Preprint of paper accepted to the JCDL 2008.

Building a Dynamic Lexicon from a Digital Library

David Bamman
The Perseus Project
Tufts University
Medford, MA
david.bamman@tufts.edu

Gregory Crane
The Perseus Project
Tufts University
Medford, MA
gregory.crane@tufts.edu

ABSTRACT

We describe here in detail our work toward creating a dynamic lexicon from the texts in a large digital library. By leveraging a small structured knowledge source (a 30,457 word treebank), we are able to extract selectional preferences for words from a 3.5 million word Latin corpus. This is promising news for low-resource languages and digital collections seeking to leverage a small human investment into much larger gain. The library architecture in which this work is developed allows us to query customized subcorpora to report on lexical usage by author, genre or era and allows us to continually update the lexicon as new texts are added to the collection.

Categories and Subject Descriptors

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

General Terms

Design, Documentation, Performance

Keywords

Lexicography, syntactic parsing, digital libraries

1. INTRODUCTION

Lexicographers have been exploiting large corpora for structured knowledge since the COBUILD project [38] of the 1980s, often in the form of extracting frequency counts and collocations – a word's frequency information is especially important to second language learners, and collocations (a word's "company") are instrumental in delimiting its meaning. This corpus-based approach to lexicon building 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 [23], 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.[1] 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 [22] also enable lexicographers to induce information about a word's grammatical behavior as well.

We are in the process now of creating a customizable dynamic lexicon from the Classical texts in the Perseus Digital Library [12, 14]. This lexicon will present a sense inventory (along with frequency information) for any Greek or Latin lexeme as it is used in any author, era or genre found in our collection, along with statistical information about its common subcategorization frames and selectional preferences as well.

While the sense inventory itself is dependent on technologies of word sense induction and disambiguation, extracting the subcategorization frames and selectional preferences for a word is based on automatic morphological tagging and syntactic parsing. State-of-the-art morphological taggers can achieve accuracy rates of over 96% for English [34, 36] and 92% for highly inflected languages like Czech [20], and dependency parsers can achieve labeled accuracy rates for the same languages of 86% [31] and 80% [11], respectively.[2] These services, however, achieve such high accuracies by being trained on large volumes of manually annotated data, usually over one million words.[3]

We have, in contrast, a Latin treebank of 30,457 words. The small training size of this dataset leads to predictably inferior tagging and overall parsing. As Church and Hovy [9] noted for machine translation, however, the evaluation of a system's performance is dependent on the application. 30,457 words may not be enough for accurate syntactic parsing as an end in itself, but the imperfectly parsed sentences that result from it are sufficient to induce strong lexical information given a large enough number of them. By using the same simple hypothesis testing techniques used to find collocations (amidst sentences full of noise), we can identify

[1] In 2006, for example, Google released the first version of its Web 1T 5-gram corpus [6], a collection of n-grams (n=1-5) and their frequencies calculated from 1 trillion words of text on the web.

[2] Unlabeled parsing accuracy (in which only the head is evaluated, not the syntactic relationship), nets higher accuracy rates of 91% for English [11] and 84% for Czech [29].

[3] The Penn Treebank [28] for instance contains over one million words (in PTB-2 style), while the Prague Dependency Treebank [21] contains 1.5 million.

JCDL'08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.

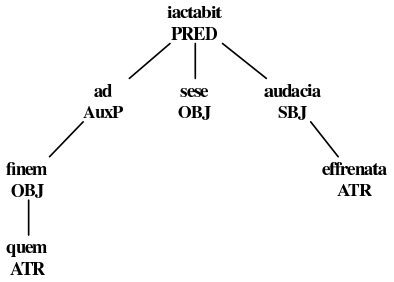


Figure 1: Dependency graph of the treebank annotation for *quem ad finem sese effrenata iactabit audiacia* ("to what end will your unbridled audacity throw itself?"), Cicero, *In Catiliam* 1.1.

```
<sentence id="74" document_id="Perseus:text:1999.02.0010" subdoc="text=Catil.:Speech=1:chapter=1" span="quem0:audacia0">
  <word id="1" form="quem" lemma="quis1" postag="p-s---ma-" head="3" relation="ATR"/>
  <word id="2" form="ad" lemma="ad1" postag="r--------" head="6" relation="AuxP"/>
  <word id="3" form="finem" lemma="finis1" postag="n-s---ma-" head="2" relation="OBJ"/>
  <word id="4" form="sese" lemma="sui1" postag="p-s---fa-" head="6" relation="OBJ"/>
  <word id="5" form="effrenata" lemma="effreno1" postag="t-srppfn-" head="7" relation="ATR"/>
  <word id="6" form="iactabit" lemma="jacto1" postag="v3sfia---" head="0" relation="PRED"/>
  <word id="7" form="audacia" lemma="audacia1" postag="n-s---fn-" head="6" relation="SBJ"/>
</sentence>
```

Figure 2: XML version of the treebank annotation for *quem ad finem sese effrenata iactabit audiacia*, Cicero, *In Catiliam* 1.1.

common selectional preferences for a word when the automatic parse is noisy itself. This is promising for other low resource languages and digital libraries seeking to leverage small structured knowledge sources against large unstructured collections. While the work here has been developed in the context of a single digital library, the supervised learning techniques we describe can be used by a collection of any size, given a small set of annotated data.

2. RESOURCES

We have two different types of resources in our digital library: a small but human-curated set of syntactically annotated data, and a far larger but unannotated corpus of texts.

2.1 Annotated data

The small structured knowledge source at our disposal is a 30,457 word treebank of Classical Latin. Now in version 1.4, the Latin Dependency Treebank is comprised of excerpts from five texts: Caesar's *Commentarii de Bello Gallico*, Cicero's *Oratio in Catilinam*, Sallust's *Bellum Catilinae*, Vergil's *Aeneid* and Jerome's *Vulgate*, as shown in table 1.

Date	Author	Words	Sentences
1st c. BCE	Caesar	1,488	71
1st c. BCE	Cicero	5,663	295
1st c. BCE	Sallust	12,311	701
1st c. BCE	Vergil	2,613	178
4th-5th c. CE	Jerome	8,382	405
	Total	30,457	1,650

Table 1: Composition of the Latin Dependency Treebank.

A treebank is large collection of sentences that have been syntactically annotated. The knowledge encoded in this structure is extremely labor intensive, as two independent annotators each annotate every sentence, and their annotations are reconciled by a third. The process of annotation itself involves specifying the exact syntactic relationship for every word in a sentence (e.g., what the subject is, what the object is, where the prepositional phrase should be attached, which adjective modifies which noun, etc.). In addition to the index of its syntactic head and the type of relation to it, each word in the treebank is also annotated with the lemma from which it is inflected (e.g., that *est* is an inflected form of the lemma *sum*) and its morphological code (e.g., that *est* is a 3rd person singular indicative active verb).

Figures 1 and 2 present two views of a syntactic annotation for a single sentence (*quem ad finem sese effrenata iactabit audacia*).[4] Figure 1 shows the conceptual structure for a dependency tree that results from the annotation (subjects and objects, for instance, are both children of the verbs they modify), and figure 2 presents an XML serialization of that tree (the format in which we release our data).

Since Latin has a highly flexible word order, we have based our annotation style on the dependency grammar used by the Prague Dependency Treebank (PDT) [19] for Czech while tailoring it for Latin via the grammar of Pinkster [33]. Dependency grammars differ from constituent-based grammars by foregoing non-terminal phrasal categories (such as NP or VP) and instead linking words themselves to their immediate head. This is an especially appropriate manner of representation for languages with a moderately free word order (such as Latin and Czech), where the linear order of constituents is broken up with elements of other constituents.

In order make our annotation style as useful as possible, we are also collaborating with other Latin treebanks (notably the Index Thomisticus [7, 32] on the works of Thomas

[4]"To what end will your unbridled audacity throw itself?" (Cicero, *In Catilinam* 1.1).

Aquinas) to create a common set of annotation guidelines to be used as a standard for Latin of any period [5]. This work has also allowed us to share our data as we annotate our respective texts [4].

2.2 Unannotated data

The set of syntactically annotated data we have in our collection is dwarfed in comparison to the size of the unannotated corpus. The Perseus Digital Library contains approximately 3.5 million words of Latin source texts, along wih 4.9 million words of Greek. While these texts are unstructured syntactically, they each possess extensive metadata detailing author and the sub-collections to which the work belongs (such as *Latin poetry* or *Latin prose*).

Our approach to extracting lexical information from this large collection involves first assigning syntactic parses to all of the sentences it contains. We cannot of course manually parse each sentence by hand, so the syntactic structure must be automatically assigned. State-of-the-art syntactic parsing is a supervised learning process in which a parser is trained on a set of human-annotated data. Parsing performance is strongly linked to the size of the training data, with large treebanks (over one million words) performing the best. Before mining this corpus, we evaluated the performance of the parsing algorithm itself on our small dataset, and of the automatic morphological tagging on which it relies.

3. EVALUATION

We evaluated the accuracy of automatic morphological tagging using the TreeTagger analyzer [36] and of automatic syntactic parsing using McDonald et al.'s dependency parser [29]. In all of the tests that follow, the accuracy rates reported are the result of a 10-fold test on our 30,457 word treebank, in which the tagger or parser is trained on 9/10 of the treebank (ca. 27,411 words) and tested on the remaining one-tenth; this test is conducted a total of ten times, once for each held-out tenth, with the reported accuracy being the average of all tests.

3.1 Morphological tagging

As part of a highly inflected language, Latin words have a morphological analysis comprised of nine features: part of speech, person, number, tense, mood, voice, gender, case and degree. The TreeTagger analyzer performed with an accuracy of 83% in correctly disambiguating the full morphological analysis. In resolving simple part of speech its performance is close to that of higher-resource languages (95%), but Latin's complex inflection presents more difficulties in resolving gender and case. Both of these features have a higher entropy in the language due to their overlapping ambiguity.[5]

3.2 Syntactic parsing

Most evaluations of parsing accuracy presume a gold standard for the underlying morphological tags in order to isolate the specific gain or loss in the parser itself. In determining the functional accuracy we might expect of a parser in assigning a syntactic analysis to all of the sentences in our corpus (for which we must automatically assign a morphological analysis as well), we present two evaluations: one for parsing a corpus with known morphological tags ("gold") and one for parsing a corpus for which the morphological tags have been automatically assigned ("automatic"). Unlabeled accuracy measures how often the syntactic head of a word is correct, while labeled accuracy also measures whether the correct syntactic tag (such as *subject* vs. *object*) has been applied as well.

Table 2: Morphological accuracy by feature

	Accuracy
Case	90.10%
Degree	99.92%
Gender	92.90%
Mood	98.68%
Number	95.15%
Part of speech	95.11%
Person	99.56%
Tense	98.62%
Voice	98.89%
All	83.10%

Table 3: Parsing Accuracy

	Unlabeled	Labeled
Gold	64.99%	54.34%
Automatic	61.49%	50.00%

As expected, the overall accuracy for the gold evaluation is much lower than that reported for languages such as English and Czech. With automatic morphological tags, we might expect to find about half of the syntactic relations in a sentence. We can break this figure down even further, however. The overall accuracy reported in table 3 is a composite of all authors, genres, and syntactic relations. If we divide those results by author (table 4), we find a strong correlation between parsing accuracy and the author's non-projectivity – the ratio with which phrasal constituents are broken up by other constituents.[6] Jerome, a prose author writing in the 4th century CE, has a low non-projectivity rate of 1.8%, while Vergil, a Golden Age poet, has the highest at 12.2%.[7] High non-projectivity is a hallmark of Latin poetry as a form of rhetorical effect (*hyperbaton*), so we can expect our lowest accuracy rates in the future to fall among the works of stylized poets and the highest to come from strict prose authors. Fortunately (in this regard), the corpus of Latin poetry is much smaller than that of prose (the Perseus Digital Library, for example, includes 593,000 words of Latin poetry and 2.9 million words of prose).

Another variable included in this overall accuracy rate is the parser's performance by individual tag. As table 5 shows, precision[8] and recall[9] are much higher for attributive

[5]For example, a word like *magna* (great) can be a feminine nominative adjective or a neuter accusative one (not to mention feminine ablative or neuter nominative as well).

[6]See Nivre [30] for a formal definition of projectivity.

[7]See Bamman and Crane [1] for a full list.

[8]We define precision here to be the number of times a tag X is correctly assigned to the correct head divided by the number of occurrences of that tag X in the automatically parsed corpus.

[9]We define recall here to be the number of times a tag X is correctly assigned to the correct head divided by the number of occurrences of that tag X in the test corpus.

Table 5: Labeled Precision/Recall by Syntactic Tag

Gold			Automatic		
	Precision	Recall		Precision	Recall
ATR	68.17%	71.20%		63.09%	62.41%
AuxP	67.38%	69.80%		63.66%	66.81%
SBJ	61.95%	62.24%		50.93%	51.10%
OBJ	59.33%	62.84%		50.90%	55.12%
ADV	53.72%	59.90%		49.24%	55.31%
AuxC	39.30%	39.00%		34.80%	36.04%
SBJ_CO	38.20%	39.81%		26.58%	29.04%
OBJ_CO	37.90%	38.48%		31.84%	30.85%
ATR_CO	34.38%	27.76%		30.35%	25.17%
ADV_CO	34.27%	27.84%		30.29%	22.22%

Table 4: Labeled Parsing Accuracy by Author

	Gold	Automatic
Jerome	61.44%	58.15%
Sallust	53.04%	46.99%
Caesar	51.34%	46.24%
Cicero	49.97%	44.41%
Vergil	48.99%	40.60%

adjectives (ATR), prepositional phrase attachment (AuxP), subjects (SBJ), objects (OBJ) and adverbs (ADV) than they are for subordinating conjunction attachment (AuxC) and any relation involved in coordination (_CO). This is a good sign for extracting selectional preferences from a corpus, since the relationships we will be looking for will be exactly these – while the precision of subjects and objects still hovers around 50%, the precision of attributes at least is higher at 63%.

4. EXTRACTING SELECTIONAL PREFERENCES

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, based on a small manual study of 100 instances of the verb [2]: 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). These are syntactic qualities since each of these arguments bears a direct syntactic relation to its head as much as it holds a semantic place within the underlying argument structure.

Selectional preferences are a variety of collocation, and can be extracted using similar methods [8] – where collocations can be found by comparing the count of two words occurring together (typically within some fixed span of words) with the independent likelihood of each occurring on its own, selectional preferences can be found by establishing the likelihood that a word bears a specific syntactic relationship to another – the most informative of these being direct objects (OBJ). Using clustering [35] or WordNet similarity metrics [10], we can then also use individual word frequencies to generalize to the class of word that a predicate prefers.

4.1 Tagging the data

In order to extract selectional preferences from our 3.5 million word Latin corpus, we first trained our tagger and parser on the full treebank, then used those trained models to morphologically tag the entire corpus and then assign syntactic structure to the automatically tagged texts.

4.2 Extracting knowledge

With the entire corpus tagged and parsed, we can now extract selectional preferences from it. Strength of association, however, is skewed by a word's overall frequency in a corpus, so that a high frequency word would naturally be a common argument for many transitive verbs. We can overcome this by using the same hypothesis testing techniques used to find common collocations. The log likelihood test (λ) [15], for example, measures how often two words occur together in a sentence compared to how often one would expect to find them together, given their frequencies in the overall corpus.[10] To adopt this measure to finding common selectional preferences, we can define the relevant counts to be the following:

$$c_1 = \text{count of lemma}_1 \text{ in the corpus}$$
$$c_2 = \text{count of lemma}_2 \text{ in the corpus}$$
$$c_{12} = \text{count of lemma}_2 \text{ depending as an argument of lemma}_1$$

With the log λ value being:

$$\log \lambda = \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)$$
$$\text{where } p = \frac{c_2}{N},\ p_1 = \frac{c_{12}}{c_1},\ p_2 = \frac{c_2 - c_{12}}{N - c_1},\ N = \textit{corpus count}$$
$$\text{and } L(a, b, c) = c^a (1 - c)^{b-a}.$$

[10]We use log likelihood as distinct from mutual information to avoid privileging collocates of low-frequency words at the expense of more well-attested pairs. For our purposes, log likelihood and χ^2 are largely interchangeable – a χ^2 evaluation of *do*, for example, provides the same identically ranked list as that found using log likelihood (table 7) below.

To achieve a confidence level of $\alpha = 0.05$ that one lemma is a common argument of another, their $-2 \log \lambda$ value must be above 3.84.

4.3 Results

The strength of hypothesis testing is that it also allows us to overcome our noisy data. Given the size of our corpus, we can be forgiving of occasional parsing errors since the counts of most lemmas are relatively high: if a word is a true common argument of another word, it will appear as that argument several times over 3.5 million words.

We can see the strength of this approach in tables 6 and 7. Table 6 presents the strongest relationships found between two words in the entire corpus (not simply selectional preferences or arguments, but all words that bear some syntactic relationship to each other). Nine of the ten pairs of words are connected with an attributive relationship and present strong collocations.

Table 6: Ten strongest syntactic collocates (uninflected root forms shown)

Latin	English	Relation	$-2 \log \lambda$
res publicus	republic	ATR	3840.0
populus romanus	Roman people	ATR	2450.8
pater conscribo	conscript father	ATR	612.6
filius Israhel	son of Israel	ATR	524.1
deus dominus	lord god	ATR	346.2
terra Aegyptius	Egypt	ATR	324.3
do opera	take pains	OBJ	254.2
rex Babylon	king of Babylon	ATR	249.0
deus immortalis	immortal god	ATR	238.6
bellum civilis	civil war	ATR	190.1

Table 7 in contrast presents the common selectional preferences for a single lemma, *do* (to give).

Table 7: Strongest OBJ of *do* (to give). Column OLD lists the entry for which it is given as an exemplary use in the *Oxford Latin Dictionary*.

Latin	English	OLD	$-2 \log \lambda$
opera	service (= take pains)	22c	254.2
obses	hostage	11a	21.8
signum	sign	-	12.6
velum	sail (= set sail)	18f	7.9
pecunia	money (= pay)	6a	7.3
negotium	business	-	6.2
poena	penalty (= suffer)	7b	5.6
possessio	possession	1c	4.8
littera	letter (for delivery)	10a	4.3
osculum	kiss	8a	4.1
tergum	back (= turn)	18d	3.4

Here we begin to see the difference between an idiomatic collocation (*do opera*) (literally, "to give work", idiomatically "to take pains"), with a log likelihood of 254.2, and selectional preferences with scores dropping below the collocational threshold (3.84) but still very typical arguments. We can judge the strength of these associations by comparing them with entries in a traditional Latin lexicon, the *Oxford Latin Dictionary* (OLD) [17]: in this we are not determining a gold-standard precision but rather the inter-annotator agreement between this automated output and a human editor. For *do*, 9 of its 11 strongest objects are cited as exemplary uses in the OLD – only *signum* and *negotium* are omitted.

Strong selectional preferences also let us distinguish between lemmas with similar meanings. The Latin words *ago* (to drive) *gero* (to bear) and *duco* (to lead) are all commonly used in a nondescript sense to specify that some action took place (*ago*, for instance, is the Latin root of the English word *agent*). This abstractness gives rise to arguments with specialized meanings. By extracting the selectional preferences of these verbs, we can compare them and isolate those arguments that distinguish them from each other. Table 8 presents all arguments of *ago*, *gero* and *duco* with a log likelihood score above 1.[11] Many of these objects form idiomatic expressions with their head (e.g., *ago* + *gratia*, "to thank"), and all but one can be found as an exemplary use of the verb in the *OLD*. The use of the verb *gero* in particular highlights the possibility of further clustering these individual words into larger classes: three of its strongest objects are official offices (*praetura*, "praetorship"; *censura*, "office of censor"; and *magistratus*, "magistracy").

Table 8: Strongest OBJ of *ago* (to drive), *gero* (to bear) and *duco* (to drive). Column OLD lists the entry for which it is given as an exemplary use in the *Oxford Latin Dictionary*.

Latin	English	OLD	$-2 \log \lambda$
ago			
gratia	thanks (= give thanks)	28b	50.6
paenitentia	repentence (= to repent)	28a	21.8
nugae	trifles (= to trifle)	22b	7.7
causa	cause (legal)	42c	1.2
aetas	age (=to be X years old)	31a	1.0
gero			
bellum	war (= make war)	8b	10.5
praetura	praetorship	10	3.2
mos	custom	8d	2.7
censura	office of censor	-	1.5
magistratus	magistracy	10	1.4
duco			
uxor	wife (= to marry)	5a	11.6
exercitus	army (= to march)	6a	1.3
pompa	parade	7a	1.0

[11] One issue that requires further research is establishing a constant threshold for discovering interesting selectional preferences – as shown above, important lexical information does fall below the standard collocational threshold (as evidenced by agreement with the human editors of the *OLD*). If we compare the extracted selectional preferences with a log likelihood score above 0.5 for the four verbs shown in tables 7 and 8 (a total of 69 arguments), we naturally find a tapering of agreement with the *OLD* – above 3.4, the agreement holds at 86.7%, and falls below 80% around 2.7 and 70% at 1.5. This of course is based on a small sample of four verbs and only begins to suggest the direction in which we should look to find an adequate threshold; in the future we plan a more comprehensive evaluation.

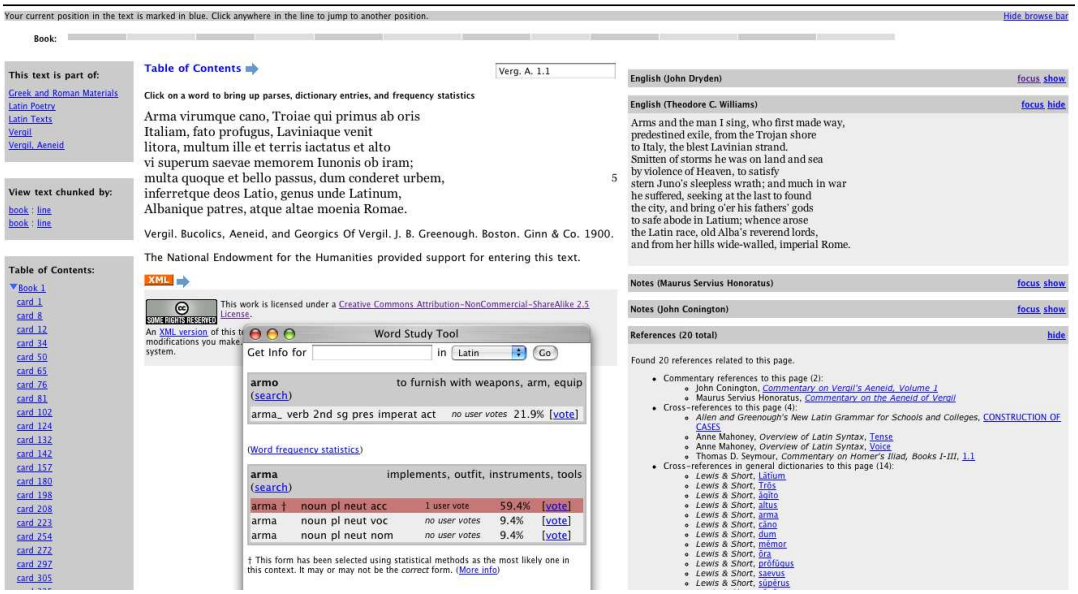


Figure 3: A screenshot of Vergil's *Aeneid* from the Perseus Digital Library.

5. LEXICAL INFORMATION IN A DIGITAL LIBRARY

A digital library architecture interacts with this knowledge in three ways: first, it lets us further contextualize our source texts for the users of our existing digital library; second, it allows us to present customized reports for word usage according to the metadata associated with the texts from which they're drawn, enabling us to create a dynamic lexicon that not only notes how a word is used in Latin in general, but also in any specific author, genre, or era (or combination of those). And third, it lets us continue to mine more texts for the knowledge they contain as they're added to the library collection, essentially making it an open-ended service.

5.1 Contextualization

Figure 3 shows a screenshot from our existing digital library. In this view, the reader is looking at the first seven lines of Vergil's *Aeneid*. The source text is provided in the middle, with contextualizing information filling the right column. This information includes:

- Translations. Here two English translations are provided, one by the 17th-century English poet John Dryden and a more modern one by Theodore Williams.
- Commentaries. Two commentaries are also provided, one in Latin by the Roman grammarian Servius, and one in English by the 19th-century scholar John Conington.
- Citations in reference works. Classical reference works such as grammars and lexica often cite particular passages in literary works as examples of use. Here, all of the citations in such reference works to any word or phrase in these seven lines are presented at the right.

Additionally, every word in the source text is linked to its morphological analysis, which lists every lemma and morphological feature associated with that particular word form. Here the reader has clicked on *arma* in the source text. This tool reveals that the word can be derived from two lemmas (the verb *armo* and the noun *arma*), and gives a full morphological analysis for each. A recommender system automatically selects the most probable analysis given the context, and users can also vote for the form they think is correct.

The selectional preference information that we have mined from our collection is another method of providing further contextual information for our users. While all of the words in a source text are linked to their lexical entries by means of their morphological analysis,[12] we are able to provide a knowledge source that complements human-curated lexica by also providing frequency information (and log likelihood scores) as a substantiation for an object's predominance.

5.2 Creating customized subcorpora

The results on the usage of the verb *do* presented above are drawn from our entire Latin corpus of 3.5 million words. The benefit of having this knowledge in a digital library is the structure that the library architecture imposes on it. The texts in our collection all have metadata associated with them that specify their author, genre, and all of the various collections to which they belong (for example, Vergil's *Aeneid* is part of the collected works of *Vergil*, which is part of *Latin poetry*, which is part of *Latin texts*). This same architecture is preserved in the automatically parsed data, so we can query and present information tailored to specific authors or genres.

Conducting this same search on three subsets of our entire corpus – all of the works authored by Caesar, Jerome, and Ovid – provides the results given in table 9. Here we clearly see the relevance of searching these selections of our entire corpus, as the word usage of the verb clearly differs according to the genre of each author. Caesar characteristically uses *do* in what can be called a "military" sense, such as with

[12]All Latin lemmas, for instance, are linked to their dictionary entries in the Lewis *Elementary Latin Dictionary* [26] and the Lewis and Short *Latin Dictionary* [27]

obses ("hostages"); Jerome, an apostolic father whose Latin works are predominantly comprised of the Vulgate Bible, uses *do* to provide drink, food, rest and glory; while the most common objects given in Ovid, a love poet, include kisses (*osculum*) and gifts (*munus*). Note that we need not simply restrict ourselves to searching by author – we can search by any element of the metadata that attends these texts, or any combination of fields (e.g., all Roman historical writing except the works of Tacitus plus all Latin elegaic poetry written before the turn of the millennium).

Table 9: Strongest OBJ of *do* by individual author

Latin	English	$-2 \log \lambda$
Caesar		
obses	hostage	18.4
opera	service	11.9
suspicio	suspicion	2.2
facultas	faculty	1.8
signum	sign	1.5
Ovid		
osculum	kiss	8.5
velum	sail	5.9
munus	gift	3.5
signum	sign	2.6
Jerome		
potus	drink	16.6
esca	food	3.3
requies	rest	3.0
gloria	glory	2.4
terra	earth	1.6

Figure 4 presents an example of what a full lexical entry would look like in the context of a digital library. While this entry shows a top-level view of the word and its use in all of Latin, all of its categories are broken down by specific subcorpora, such as its use in individual authors.

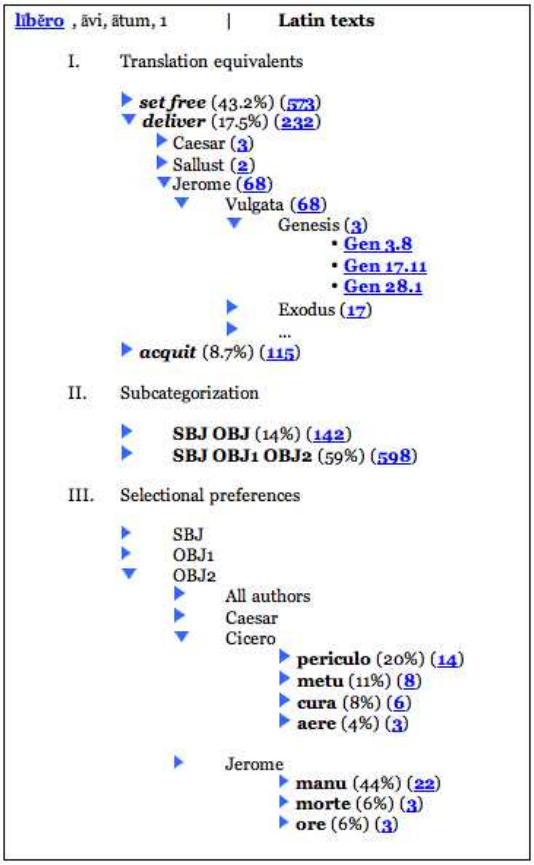


Figure 4: Mock-up of a dynamic lexicon entry for the Latin verb *libero* (to free).

5.3 Open collection

The Perseus Digital Library itself contains only a very small subset of Latin – its collections are comprised mainly of texts from the Classical era (ca. 200 BCE to 200 CE) with a handful that date beyond (Jerome's *Vulgate*, for instance, was composed in the 4th century CE). The texts that survive from this period generally form part of a fixed canon; in this respect they form a closed collection, and are similar to any number of the controlled linguistic corpora that have come into existence over the past 40 years (such the Brown corpus [24] or the British National Corpus [25]) – they provide a balanced and well-delineated set of test cases on which to conduct repeatable experiments, but their scope is extremely small compared to the volumes of texts that exist outside of them.

While the "Golden Age" of Latin literature flourished near the turn of the millennium (broadly spanning from the first century BCE through the first century CE), Latin continued to be a productive language for the ensuing two millennia. As a lingua franca, its use cut across both national boundaries and genres alike. Even into the beginnings of the modern era, it is the language not only of important scientific works such as Johannes Kepler's *Astronomia nova* (1609) or Carolus Linnaeus' *Systema Naturae* (1735) and religious treatises such as the writings of Desiderius Erasmus and Martin Luther, but of thousands of obscure and anonymous works as well. Figures 5, 6 and 7 present three such examples drawn from Google Books – a mathematical treatise written by Robert Simson in 1735 [37], a religious history written by Johann Friedrich Gruner in 1764 [18] and a philosophy dissertation written in 1836 [16]. These works represent only three such examples of the thousands of Latin works that fall outside of the controlled canon of the Classical era but can still be found in existing digital libraries.

The volume of Latin texts to be found in digital libraries is orders of magnitude larger than that found in any controlled corpus. This volume presents us with an opportunity for a dynamic lexicon. As shown above, even subpar automatic parsing is more than compensated for by the volume of texts that are parsed – the more data, the better the performance. Additionally, a wider sample of Latin from different eras and authors also lets us isolate those features in word usage that make any individual author distinct – how Caesar uses the word *do* differs from Ovid's use, and we might rightly imagine that their use together is distinct from that of a non-native author writing in the 18th century. By analyzing texts such as 18th-century mathematical treatises and 19th-century philosophical dissertations, we are able to significantly broaden the scope of our lexicon.

A digital library also differs from a controlled corpus of texts in that its collection is dynamic: while a corpus is

SECTIONUM
CONICARUM
LIBER PRIMUS.

De Parabola.

DEFINITIONES.

I. LIT recta linea AB, & punctum extra ipſam C; Fig. 1. & plano in quo ſunt recta & punctum impoſitum norma DEF, ita ut latus ipſius DE applicetur rectæ AB, alterum vero EF ſit ad eas partes ipſius AB ad quas eſt C; & extremitati F lateris EF annectatur extremitas una fili FGC ejuſdem longitudinis cum eo latere, altera vero fili extremitas in puncto C figatur; & adducatur pars fili FG ope paxilli G ad latus normæ EF, & juxta ipſum tendatur; dein moveatur normæ latus DE ſecundum rectam AB, & interea filum paxillo diſtentum ſemper lateri EF appli-

A

Figure 5: Excerpt from Robert Simson's 1735 mathematical treatise, *Sectionum Conicarum Libri V*, from Google Books.

XXIV DE ORIG. EPISCOP. EORVMQ. caet.

Deo ita ſtudeam, ut fidem certe, diligentiam atque induſtriam deſiderari in me nunquam patiar. Eum animum meum, academiae huic, eiusque Ciuium Ornatiſſimorum commodis plane deuotum, ne uerbis tantum oſtendere uidear; conabor re et opere, quantum quidem praeſentis temporis ratio permittit, declarare. Igitur, quod Deus ter Optimus Maximus felix fauſtumque eſſe iubeat, nouum munus *praelectionibus exegeticis in epiſtolam* PAVLI *Apoſtoli ad Romanos* auſpicabor. Quo quidem in labore ita uerſari ſtudebo, ut ſenſu uerborum legitime indagato, uia Auditoribus Lectiſſimis ad rerum, quae diuino illo libro continentur, intelligentiam muniatur; memor magni olim uiri, PHIL. MELANCHTHONIS effati, „non poſſe „ ſcripturam intelligi theologice, niſi intellecta ante ſit grammatice.„ Neque uero iis deero, qui *diſputandi exercitio* acuere ingenium uelint. Faxit Deus, ut his meis, tenuibus licet, conatibus nominis ipſius gloria illuſtretur! Scribebam in Regia Fridericiana,
a. d. XXI. m. Nouembris, CIↃIↃCCLXIV.

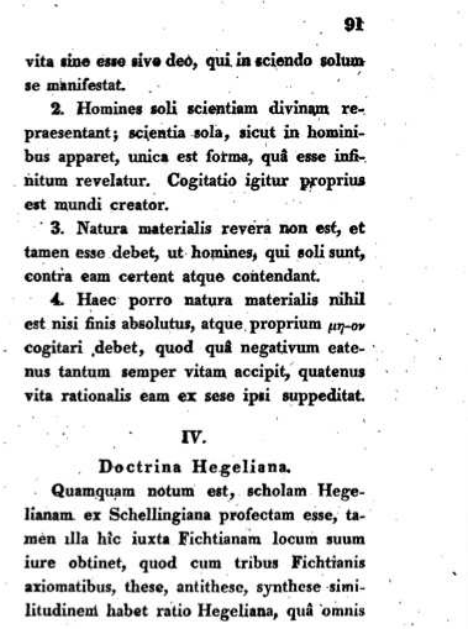
Figure 6: Excerpt from Johann Friedrich Gruner's 1764 religious history, *De origine episcoporum eorumque in ecclesia primitiva iure exercitatio*, from Google Books.

carefully curated by hand to present a balance of texts that reflect current usage, a digital library is constantly adding new texts to its collection. Without a fixed corpus on which to draw its knowledge, a lexicon that automatically parses every new text that's added to a digital collection is always up to date; by simply adding new texts – however obscure – we can gather information about how their authors use language in a way that's similar to (or radically different from) the other authors in the collection. Parsing a text and including its lexical information in a larger reference work is simply another way of contextualizing it.

6. CONCLUSION

The application of structured knowledge to much larger but unstructured collections addresses a gap left by the massive digitization efforts of groups such as Google and the Open Content Alliance (OCA). While these large projects are creating truly million book collections, the services they provide are general (e.g., key term extraction, named entity analysis, related works) and reflect the wide array of texts and languages they contain. By applying the language-specific knowledge of experts (as encoded in our treebank), we are able to create more specific services to complement these general ones already in place. In creating a dynamic lexicon built from the intersection of a 3.5 million word corpus and a 30,457 word treebank, we are highlighting the immense role than even very small structured knowledge sources can play.

In the future we plan to further investigate the knowledge and services that can arise from this interaction between small structured data and large unstructured collections (we have also just used treebanks, for instance, to inform the au-



91

vita sine esse sive deo, qui in sciendo solum se manifestat.

2. Homines soli scientiam divinam repraesentant; scientia sola, sicut in hominibus apparet, unica est forma, quâ esse infinitum revelatur. Cogitatio igitur proprius est mundi creator.

3. Natura materialis revera non est, et tamen esse debet, ut homines, qui soli sunt, contra eam certent atque contendant.

4. Haec porro natura materialis nihil est nisi finis absolutus, atque proprium μη-ον cogitari debet, quod quâ negativum eatenus tantum semper vitam accipit, quatenus vita rationalis eam ex sese ipsi suppeditat.

IV.

Doctrina Hegeliana.

Quamquam notum est, scholam Hegelianam ex Schellingiana profectam esse, tamen illa hic iuxta Fichtianam locum suum iure obtinet, quod cum tribus Fichtianis axiomatibus, these, antithese, synthese similitudinem habet ratio Hegeliana, quâ omnis

Figure 7: Excerpt from M. Freystadt's 1832 philosophy dissertation, *Philosophia cabbalistica et pantheismus*, from Google Books. Note the neologisms coined from the names of 19th-century German philosophers (*Hegeliana*, *Schellingiana* and *Fichtiana*.

tomatic discovery of allusions in texts [3]). Also important will be evaluating this lexicon in its end role as a resource within our digital library, including the opportunities that exist there for community-driven improvement. The morphological and dictionary services that currently exist within Perseus already provide the ability for users reading a text to "vote" on the morphological analysis or word sense that is appropriate given the surrounding context, with the accuracy improving with the greater number of votes cast [13]. With this sort of human interaction, we should be able to improve the overall resource by noting where our system has made errors so we can focus on correcting them automatically in the future.

Additionally, since the lexicon is built from modular technologies, it stands to benefit from any improvement in those individual services (such as morphological tagging or syntactic parsing), and since tagging and parsing accuracy are generally dependent on the size of their training corpus, we expect further improvements as our treebank grows. We are currently in the process now of adding Petronius (a late Latin prose author), and several texts of Ovid and Propertius (both Golden Age poets) as well.

The work described to date has also focussed exclusively on Latin, but the texts in the Perseus Digital Library contain a far larger collection of Greek (4.9 million words). Our goal in developing this work is to design an architecture that can just as easily be applied to both languages – all we need to extract selectional preferences for Greek is a large enough treebank with which to train a statistical parser, and we are in the initial stages of developing one now. Indeed, the technologies described above are not language or even library specific: they simply depend on a small structured knowledge source and a large textual collection. As million book libraries are proving, large textual collections in many different languages are now beginning to emerge; what still remains, however, are the knowledge sources that can only be created by practitioners in the field.

7. ACKNOWLEDGMENTS

Grants from the Digital Library Initiative Phrase 2 (IIS-9817484), the National Science Foundation (BCS-0616521) and the Andrew W. Mellon Foundation (#40700635) provided support for this work. Thanks are due also to Meg Luthin and Skylar Neil for their invaluable research assistance.

8. REFERENCES

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

[2] David Bamman and Gregory Crane. The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 33–40, Prague, 2007. Association for Computational Linguistics.

[3] David Bamman and Gregory Crane. The logic and discovery of textual allusion. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008).*

[4] David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. A collaborative model of treebank development. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT2007)*, pages 1–6, Bergen, 2007.

[5] David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. Guidelines for the syntactic annotation of Latin treebanks, version 1.3. Technical report, Tufts Digital Library, Medford, 2007.

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

[7] Roberto Busa. *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.* Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1974–1980.

[8] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 76–83, Morristown, NJ, USA, 1989. Association for Computational Linguistics.

[9] Kenneth Ward Church and Eduard H. Hovy. Good applications for crummy machine translation. *Machine Translation*, 8(4):239–258, 1993.

[10] Stephen Clark and David Weir. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206, 2002.

[11] Michael Collins, Lance Ramshaw, Jan Hajič, and Christoph Tillmann. A statistical parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 505–512, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

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

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

[14] Gregory Crane, Robert F. Chavez, Anne Mahoney, Thomas L. Milbank, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman. Drudgery and deep thought. *Communications of the ACM*, 44(5):34–40, 2001.

[15] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74, 1993.

[16] M. Freystadt. *Philosophia cabbalistica et pantheismus.* Regimontii Prussorum, Borntraeger, 1832.

[17] P. G. W. Glare, editor. *Oxford Latin Dictionary.* Oxford University Press, Oxford, 1968–1982.

[18] Johann Friedrich Gruner. *De origine episcoporum eorumque in Ecclesia primitiva iure exercitatio.* Litteris Grunertianis, Halae Magdeburgicae, 1764.

[19] Jan Hajič. Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press, 1998.

[20] Jan Hajič and Barbora Hladká. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *COLING-ACL*, pages 483–490, 1998.

[21] Jan Hajič, Jarmila. Panevová, Eva Buráňová, Zdenka Urešová, and Alla Bémová. Annotations at analytical level: Instructions for annotators (English translation by Z. Kirschner). Technical report, ÚFAL MFF UK, Prague, Czech Republic, 1999.

[22] Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. The sketch engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, 2004.

[23] Annette Klosa, Ulrich Schnörch, and Petra Storjohann. ELEXIKO – a lexical and lexicological, corpus-based hypertext information system at the Institut für deutsche Sprache, Mannheim. In *Proceedings of the 12th Euralex International Congress*, 2006.

[24] H. Kučera and W. N. Francis. *Computational Analysis of Present-Day American English.* Brown University Press, Providence, RI, 1967.

[25] Geoffrey Leech, Roger Garside, and Michael Bryant. CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics*, pages 622–628, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[26] Charles T. Lewis, editor. *An Elementary Latin Dictionary.* Clarendon Press, Oxford, 1891.

[27] Charles T. Lewis and Charles Short, editors. *A Latin Dictionary.* Clarendon Press, Oxford, 1879.

[28] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1994.

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

[30] Joakim Nivre. Constraints on non-projective dependency parsing. In *EACL*. The Association for Computer Linguistics, 2006.

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

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

[33] Harm Pinkster. *Latin Syntax and Semantics.* Routledge, London, 1990.

[34] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey, 1996.

[35] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 104–111, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[36] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[37] Roberto Simson. *Sectionum Conicarum Libri V.* T. and W. Ruddimannos, Edinburgh, 1735.

[38] John M. Sinclair, editor. *Looking Up: an account of the COBUILD project in lexical computing.* Collins, 1987.