing the reference term. Similarly, we can also weed out words that generally occur with low frequency, and whose frequency is not significantly higher in our selected sections.

This term weighting scheme in general gives the desired results. When an English word x is a stop word, such as *the*, or appears no more frequently in a given section than in the full corpus, it has no sense-discriminating value for Greek word y. However, when x occurs in our section with higher frequency than in the total corpus, then it may indeed have sense-disambiguating information for y. The higher this difference is, the more strongly the two words are connected. The term weighting scheme filters out noise and uncorrelated terms, and emphasizes those translated words that most likely correspond to a stable sense in the original language. Additionally, we can also use clustering techniques to collect these word senses into like groups.

A.3 Parsing

Two of the features we plan to incorporate into our dynamic lexicon are based on a word's role in syntax: subcategorization and selectional preference. A verb's subcategorization frame is the set of possible combinations of surface syntactic arguments it can appear with. In linear, unlabeled phrase structure grammars, these frames take the form of, for example, *NP PP* (requiring a direct object + prepositional phrase, as *I gave a book to John*) or *NP NP* (requiring two objects, as in *I gave John a book*). In a labeled dependency grammar, we can express a verb's subcategorization as a combination of syntactic roles (e.g., OBJ OBJ).

A predicate's selectional preference specifies the type of argument it generally appears with. The verb *to eat*, for example, typically requires its object to be a thing that can be eaten and its subject to have animacy, unless used metaphorically. Selectional preference, however, can also be much more detailed, reflecting not only a word class (such as *animate* or *human*), but also individual words themselves. For instance, the kind of arguments used with the Latin verb *libero* (to free) are very different in Cicero and Jerome: Cicero, as an orator of the republic, commonly uses it to speak of liberation from *periculum* (danger), *metus* (fear), *cura* (care) and *aere alieno* (debt); Jerome, on the other hand, uses it to speak of liberation from a very different set of things, such as *manus Aegyptorum* (the hand of the Egyptians), *os leonis* (the mouth of the lion), and *mors* (death). [6] These are syntactic qualities since each of these arguments bears a direct syntactic relation to their head as much as they hold a semantic place within the underlying argument structure.

In order to extract this kind of subcategorization and selectional information from unstructured text, we first need to impose syntactic order on it. One option for imposing this kind of order is through manual annotation, but this option is not feasible here due the sheer volume of data that must be annotated – even the more resourceful of such endeavors (such as the Penn Treebank (Marcus et al. 1994) or the Prague Dependency Treebank (Hajič 1994)) take years to complete.

A second, more practical option is to assign syntactic structure to a sentence using automatic methods. Great progress has been made in recent years in the area of syntactic parsing, both for phrase structure grammars (Charniak 2000, Collins 1999) and dependency grammars (Nivre et al. 2006, McDonald et al. 2005), with labeled dependency parsing achieving an accuracy rate approaching 90% for English (a high resource, fixed word order language) and 80% for Czech (a relatively free word order language like Latin and Greek). Automatic parsing generally requires

[6] See Bamman and Crane (2007) for a summary of this work.

the presence of a treebank – a large collection of manually annotated sentences – and a treebank's size directly correlates with parsing accuracy: the larger the treebank, the better the automatic analysis.

Under funding from the NSF, we are currently in the process of creating a treebank for Latin. Now in version 1.3, the Latin Dependency Treebank[7] is comprised of excerpts from four texts: Cicero's *Oratio in Catilinam*, Caesar's *Commentarii de Bello Gallico,* Vergil's *Aeneid* and Jerome's *Vulgate*. Each sentence in the treebank has been manually annotated so that every word is assigned a syntactic relation, along with the lemma from which it is inflected and its morphological code (a composite of nine different morphological features: part of speech, person, number, tense, mood, voice, gender, case and degree). Based predominantly on the guidelines used for the Prague Dependency Treebank, our annotation style is also influenced by the Latin grammar of Pinkster (1990), and is founded on the principles of dependency grammar (Mel'čuk 1988). Dependency grammars differ from phrase-structure grammars in that they forego non-terminal phrasal categories and link words themselves to their immediate heads. This is an especially appropriate manner of representation for languages with a free word order (such as Latin and Czech), where the linear order of constituents is broken up with elements of other constituents. A dependency grammar representation, for example, of *ista meam norit gloria canitiem* (Propertius I.8.46) – "that glory would know my old age" – would look like the following:

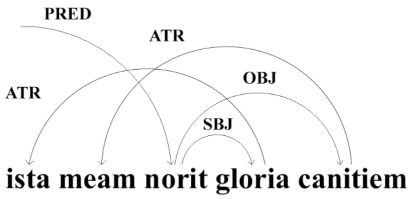


Figure 7: Dependency grammar representation of *ista meam norit gloria canitiem* ("that glory would know my old age")

While this treebank is still in its infancy, we can still use it to train a parser to parse the volumes of unstructured Latin in our collection. Our treebank is still too small to achieve state-of-the-art results in parsing but we can still use automatic methods to assign highly accurate but only partial parses to the sentences. Using Nivre et al.'s (2006) parser trained on our Latin data, we are able to automatically tag a smaller subset of the words in a Latin sentence (32.2%) with a higher accuracy rate (73.9%) than if we tried to parse them all. The syntactic dependencies that we uncover using these partial parses are in fact the ones that we are interested in – we can, for instance, achieve a precision of 87.0% for attributive modifiers (ATR) with a higher recall rate (52.2%) as well. As part of this project, we will work to further improve these automatic methods for the shallow parsing of Latin, and work as well to develop similar methods for Greek.

B. Cyberinfrastructure

The large "million book" projects that are now taking shape are more than simple stocks of incunabular page images – they are large collections of texts that have the potential to communicate with one another. For the past twenty years, the Perseus Project has been

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

concerned with the varieties of knowledge that emerge when texts are given this opportunity to speak. Our digital library of Greek and Latin texts has been a testbed for research in morphology, syntax and named entity analysis. As new texts are added to this library, they are subjected to a variety of automatic processes – a morphological analyzer inspects each source word and presents a list of possible parses, while a tagger selects the most probable one based on the other texts in our collection; a named entity analyzer that has been trained on these texts does the same for all proper names found therein. All Greek or Latin source words are linked to their respective dictionary entries, and all canonical citations are linked to their source text. Every time a new text is added, it is analyzed by systems that have been trained on the texts that are already there, and it becomes part of the cyberinfrastructure itself.

The million book projects that are now emerging have the potential to significantly transform these processes by their sheer volume alone. We have been able to make great progress with a classical collection of nine million words – over 400,000 distinct users rely on our services each month – but we stand to go much further with a collection several hundred times that size.

Each of the technologies described above will let us identify significant patterns in a text for individual words. One main advantage of a digital, dynamic lexicon, however, is that we can identify these patterns in subcorpora of varying size – comparing, for instance, how Thucydides' use of the word *archê* differs from that of Herodotus and from Greek literature at large.

The technologies used to uncover this information are only one part of the problem, however: we must also investigate ways to present this information to scholars to enable research of their own, and to automatically update it to reflect changes in the cyberinfrastructure of which it is a part. This involves two separate but related endeavors: creating a headword-based lexicon for browsing, and integrating that lexicon with our classical collection to create an intelligent search index.

B.1 Lexicon

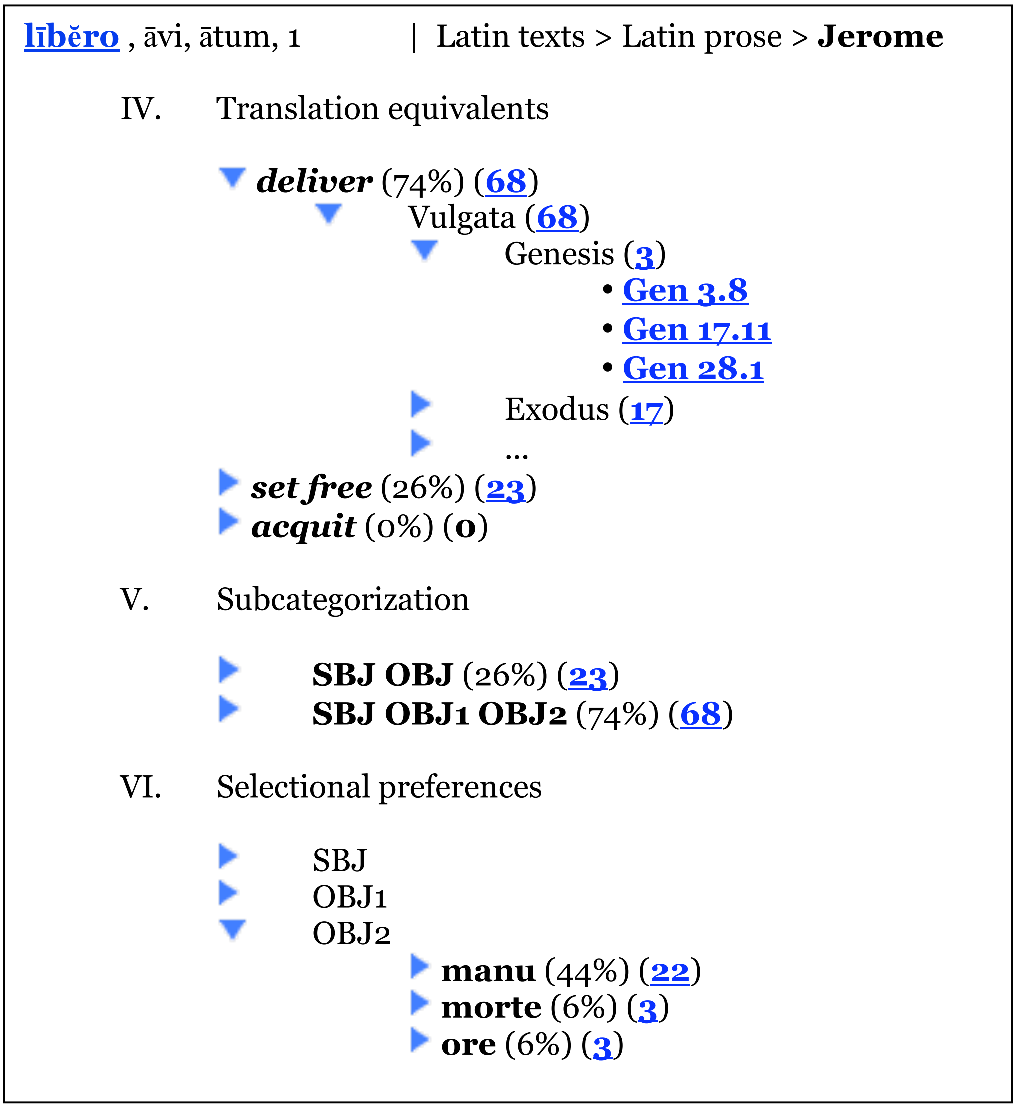
The Perseus Digital Library already contains several digitized lexica, including Lewis & Short's *Latin Dictionary* (Lewis and Short 1879) and the Liddell and Scott *Greek-English Lexicon* (Liddell et al. 1940). While these reference works have been scanned, encoded in XML and carefully edited to preserve the different categories of information found in their print counterparts, they remain static works. Our dynamic lexica for Greek and Latin will be responsive to their users' demands, drawing on the complete texts in our collection to present information about any part – or whole – of the entire corpus.

Figure 1 presented one possible view of a lexicon entry – a top-level view of a word from the perspective of the entire corpus (i.e., not how a word has been used in a single author or set of authors, but in all Latin texts). Three types of information are presented: a list of translation equivalents (found using the two word sense technologies described above), a list of possible subcategorizations and a list of predominant selectional preferences. Since we induce this information from the actual texts themselves, we are also in a position to present statistical information about the results (how frequently each phenomenon actually appears).

In Figure 1, the Latin verb *libero* is assigned three different translation equivalents, which correspond to coarse-grained sense distinctions: *set free*, *deliver* and *acquit*. Using word sense disambiguation on the texts without translations (and simple counting for the texts that do), we have found (hypothetically) that *set free* is used as a translation/sense in 573 instances in all of the

Latin texts in our collection, while *deliver* is found 232 times and *acquit* 115 times. These counts are further broken down by author and work and are all ultimately available as a canonical citation linked to the text itself.

The mock-up presented in Figure 8 presents a similar but slightly different presentation of the same word, limited this time to its use within one specific author. While Figure 1 presented a view of *libero* from the point of view of its use in all Latin texts, this view is more specific, and lets us examine how one specific author distributes the different senses.



līběro , āvi, ātum, 1 | Latin texts > Latin prose > **Jerome**

IV. Translation equivalents

▼ ***deliver*** (74%) (**68**)
  ▼ Vulgata (**68**)
    ▼ Genesis (**3**)
      - **Gen 3.8**
      - **Gen 17.11**
      - **Gen 28.1**

    ▶ Exodus (**17**)
    ▶ ...

▶ ***set free*** (26%) (**23**)
▶ ***acquit*** (0%) (**0**)

V. Subcategorization

▶ **SBJ OBJ** (26%) (**23**)
▶ **SBJ OBJ1 OBJ2** (74%) (**68**)

VI. Selectional preferences

▶ SBJ
▶ OBJ1
▼ OBJ2
  ▶ **manu** (44%) (**22**)
  ▶ **morte** (6%) (**3**)
  ▶ **ore** (6%) (**3**)

Figure 8: Mock-up of sample dynamic lexicon entry (author-specific)

These two figures together illustrate the inherent value of a customizable and dynamic lexicon – like a traditional lexicon, we can present large-scale information about a word's use in the entire corpus of classical literature, but we can also instantly present a view of the word tailored for one specific author – in essence, combining the virtues of both a large scale period reference work such as the *Mediae Latinitatis Lexikon Minus* with an individual one such as the *Thomas-Lexikon*. This customization is available on any level, so that we will not only be able to present information about a word's usage in a specific author, but also combinations of authors (e.g., the use of *archê* in both Thucydides and Herodotus) and genres (e.g., Greek prose) as well.

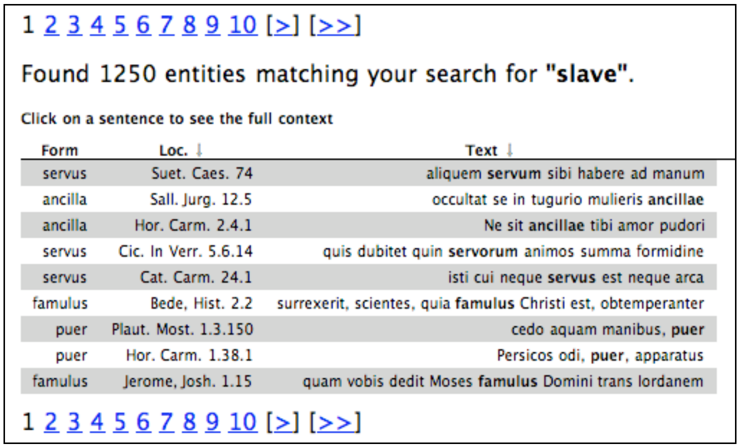
B.2 Searching

The dynamic lexicon resembles its more traditional print counterparts in that it is a work designed to be browsed: one looks up an individual headword and then reads its lexical entry. The technologies that will build this lexicon, however, do so by processing a Greek and Latin corpus of roughly nine million words. As the figures show, each entry in the lexicon ultimately ends with a list of canonical citations to fixed passages in the text. These citations are a natural index to a corpus, and provide the foundation for advanced methods of searching.

This ability to search through a collection of texts on a variety of levels – from individual word form (as in typical search engines) to word sense and subcategorization – will enable scholars to interact with a text in a way that was possible before only with extreme manual effort, if at all.

- *searching by word sense*

The ability to search a Latin or Greek text by an English translation equivalent is a close approximation to real cross-language information retrieval. Consider scholars researching Roman slavery: they could compare all passages where any number of Latin "slave" words appear, but this would lead to separate searches for *servus*, *serva*, *ancilla*, *famulus*, *famula*, *minister*, *ministra*, *puer*, *puella* etc. (and all of their inflections), plus many other less-common words. By searching for word sense, however, a scholar can simply search for *slave* and automatically be presented with all of the passages for which this translation equivalent applies. Figure 9 presents a mock-up of what such a service could look like.



1 2 3 4 5 6 7 8 9 10 [>] [>>]

Found 1250 entities matching your search for "**slave**".

Click on a sentence to see the full context

Form	Loc. ↓	Text ↓
servus	Suet. Caes. 74	aliquem **servum** sibi habere ad manum
ancilla	Sall. Jurg. 12.5	occultat se in tugurio mulieris **ancillae**
ancilla	Hor. Carm. 2.4.1	Ne sit **ancillae** tibi amor pudori
servus	Cic. In Verr. 5.6.14	quis dubitet quin **servorum** animos summa formidine
servus	Cat. Carm. 24.1	isti cui neque **servus** est neque arca
famulus	Bede, Hist. 2.2	surrexerit, scientes, quia **famulus** Christi est, obtemperanter
puer	Plaut. Most. 1.3.150	cedo aquam manibus, **puer**
puer	Hor. Carm. 1.38.1	Persicos odi, **puer**, apparatus
famulus	Jerome, Josh. 1.15	quam vobis dedit Moses **famulus** Domini trans Iordanem

1 2 3 4 5 6 7 8 9 10 [>] [>>]

Figure 9: Searching Latin texts by English word sense

Searching by word sense also allows us to investigate problems of changing orthography – both across authors and time: as Latin passes through the Middle Ages, for instance, the spelling of words changes dramatically even while meaning remains the same. So, for example, the diphthong *ae* is often reduced to *e*, and prevocalic *ti* is changed to *ci*. Even within a given time frame, spelling can vary, especially from poetry to prose. By allowing users to search for a sense rather than a specific word form, we can return all passages containing *saeculum*, *saeclum, seculum* and *seclum* – all valid forms for *era*. Additionally, we can automate this process to discover common words with multiple orthographic variations, and include these in our dynamic lexicon as well.

- *searching by selectional preference*

The ability to search by a predicate's selectional preference is also a step toward semantic searching – the ability to search a text based on what it "means." In building the lexicon, we automatically assign an argument structure to all of the verbs. Once this structure is in place, it can stay attached to our texts and thereby be searchable in the future, allowing us to search a text for the subjects and direct objects of any verb. Our scholar researching Roman slavery can use this information to search not only for passages where any slave has been freed (i.e., when any Latin variant of the English translation *slave* is the direct object of the active form of the verb *libero*), but also who was doing the freeing (who in such instances is the subject of that verb). This is a powerful resource that can give us much more information about a text than simple search engines currently allow.

C. Evaluation

An evaluation of this project involves two different sub-problems: evaluating the performance of the automatic methods we use to create the lexicon, and evaluating the overall significance that this resource has on the scholarly community.

C.1 Evaluating Performance

The three technologies on which this work is based – automatic sense induction, word sense disambiguation, and syntactic parsing – are all stochastic processes with measurable accuracy rates. The accuracy of word sense induction is a composite of the levels of alignment used to create it, and in evaluating this technology we will evaluate our accuracy at aligning source texts with their translations at the level of the chapter, sentence and individual word.

Word senses have proven to be recalcitrant to evaluation since even human annotators often disagree about a single meaning for a given word in context. The senses that we intend to disambiguate as part of this lexicon, however, will be coarse-grained, and we will also include information on human inter-annotator agreement for the test corpus on which we will evaluate the automatic processes. We will measure the accuracy of syntactic parsing using the PARSEVAL measures (Black et al. 1991) applied to dependency grammars (where labeled recall equals the number of correct dependencies in the candidate parse divided by the total dependencies in the gold standard parse, and labeled precision equals the number of correct dependencies in the candidate parse divided by the total number of dependencies attempted).

C.2 Evaluating Significance

The ultimate evaluation of a dynamic lexicon, however, must also go beyond an evaluation of the technologies used to create it. If part of our goal is to help traditional lexicographic work and the labors of students, then the evaluation must also include a measure of *significance*. For this we will we collaborate with our advisory board, which includes professional lexicographers as well as pedagogues, to judge the utility of this work as a professional instrument and as a vehicle for academic research.

We will elicit feedback as broadly as possible on two questions: first, what are the potential benefits and how can we evaluate the extent to which those benefits have been realized? We will study three spheres of impact:

- Impact within an established field: To what extent can the existing research community ask new questions and/or pursue its existing research more effectively?
- Impact across disciplines: To what extent can researchers from one field make better use of sources and/or results from another field?
- Impact upon the broader public: To what extent can we open up avenues of inquiry to new audiences who are not professional researchers?

Concrete instruments include surveys, structured interviews and formal meetings.

Over the course of the past year, we have held a two-day meeting involving projects actively developing tools and collections for Greek and Latin, a four-hour meeting of librarians to discuss scholarly services for the Open Content Alliance, a six-hour meeting aimed at recently tenured classicists to discuss the implications of emerging technologies for teaching and research in their field, and a three-day workshop bringing together humanists and computer scientists to discuss the implications of the emerging million book collections. Each of these meetings has set in motion a longer term conversation that will extend through the course of the project.

V. Work Plan

The work plan can be divided into three sections: one centering around the technologies required to create a dynamic lexicon, one exploring the place of that lexicon within the emerging cyberinfrastructure, and one focusing on the evaluation and dissemination of our work.

A. Technologies (Month 1-12)

While we have already completed a great deal of research on the three technologies needed to create a dynamic lexicon, there is still much more to be done. We will spend the first year of this project investigating word sense induction, word sense disambiguation and syntactic parsing and their role in inducing information for a dynamic lexicon.

Month 1-3: Word sense induction

Of the three technologies, we have already made the most progress on word sense induction, and will only spend the first three months applying our previous work on Latin to Greek, and in improving our performance on sentence alignment to increase the number of parallel sentences we can then use for the task of word alignment.

Month 4-8: Word sense disambiguation

In part, our progress on word sense induction in the first three months will largely feed directly into the task of word sense disambiguation, since the same techniques we use to improve sense induction can also improve disambiguation as well. Disambiguation, however, is harder than simple induction, so we will devote a larger fraction of our research time to this issue. Here we will investigate different methods for sense disambiguation given the resources available in our digital library.

Month 9-12: Automatic parsing

We have already been working on automatic parsing methods under our NSF funded project on creating a Latin treebank. While this research has been focused on the use of partial automatic parses as an aid for human annotators, we will focus our research in these last four months of our first year on the accuracy of such automatic methods for extracting lexical information, and in improving the parsing algorithms to maximize their effectiveness for this task. While we can use our Latin treebank to train an automatic parser for that language, we will also need to create a small treebank of ancient Greek to bootstrap a parser for it as well. For this we will modify the tools already in use for Latin and support two summer graduate students to annotate the Greek data.

B. Cyberinfrastructure (Month 13-22)

Once this work has been completed, we will next proceed to investigating the role of the lexicon in the cyberinfrastructure that forms the Perseus Digital Library, as a representative example of the kind of work that would need to be completed by any collection (historical or modern) to create a dynamic reference work of its own.

Month 13-17: Dynamic lexicon

The first five months will be spent devising methods to automatically create a lexicon from the texts in a digital library and to updating it when new texts are added.

Month 18-22: Search index

The next five months will be spent interfacing the dynamic lexicon with the texts in our collection to enable researchers to query our collection by word sense, subcategorization and selectional preference.

C. Dissemination and evaluation (Month 23-24)

Throughout this project, we will continually evaluate our work using the methods described in section IV.C: by calculating the accuracy rates for the various technologies in use, and by constant consultation with scholars to establish best practices for the field. While part of our dissemination will involve aspects of cyberinfrastructure completed during months 13-22, we will spend the final two months of this project evaluating our progress, disseminating all of the data (as described in section VII) and publishing our results to enable other digital collections to reproduce it.

VI. Staff

Gregory Crane will serve as project director and principal investigator for the project. We will also support two analysts: a programmer analyst with a background in computational humanities to perform work on the tools, including development, programming, and testing, and a research analyst who will be responsible for evaluation of the work, focus group interviews, authorship of white papers and journal articles, as well as conference presentations and outreach.

Since we will need to create a small Greek treebank to bootstrap a Greek parser, we will also support two summer graduate students to annotate this work (as we have in the past used such students to create a treebank for Latin).

We have created a large and disparate advisory committee — more than we can physically bring to Tufts. Some are local and most will be consulted at professional meetings and conferences. Much of our interaction will have to take place remotely by pointing advisors to services, marked-up data and written documents.

VII. Dissemination

Our recent grant-funded projects on named entity identification and constructing a Latin treebank demonstrate some of the methods of dissemination that we will use in this project as well. We released both the Latin treebank and the tagged output of our named entity analysis work as an XML corpus, available under a creative commons license; we published in the Tufts university repository several open access technical reports documenting how both systems worked and evaluating various elements of its effectiveness; we published articles describing this work in the proceedings of the Joint Conference on Digital Libraries, *D-Lib Magazine*, the European Conference on Digital Libraries, the Treebank and Linguistic Theories workshop, and the LaTeCH workshop sponsored by the Association for Computational Linguistics. These tangible publications, however, need to be reinforced by personal contacts, presentations, targeted demonstrations, small group discussions, e-mail correspondence, phone conversations, and other mechanisms of social networking. We have allocated an extensive amount of our budget to travel to facilitate such contacts as well as for small group evaluations.

Data released under a Creative Commons license

As in our prior NEH-funded project to digitize the Liddell, Scott and Jones Greek-English lexicon, we will also release the dynamic lexicon created in this project as an XML file under a creative commons license, encouraging others to study, augment, revise or otherwise build upon what we have done. As described above, this XML file will include information about a word's possible translation equivalents, weighted according to their usage in the corpus, along with information about common subcategorizations and selectional preferences. Each of those categories will also include a list of canonical citations documenting their appearance in the text.

Technical Reports

Technical reports play a key role in disseminating results. These allow us to provide timely, full and consistent coverage of key topics:

- <u>What methods were used</u>? How would others be able to replicate what we did, whether or not they use our code.

- <u>What other methods are available</u>? We cannot implement all approaches but we will provide a literature review pointing to what approaches and software packages are available.

- <u>Evaluation measures</u>: How well can we measure the performance of the techniques employed? How well do these measures compare with what the literature suggests we should be able to achieve?

- <u>Lessons learned</u>: What would need to be done next? What might we have done differently had we known ahead of time what would happen when we actually implemented techniques at scale?

Technical reports will be made freely available as a permanent part of the Tufts Digital Library. We expect at least two technical reports (one on the technologies involved in creating this work, and one on its place in a digital environment) but suspect that we may produce others.

Conferences, panels, small group discussions and lectures

We have budgeted travel expenses to attend and to hold discussion sessions in major conferences, including the APA/AIA, the MLA, the SAA and the RSA. Participation in panel sessions and presentation of papers are also important mechanisms.

Project members spend substantial time traveling and presenting their work at invited forums. The Principal Investigator allots time for up to three or four substantial lecture visits per semester.

Publication via Perseus

The most effective channel for communication may be the Perseus Digital Library itself, which serves 15,000,000 pages per month. This puts us in touch with a large and diverse community dedicated to the humanities and accustomed to digital tools. The services that we are able to implement will, where possible, become part of the overall Perseus DL infrastructure.

References

Bamman, David and Gregory Crane (2007), “The Latin Dependency Treebank in a Cultural Heritage Digital Library,” *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)* (Prague: Assocation for Computational Linguistics), pp. 33-40.

Banerjee, Sid and Ted Pedersen (2002). “An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet,” *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, pp. 136-145.

Bible Foundation and On-Line Book Initiative (1996), *Jerome’s Vulgate* (http://world.std.com/obi/Religion/Vulgate).

Black, E., S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, T. Strzalkowski (1991), "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars," in: *Proceedings of the DARPA Speech and Natural Language Workshop* (Pacific Grove, CA).

Bourne, Ella (1916), “The Messianic Prophecy in Vergil’s Fourth Eclogue,” *The Classical Journal* 11.7, pp. 390-400.

Brown, Peter F., Stephen A. Della Pietra ,Vincent J. Della Pietra and Robert L. Mercer, (1991), “Word-sense disambiguation using statistical methods,” *Proceedings of the 29th Conference of the Association for Computational Linguistics*, pp. 264-270.

Busa, R. (2004), “Foreword: Perspectives on the Digital Humanities,” in: Ray Siemens et al., *Blackwell Companion to Digital Humanities* (Oxford: Blackwell), pp. xvi-xxi.

Busa, R. (1974-1980), *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SI* (Stuttgart-Bad Cannstatt: Frommann-Holzboog).

Charniak, Eugene (2000), “A Maximum-Entropy-Inspired Parser,” in: *Proceedings of NAACL* (San Francisco).

Clark, Albert Curtis (ed.) (1908), *M. Tulli Ciceronis Orationes* (Oxford, Clarendon Press).

Collins, Michael (1999), “Head-Driven Statistical Models for Natural Language Parsing,” Ph.D. thesis (Philadelphia: University of Pennsylvania).

Crane, Gregory, David Bamman, Lisa Cerrato, Alison Jones, David M. Mimno, Adrian Packel, David Sculley and Gabriel Weaver (2006), “Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries,” *ECDL 2006*, pp. 353-366.

Crane, Gregory (1987), “From the Old to the New: Integrating Hypertext into Traditional Scholarship,” in: *Hypertext ’87: Proceedings of the 1st ACM Conference on Hypertext*, pp. 51-56.

Gale, William, Kenneth W. Church and David Yarowsky (1992), “Using bilingual materials to develop word sense disambiguation methods,” *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 101-112.

Greenough, J. B. (ed.) (1900), *Bucolics, Aeneid and Georgics of Vergil* (Boston: Ginn and Co.).

Grozea, Christian (2004), “Finding Optimal Parameter Settings for High Performance Word Sense Disambiguation,” *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp. 125-128.

Hajič, Jan (1999), “Building a Syntactically Annotated Corpus: The Prague Dependency Treebank,” in: E. Hajičová (ed.), *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová* (Prague, Czech Republic: Charles University Press).

Holmes, T. Rice (ed.) (1914), *C. Iuli Commentarii Rerum in Gallia Gestarum VII A. Hirti Commentarius VII* (Oxford: Clarendon Press).

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell (2004), “The Sketch Engine,” Proceedings of EURALEX 2004, pp. 105-116.

Klosa, A., U. Schnörch, P. Storjohann (2006), “ELEXIKO – A Lexical and Lexicological, Corpus-based Hypertext Information System at the Institut für deutsche Sprache, Mannheim,” *Proceedings of the 12th Euralex International Congress, Torino, Italy*, pp. 425-430.

Lesk, Michael (1986), “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone,” *Proceedings of the ACM-SIGDOC Conference*.

Lewis, Charles T., and Charles Short (eds.) (1879), *A Latin Dictionary* (Clarendon Press, Oxford).

Liddell, Henry George, and Robert Scott (1940), *A Greek-English Lexicon, revised and augmented throughout by Sir Henry Stuart Jones* (Oxford: Clarendon Press).

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1994), “Building a Large Annotated Corpus of English: The Penn Treebank,” *Computational Linguistics* 19.2, p. 313-330.

McCarthy, Diana, Rob Koeling, Julie Weeds and John Carroll (2004), “Finding Predominant Senses in Untagged Text,” *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 577-583.

McDonald, R., F. Pereira, K. Ribarov, and J. Hajič (2005), "Non-projective Dependency Parsing using Spanning Tree Algorithms," *Proceedings of HLT/EMNLP 2005*.

Mel’čuk, Igor A. (1988), *Dependency Syntax: Theory and Practice* (Albany: State University of New York Press).

Mihalcea, Rada and Philip Edmonds (eds.) (2004), *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (Barcelona, Spain).

Miller, George (1995), “Wordnet: A Lexical Database,” *Commnications of the ACM* 38.11, pp. 39-41.

Miller, George, Claudia Leacock, Randee Tengi and Ross Bunker (1993), “A Semantic Concordance,” *Proceedings of the ARPA Workshop on Human Language Technology*, pp. 303-308.

Moore, Robert C. (2002), “Fast and Accurate Sentence Alignment of Bilingual Corpora,” *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation*, pp. 135-144.

Niermeyer, Jan Frederick (1976), *Mediae Latinitais Lexicon Minus* (Leiden: Brill).

Nivre, J., Hall, J. and Nilsson, J. (2006), “MaltParser: A Data-Driven Parser-Generator for Dependency Parsing,” *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006),* pp. 2216-2219.

Och, Franz Josef, and Hermann Ney (2003), “A Systematic Comparison of Various Statistical Alignment Models,” *Computational Linguistics* 29.1, pp. 19-51.

Pinkster, Harm (1990), Latin Syntax and Semantics (London: Routledge).

Rundell, M. (ed.) (2002). *Macmillan English Dictionary for Advanced Learners*. Macmillan.

Rydberg-Cox, Jeffrey A. (2002), “From Lexicon To Commentary and Back Again,” *New England Classical Journal* 29.3:159-167.

Schütz, Ludwig (1895), *Thomas-Lexikon* (Paderborn: F. Schoningh).

Sinclair, J. M. (ed.) (1987). *Looking Up: an account of the COBUILD project in lexical computing*. Collins.

Singh, Anil Kumar and Samar Husain (2005), “Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs,” *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pp. 99-106.

Stewart, Gordon, Gregory Crane and Alison Babeu (2007), “A New Generation of Textual Corpora: Mining Corpora from Very Large Collections,” *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 356-365.

Tufis, Dan, Radu Ion and Nancy Ide (2004), “Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets,” *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pp. 1312-1318.