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

Computational Linguistics and Classical Lexicography

David Bamman <David_dot_Bamman_at_tufts_dot_edu>, Tufts University
Gregory Crane <gregory_dot_crane_at_tufts_dot_edu>, Tufts University

Abstract

Manual lexicography has produced extraordinary results for Greek and Latin, but it cannot in the immediate future provide for all texts the same level of coverage available for the most heavily studied materials. As we build a cyberinfrastructure for Classics in the future, we must explore the role that automatic methods can play within it. Using technologies inherited from the disciplines of computational linguistics and computer science, we can create a complement to these traditional reference works - a dynamic lexicon that presents statistical information about a word's usage in context, including information about its sense distribution within various authors, genres and eras, and syntactic information as well.

> ...Great advances have been made in the sciences on which lexicography depends. Minute research in manuscript authorities has largely restored the texts of the classical writers, and even their orthography. Philology has traced the growth and history of thousands of words, and revealed meanings and shades of meaning which were long unknown. Syntax has been subjected to a profounder analysis. The history of ancient nations, the private life of the citizens, the thoughts and beliefs of their writers have been closely scrutinized in the light of accumulating information. Thus the student of to-day may justly demand of his Dictionary far more than the scholarship of thirty years ago could furnish. (Advertisement for the Lewis & Short Latin Dictionary, March 1, 1879.)

The "scholarship of thirty years ago" that Lewis and Short here distance themselves from is Andrews' 1850 Latin-English lexicon, itself largely a translation of Freund's German Wörterbuch published only a decade before. As we design a cyberinfrastructure to support Classical Studies in the future, we will soon cross a similar milestone: the Oxford Latin Dictionary (1968-1982) has begun the slow process of becoming thirty years old (several of the earlier fascicles have already done so) and by 2012 the eclipse will be complete. Founded on the same lexicographic principles that produced the juggernaut *Oxford English Dictionary*, the *OLD* is a testament to the extraordinary results that rigorous manual labor can provide. It has, along with the Thesaurus Linguae Latinae, provided extremely thorough coverage for the texts of the Golden and Silver Age in Latin literature and has driven modern scholarship for the past thirty years. 1

Manual methods, however, cannot in the immediate future provide for all texts the same level of coverage available for the most heavily studied materials, and as we think toward Classics in the next ten years, we must think not only of desiderata, but also of the means that would get us there. Like Lewis and Short, we can also say that great advances have been made over the past thirty years in the sciences underlying lexicography; but the "sciences" that we group in that statement include not only the traditional fields of paleography, philology, syntax and history, but computational linguistics and computer science as well. 2

Lexicographers have long used computers as an aid in dictionary production, but the recent rise of statistical language 3

processing now lets us do far more: instead of using computers to simply expedite our largely manual labor, we can now use them to uncover knowledge that would otherwise lie hidden in expanses of text. Digital methods also let us deal well with scale. For instance, while the *OLD* 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.[1] 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.

4 In deciding how we want to design a cyberinfrastructure for Classics over the next ten years, there is an important question that lurks between "where are we now?" and "where do we want to be?": where are our colleagues already? Computational linguistics and natural language processing generally perform best in high-resource languages — languages like English, on which computational research has been focusing for over sixty years, and for which expensive resources (such as treebanks, ontologies and large, curated corpora) have long been developed. Many of the tools we would want in the future are founded on technologies that already exist for English and other languages; our task in designing a cyberinfrastructure may simply be to transfer and customize them for Classical Studies. Classics has arguably the most well-curated collection of texts in the world, and the uses its scholars demand from that collection are unique. In the following I will document the technologies available to us in creating a new kind of reference work for the future — one that complements the traditional lexicography exemplified by the *OLD* and the *TLL* and lets scholars interact with their texts in new and exciting ways.

Where are we now?

5 In answering this question, I am mainly concerned with two issues: the production of reference works (i.e., the act of lexicography) and the use that scholars make of them.

6 All of the reference works available in Classics are the products of manual labor, in which highly skilled individuals find examples of a word in context, cluster those examples into distinguishable "senses," and label those senses with a word or phrase in another language (like English) or in the source language (as with the *TLL*). In the past thirty years, computers have allowed this process to be significantly expedited, even in such simple ways as textual searching. Rather than relying on a vast network of volunteer readers to read through scores of books and write down "apt" sentences as they come across them (as with the *OED*), we can simply search our electronic corpora, find all examples of a word in context, and winnow through them sequentially to find those that most clearly illuminate the meaning of any given sense. This approach has been exploited most recently by the Greek Lexicon Project[2] at the University of Cambridge, which has been developing a *New Greek Lexicon* since 1998 using a large database of electronically compiled slips (with a target completion date of 2010). Here the act of lexicography is still very manual, as each dictionary sense is still heavily curated, but the tedious job of citation collection is not.

7 We can contrast this computer-assisted lexicography with a new variety — which we might more properly call "computational lexicography" — that has emerged with the COBUILD project [Sinclair 1987] of the late 1980s. The *COBUILD English Language Dictionary* (1987) is a learner's dictionary centered around a word's use in context, and is created from an analysis of an evolving English textual corpus (the Bank of English, on which current editions of the COBUILD dictionary are based, was officially launched in 1991 and now includes 524 million words[3]). This corpus evidence allows lexicographers to include frequency information as part of a word's entry (helping learners concentrate on common words) and also to include sentences from the corpus that demonstrate a word's common collocations — the words and phrases that it frequently appears with. By keeping the underlying corpus up to date, the editors are also able to add new headwords as they appear in the language, and common multi-word expressions and idioms (such as *bear fruit*) can also be uncovered as well.

8 This corpus-based approach has since been augmented in two dimensions. On the one hand, dictionaries and

lexicographic resources are being built on larger and larger textual collections: the German *elexiko* project [Klosa et al. 2006], for instance, is built on a modern German corpus of 1.3 billion words, and we can expect much larger projects in the future as the web is exploited as a corpus.[4] At the same time, researchers are also subjecting their corpora to more complex automatic processes to extract more knowledge from them. While word frequency and collocation analysis is fundamentally a task of simple counting, projects such as Kilgarriff's Sketch Engine [Kilgarriff et al. 2004] also enable lexicographers to induce information about a word's grammatical behavior as well.

9 In their ability to include statistical information about a word's actual use, these contemporary projects are exploiting advances in computational linguistics that have been made over the past thirty years. Before turning, however, to how we can adapt these technologies in the creation of a new and complementary reference work, we must first address the use of such lexica.

10 Like the *OED*, Classical lexica generally include a list of citations under each headword, providing testimony by real authors for each sense. Of necessity, these citations are usually only exemplary selections, though the *TLL* provides comprehensive listings by Classical authors for many of its lemmata. These citations essentially function as an index into the textual collection. If I am interested in the places in Classical literature where the verb *libero* means *to acquit*, I can consult the *OLD* and then turn to the source texts it cites: Cic. *Ver*. 1.72, Plin. *Nat*. 6.90, etc. For a more comprehensive (but not exhaustive) comparison, I can consult the *TLL*.

11 This is what we might consider a manual form of "lemmatized searching." The Perseus Digital Library[5] and the Thesaurus Linguae Graecae[6] both provide a form of lemmatized searching for their respective texts, but it is a fuzzier variety than that presented here: a user can search for a word form such as *edo* (*to eat*) and simultaneously search the texts for all of its various inflections, but ambiguity is rampant - a lemmatized search for *edo* would also search for *est*, which is also an inflection of the far more common *sum* (*to be*). The search results are thus significantly diluted by a large number of false positives.

12 The advantage of the Perseus and TLG lemmatized search is that it gives scholars the opportunity to find all the instances of a given word form or lemma in the textual collections they each contain. The *TLL* may be built on a comprehensive collection of 10 million slips containing all of Latin literature up to 200 CE and selections beyond, but that complete collection can only be found housed in their archives; what we have in print and on CD-ROM is still only a sample. The *TLL*, however, is impeccable in precision, while the Perseus and TLG results are dirty. What we need is a resource to combine the best of both.

Where do we want to be?

13 The *OLD* and *TLL* are not likely to become obsolete anytime soon; as the products of highly skilled editors and over a century of labor, the sense distinctions within them are highly precise and well substantiated. What we can provide in the near future, however, is a complement to these resources, one that presents statistics about a word's actual usage in texts — and not only in texts from the Classical period, but from any era for which we have electronic corpora. Heavily curated reference works provide great detail for a small set of texts; our complement is to provide lesser detail for *all* texts.

14 In order to accomplish this, we need to consider the role that automatic methods can play within our emerging cyberinfrastructure. I distinguish cyberinfrastructure from the vast corpora that exist for modern languages 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 (as mainly newswire), Classical 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.[7] 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.

We therefore must 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? 15

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* (to give) requires a subject, direct object and indirect object), and selectional preferences (e.g., that the subject of the verb *amo* (to love) is typically animate). With clustering techniques, we can establish the semantic similarity between two words based on their appearance in similar contexts. 16

If we leverage all of these techniques to create a lexicon for both Latin and Greek, the lexical entries in each reference work could include the following: 17

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; and
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 also 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. 18

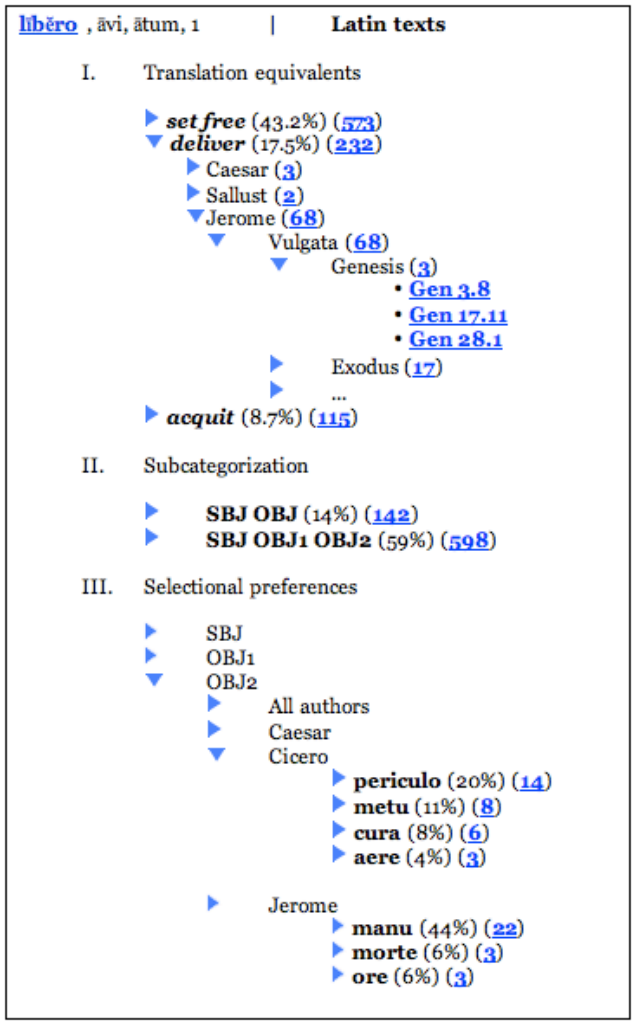


Figure 1. Mock-up of a sample entry in a dynamic lexicon

How do we get there?

We have already begun work on a dynamic lexicon like that shown in Figure 1 [Bamman and Crane 2008]. Our approach is to use already established methods in natural language processing; as such, our methodology involves the application of three core technologies: 19

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 the Perseus Project and in the natural language processing community at large. In the following I will detail how we can leverage them all to uncover large-scale usage patterns in a text. 20

Word Sense Induction

Our work on building a Latin sense inventory from a small collection of parallel texts in our digital library 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 LSJ definitions I.1 and 21

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 2. 22

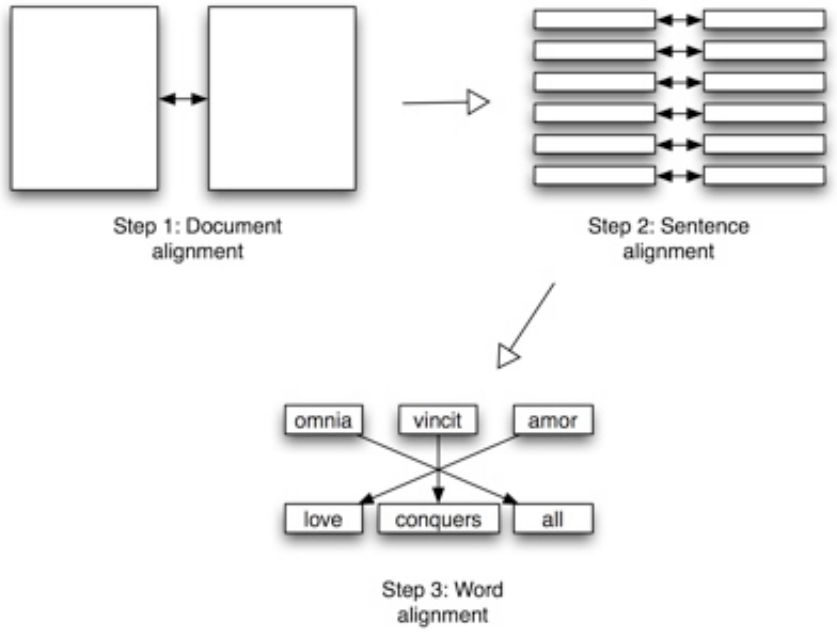


Figure 2. Alignment workflow

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]). 23

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 (MWEs) in the source language. 24

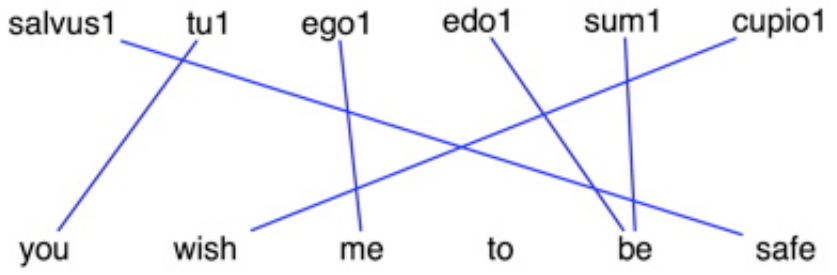


Figure 3. Sample word alignment from GIZA++

Figure 3 shows the result of this word alignment (here with English as the source language). The original, pre- 25

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 3), 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 4. 26

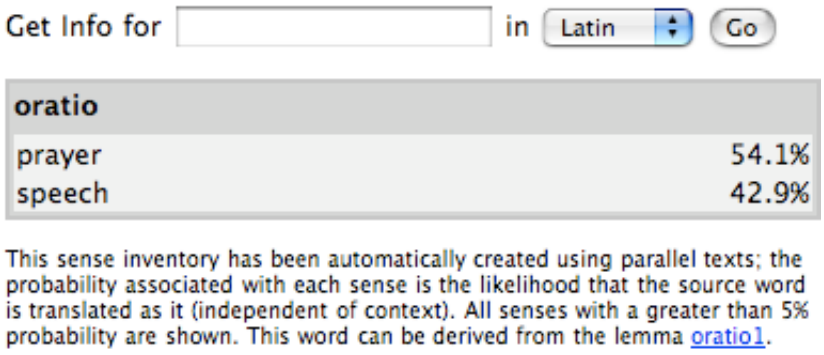


Figure 4. Sense inventory for *oratio* induced from parallel texts

The weighted list of translation equivalents we identify using this technique can provide the foundation for our further 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*. 27

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. 28

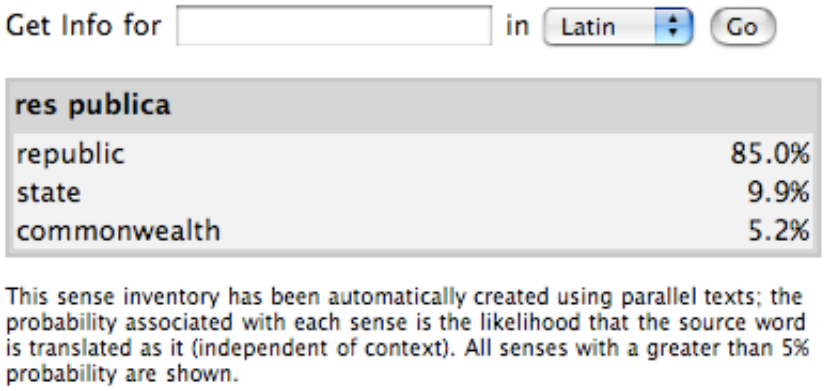


Figure 5. Sense inventory for the multi-word expression *res publica* induced from parallel texts

Second, our word alignment also maps multi-word expressions, so we can include significant collocations in our lexicon 29

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*).

Word Sense Disambiguation

Approaches to word sense disambiguation generally come in three varieties: 30

1. 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];
2. 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
3. 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.[8] 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. 31

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 using parallel texts for word sense disambiguation. This 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). 32

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. 33

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 filter is trained by a user explicitly labeling a message as spam, this classifier can be trained simply by the presence of an aligned translation. 34

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. 35

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. 36

Parsing

Two of the features we would like to incorporate into a 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 in *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). 37

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 *aes alienum* (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).[9] 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. 38

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 to the sheer volume of data involved — even the more resourceful of such endeavors (such as the Penn Treebank [Marcus et al. 1993] or the Prague Dependency Treebank [Hajič 1999]) take years to complete. 39

A second, more practical option is to assign syntactic structure to a sentence using automatic methods. Great progress 40

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 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.

41 We are currently in the process of creating a treebank for Latin, and have just begun work on a one-million-word treebank of Ancient Greek. Now in version 1.5, the Latin Dependency Treebank[10] is composed of excerpts from eight texts, including Caesar, Cicero, Jerome, Ovid, Petronius, Propertius, Sallust and Vergil. 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:

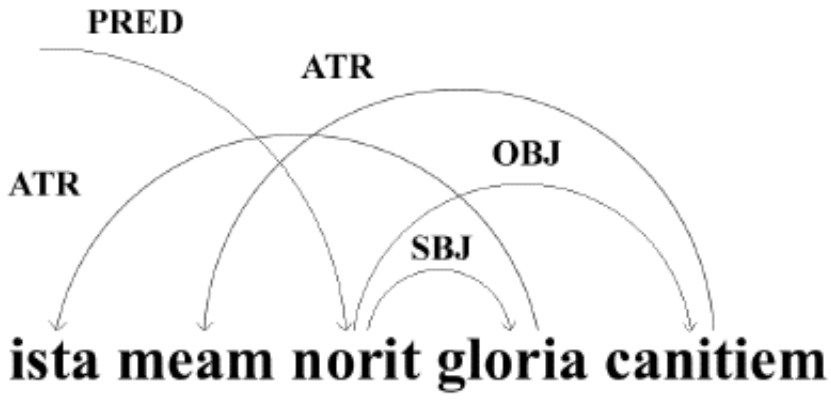


Figure 6. Dependency grammar representation of *ista meam norit gloria canitiem* ("that glory would know my old age")

42 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 induce valuable lexical information from its output by using a large corpus and simple hypothesis testing techniques to outweigh the noise of the occasional error [Bamman and Crane 2008]. The key to improving this parsing accuracy is to increase the size of the annotated treebank: the better the parser, the more accurate the syntactic information we can extract from our corpus.

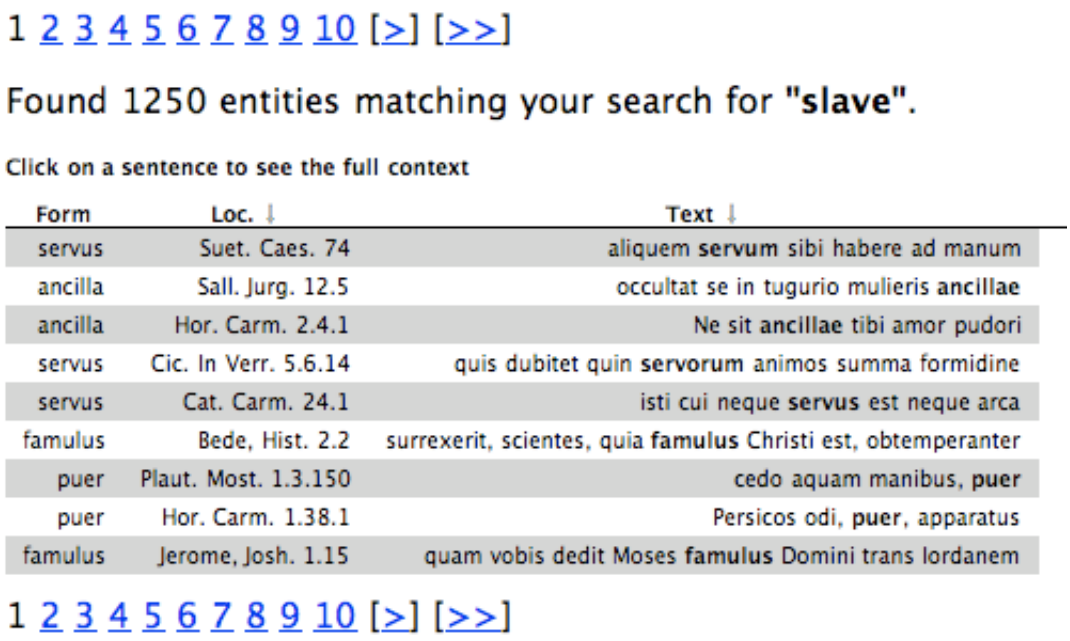
Beyond the lexicon

43 These technologies, borrowed from computational linguistics, will give us the grounding to create a new kind of lexicon, one that presents information about a word's actual usage. This 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 reference work, however, do so by processing a large Greek and Latin textual corpus. The results of this automatic processing go far beyond the construction of a single lexicon.

44 I noted earlier that all scholarly dictionaries include a list of citations illustrating a word's exemplary use. As Figure 1

shows, each entry in this new, dynamic lexicon ultimately ends with a list of canonical citations to fixed passages in the text. These citations are again a natural index to a corpus, but since they are based in an electronic medium, they provide the foundation for truly advanced methods of textual searching — going beyond a search for individual word form (as in typical search engines) to word sense.

Searching by word sense



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 7. Mock-up of a service to search Latin texts by English 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 7 presents a mock-up of what such a service could look like. 45

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. 46

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 47

about a text than simple search engines currently allow.

Conclusion

Manual lexicography has produced fantastic results for Classical languages, but as we design a cyberinfrastructure for Classics in the future, our aim must be to build a scaffolding that is essentially enabling: it must not only make historical languages more accessible on a functional level, but intellectually as well; it must give students the resources they need to understand a text while also providing scholars the tools to interact with it in whatever ways they see fit. In this a dynamic lexicon fills a gap left by traditional reference works. By creating a lexicon directly from a corpus of texts and then situating it within that corpus itself, we can let the two interact in ways that traditional lexica cannot. 48

Even driven by the scholarship of the past thirty years, however, a dynamic lexicon cannot yet compete with the fine sense distinctions that traditional dictionaries make, and in this the two works are complementary. Classics, however, is only one field among many concerned with the technologies underlying lexicography, and by relying on the techniques of other disciplines like computational linguistics and computer science, we can count on the future progress of disciplines far outside our own. 49

Notes

[1] 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).

[2] See http://people.pwf.cam.ac.uk/blf10/GLP/Greek_Lexicon_Project.htm.

[3] See http://www.collins.co.uk/books.aspx?group=153.

[4] In 2006, for example, Google released the first version of its Web 1T 5-gram corpus [Brants and Franz 2006] — a collection of n-grams (n=1-5) and their frequencies calculated from 1 trillion words of text on the web.

[5] See http://www.perseus.tufts.edu/hopper/.

[6] See http://www.tlg.uci.edu/.

[7] See [Bourne 1916] for an overview of *puer* in *Ec.* IV.

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

[9] See [Bamman and Crane 2007] for a summary of this work.

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

Works Cited

Andrews 1850 Andrews, Ethan Allen, ed. *A Copious and Critical Latin-English Lexicon: Founded on the Larger Latin-German Lexicon of Dr. William Freund;*. New York: Harper and Bros, 1850.

Bamman and Crane 2007 Bamman, David, and Gregory Crane. “The Latin Dependency Treebank in a Cultural Heritage Digital Library”. Presented at *LaTeCH 2007*. *Proceedings of the Workshop on Language Technology for Cultural Heritage Data* (2007), pp. 33-40. http://dl.tufts.edu/view_pdf.jsp?pid=tufts:PB.001.002.00002.

Bamman and Crane 2008 Bamman, David, and Gregory Crane. “Building a Dynamic Lexicon from a Digital Library”. Presented at *JCDL 2008*. *Proceedings of the Eighth ACM/IEE-CS Joint Conference on Digital Libraries* (2008), pp. 11-20.

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

Bourne 1916 Bourne, Ella. “The Messianic Prophecy in Vergil's Fourth Eclogue”. *The Classical Journal* 11: 7 (1916).

Brants and Franz 2006 Brants, Thorsten, and Alex Franz. *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium, 2006.

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

Busa 1974-1980 BusaRoberto. *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, 1974-1980.

Busa 2004 Busa, Roberto. “Foreword: Perspectives on the Digital Humanities”. In Susan Schreibman Ray Siemens and John Unsworth, eds., *Blackwell Companion to Digital Humanities*. Oxford: Blackwell Publishing, 2004.

Charniak 2000 Charniak, Eugene. “A Maximum-Entropy-Inspired Parser”. Presented at *NAACL 2000*. *Proceedings of the 2000 Conference of the North American Chapter of the Association for Computational Linguistics* (2000).

Collins 1999 CollinsMichael. *Head-Driven Statistical Models for Natural Language Parsing*. Thesis, University of Pennsylvania: 1999.

Freund 1840 Freund, Wilhelm. *Wörterbuch der lateinischen Sprache: nach historisch-genetischen Principien, mit steter Berücksichtigung der Grammatik, Synonymik und Alterthumskunde*. Leipzig: Teubner, 1840.

Gale et al. 1992a Gale, William, Kenneth W. Church and David Yarowsky. “Using Bilingual Materials to Develop Word Sense Disambiguation Methods”. Presented at *TMI 1992*. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation* (1992).

Glare 1982 GlareP. G. W., ed. *Oxford Latin Dictionary*. Oxford: Oxford University Press, 1968-1982.

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

Hajič 1999 Hajic, Jan. “Building a Syntactically Annotated Corpus: The Prague Dependency Treebank”. In Eva Hajicová, ed., *Issues of Valency and Meaning: Studies in Honour of Jarmila Panevová*. Prague: Charles University Press, 1999.

Kilgarriff et al. 2004 Kilgarriff, A., P. Rychly, P. Smrz and D. Tugwell. “The Sketch Engine”. Presented at *Eleventh Euralex Congress*. (2004), pp. 105-116.

Klosa et al. 2006 Klosa, Annette, Ulrich Schnörch and Petra Storjohann. “ELEXIKO -- A Lexical and Lexicological, Corpus-based Hypertext Information System at the Institut für deutsche Sprache, Mannheim”. Presented at *Euralex 2006*. *Proceedings of the Twelfth Conference of the European Association for Lexicography* (2006).

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

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

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

Marcus et al. 1993 Marcus, M., B. Santorini and M. Marcinkiewicz. “Building a Large Annotated Corpus of English: The Penn Tree Bank”. *Computational Linguistics* 19: 2 (1993), pp. 313-330.

McCarthy et al. 2004 McCarthy, Diana, Rob Koeling, Julie Weeds and John Carroll. “Finding Predominant Senses in Untagged Text”. Presented at *ACL 2004. Proceedings of the Forty-Second Annual Meeting o f the Association for Computational Linguistics* (2004).

McDonald et al. 2005 McDonald, Ryan, Fernando Pereira, Kiril Ribarov and Jan Hajic. “Non-projective Dependency Parsing using Spanning Tree Algorithms”. Presented at *HLT/EMNLP 2005. Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing* (2005).

Mel’čuk 1988 Mel’čuk, Igor A. *Dependency Syntax: Theory and Practice*. Albany: SUNY Press, 1988.

Mihalcea and Edmonds 2004 Mihalcea, Rada, and Phillip Edmonds, eds. *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (2004*. 2004. http://www.senseval.org/.

Miller 1995 MillerGeorge. “Wordnet: A Lexical Database”. *Communications of the ACM* 38: 11 (1995).

Miller et al. 1993 Miller, George, Claudia Leacock, Randee Tengi and Ross Bunker. “A Semantic Concordance”. Presented at *HLT 1993*. *Proceedings of the ARPA Workshop on Human Language Technology* (1993).

Moore 2002 Moore, Robert C. “Fast and Accurate Sentence Alignment of Bilingual Corpora”. Presented at *AMTA 2002*. *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas on Machine Translation* (2002).

Niermeyer 1976 Niermeyer, Jan Frederick. *Mediae Latinitatis Lexicon Minus*. Leiden: Brill, 1976.

Nivre et al. 2006 Nivre, Joakim, Johan Hall and Jens Nilsson. “MaltParser: A Data-Driven Parser-Generator for Dependency Parsing”. Presented at *LREC 2006*. *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (2006).

Och and Ney 2003 Och, F.J., and H. Ney. “A Systematic Comparison of Various Statistical Alignment Models”. *Computational Linguistics* 29 (2003), pp. 19-51.

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

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

Sinclair 1987 Sinclair, John M., ed. *Looking Up: an account of the COBUILD project in lexical computing*. London and Glasgow: Collins, 1987.

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

TLL Warning: Biblio formatting not applied. *Thesaurus Linguae Latinae, fourth electronic edition*. Munich. K. G. Saur. 2006.

Tufis et al. 2004 Tufis, Dan, Radu Ion and Nancy Ide. “Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets”. Presented at *COLING 2004*. *Proceedings of the Twentieth International Conference on Computational Linguistics* (2004).