The Latin Dependency Treebank in a Cultural Heritage 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

This paper describes the mutually beneficial relationship between a cultural heritage digital library and a historical treebank: an established digital library can provide the resources and structure necessary for efficiently building a treebank, while a treebank, as a language resource, is a valuable tool for audiences traditionally served by such libraries.

1 Introduction

The composition of historical treebanks is fundamentally different from that of modern ones. While modern treebanks are generally comprised of newspaper articles,[1] historical treebanks are built from texts that have been the focus of study for centuries, if not millennia. The Penn-Helsinki Parsed Corpus of Middle English (Kroch and Taylor, 2000), for example, includes Chaucer's 14th-century *Parson's Tale*, while the York Poetry Corpus (Pintzuk and Leendert, 2001) includes the entire text of *Beowulf*. The scholarship that has attended these texts since their writing has produced a wealth of contextual materials, including commentaries, translations, and linguistic resources.

For the past twenty years, the Perseus digital library (Crane, 1987; Crane et al., 2001) has collected materials of this sort to create an open reading environment for the study of Classical texts. This environment presents the Greek or Latin source text and contextualizes it with secondary publications (e.g., translations, commentaries, references in dictionaries), along with a morphological analysis of every word in the text and variant manuscript readings as well (when available).

We have recently begun work on syntactically annotating the texts in our collection to create a Latin Dependency Treebank. In the course of developing this treebank, the resources already invested in the digital library have been crucial: the digital library provides a modular structure on which to build additional services, contains a large corpus of Classical source texts, and provides a wealth of contextual information for annotators who are non-native speakers of the language.

In this the digital library has had a profound impact on the creation of our treebank, but the influence goes both ways. The digital library is a heavily trafficked website with a wide range of users, including professional scholars, students and hobbyists. By incorporating the treebank as a language resource into this digital library, we have the potential to introduce a fundamental NLP tool to an audience outside the traditional disciplines of computer science or computational linguistics that would normally use it. Students of the language can profit from the syntactic information encoded in a treebank, while traditional scholars can benefit from the textual searching it makes possible as well.

[1] To name just three, the Penn Treebank (Marcus et al., 1994) is comprised of texts from the *Wall Street Journal*; the German TIGER Treebank (Brants et al., 2002) is built from texts taken from the *Frankfurter Rundschau*; and the Prague Dependency Treebank (Hajič, 1998) includes articles from several daily newspapers (*Lidové noviny* and *Mladá fronta Dnes*), a business magazine (*Českomoravský Profit*) and a scientific journal (*Vesmír*).

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 33–40,
Prague, 28 June 2007.

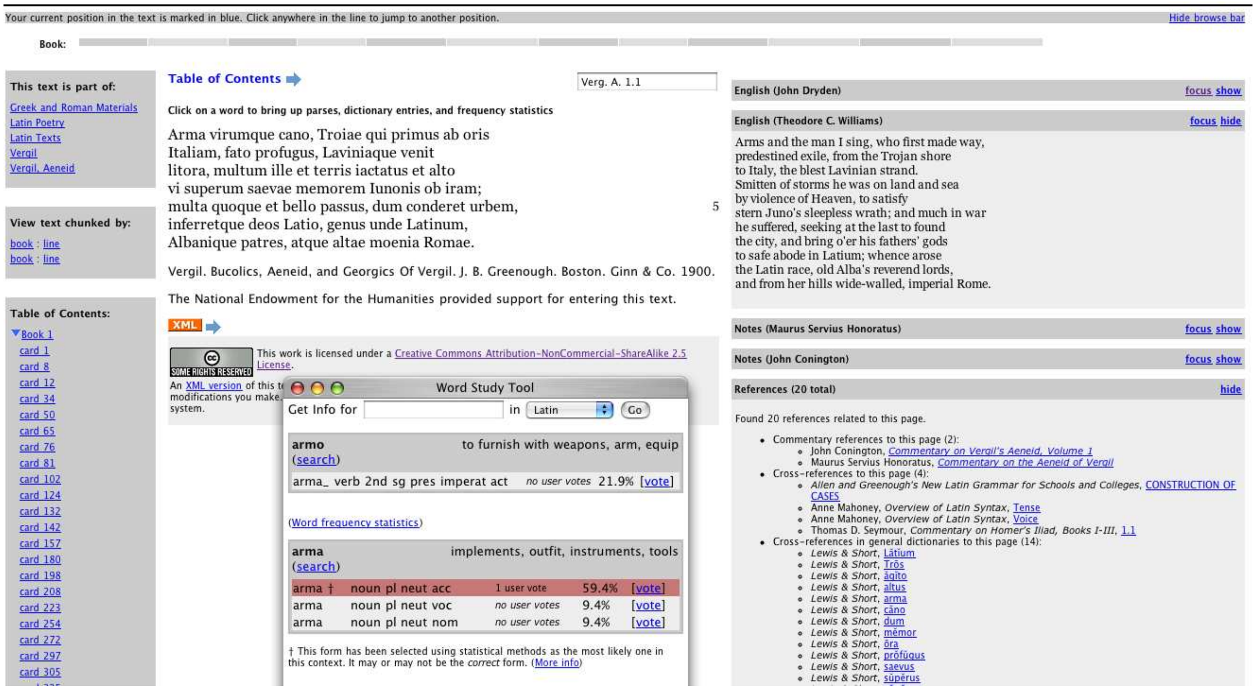

Figure 1: A screenshot of Vergil's *Aeneid* from the Perseus digital library.

2 The Perseus Digital Library

Figure 1 shows a screenshot from our 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 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 for a word given its surrounding context, and users can also vote for the form they think is correct.[2]

3 Latin Dependency Treebank

Now in version 1.3, the Latin Dependency Treebank is comprised of excerpts from four texts: Cicero's *Oratio in Catilinam*, Caesar's *Commentarii de Bello Gallico*, Vergil's *Aeneid* and Jerome's *Vulgate*.

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) (Hajič, 1998) for Czech (another non-projective language) while tailoring it for Latin via

[2]These user contributions have the potential to significantly improve the morphological tagging of these texts: any single user vote assigns the correct morphological analysis to a word 89% of the time, while the recommender system does so with an accuracy of 76% (Crane et al., 2006).

Date	Author	Words
63 BCE	Cicero	1,189
51 BCE	Caesar	1,486
19 BCE	Vergil	2,647
405 CE	Jerome	8,382
	Total:	13,683

Table 1: Treebank composition by author.

the grammar of Pinkster (1990).[3]

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 and its morphological code. We plan to release the treebank incrementally with each new major textual addition (so that version 1.4, for instance, will include the treebank of 1.3 plus Sallust's *Bellum Catilinae*, the text currently in production).

4 The Influence of a Digital Library

A cultural heritage digital library has provided a fertile ground for our historical treebank in two fundamental ways: by providing a structure on which to build new services and by providing reading support to expedite the process of annotation.

4.1 Structure

By anchoring the treebank in a cultural heritage digital library, we are able to take advantage of a structured reading environment with canonical standards for the presentation of text and a large body of digitized resources, which include XML source texts, morphological analyzers, machine-readable dictionaries, and an online user interface.

Texts. Our digital library contains 3.4 million words of Latin source texts (along with 4.9 million words of Greek). The texts are all public-domain materials that have been scanned, OCR'd and formatted into TEI-compliant XML. The value of this prior labor is twofold: most immediately, the existence of clean, digital editions of these texts has saved us a considerable amount of time and resources, as we would otherwise have to create them before annotating them syntactically; but their encoding as repurposeable XML documents in a larger library also allows us to refer to them under standardized citations. The passage of Vergil displayed in Figure 1 is not simply a string of unstructured text; it is a subdocument (*Book=1:card=1*) that is itself part of a larger document object (*Perseus:text:1999.02.0055*), with sisters (*Book=1:card=8*) and children of its own (e.g., *line=4*). This XML structure allows us to situate any given treebank sentence within its larger context.

Morphological Analysis. As a highly inflected language, Latin has an intricate morphological system, in which a full morphological analysis is the product of nine features: part of speech, person, number, tense, mood, voice, gender, case and degree. Our digital library has included a morphological analyzer from its beginning. This resource maps an inflected form of a word (such as *arma* above) to all of the possible analyses for all of the dictionary entries associated with it. In addition to providing a common morphological standard, this mapping greatly helps to constrain the problem of morphological tagging (selecting the correct form from all possible forms), since a statistical tagger only needs to consider the morphological analyses licensed by the inflection rather than all possible combinations.

User interface. The user interface of our library is designed to be modular, since different texts have different contextual resources associated with them (while some have translations, others may have commentaries). This modularity allows us to easily introduce new features, since the underlying architecture of the page doesn't change – a new feature can simply be added.

Figure 2 presents a screenshot of the digital library with an annotation tool built into the interface. In the widget on the right, the source text in view (the first chunk of Tacitus' *Annales*) has been automatically segmented into sentences; an annotator can click on any sentence to assign it a syntactic annotation. Here the user has clicked on the first sentence (*Vrbem Romam a principio reges habuere*); this action brings up an annotation screen in which a partial automatic parse is provided, along with the most likely morphological analysis for each word. The annotator can then correct this automatic output

[3] We are also collaborating with other Latin treebanks (notably the Index Thomisticus on the works of Thomas Aquinas) to create a common set of annotation guidelines to be used as a standard for Latin of any period (Bamman et al., 2007).

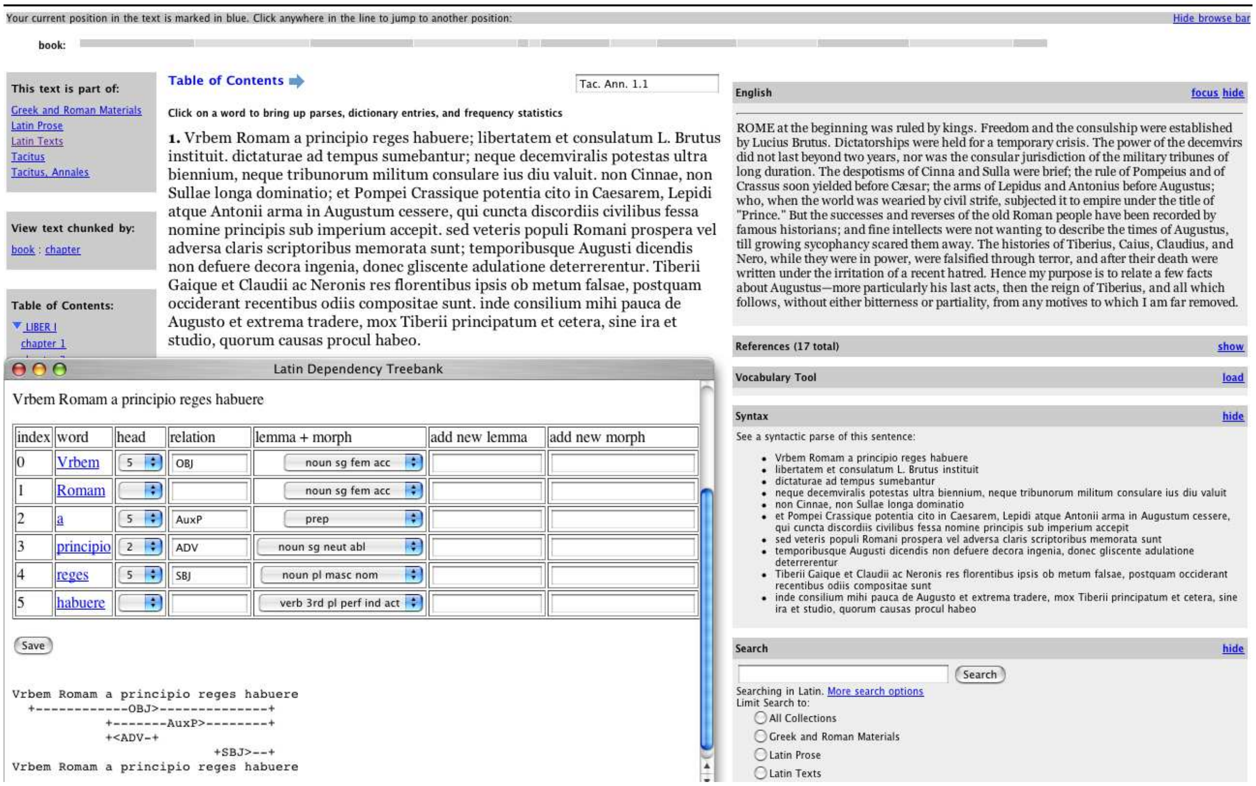

Figure 2: A screenshot of Tacitus' *Annales* from the Perseus digital library.

and move on to the next segmented sentence, with all of the contextual resources still in view.

4.2 Reading support

Modern treebanks also differ from historical ones in the fluency of their annotators. The efficient annotation of historical languages is hindered by the fact that no native speakers exist, and this is especially true of Latin, a difficult language with a high degree of non-projectivity. While the Penn Treebank can report a productivity rate of between 750 and 1000 words per hour for their annotators after four months of training (Taylor et al., 2003) and the Penn Chinese treebank can report a rate of 240-480 words per hour (Chiou et al., 2001), our annotation speeds are significantly slower, ranging from 90 words per hour to 281. Our best approach for Latin is to develop strategies that can speed up the annotation process, and here the resources found in a digital library are crucial. There are three varieties of contextual resources in our digital library that aid in the understanding of a text: translations, commentaries, and dictionaries. These resources shed light on a text, from the level of sentences to that of individual words.

Translations. Translations provide reading support on a large scale: while loose translations may not be able to inform readers about the meaning and syntactic role of any single word, they do provide a broad description of the action taking place, and this can often help to establish the semantic structure of the sentence – who did what to whom, and how. In a language with a free word order (and with poetry especially), this kind of high-level structure can be important for establishing a quick initial understanding of the sentence before narrowing down to individual syntactic roles.

Commentaries. Classical commentaries provide information about the specific use of individual words, often noting morphological information (such as case) for ambiguous words or giving explanatory information for unusual structures. This information often comes at crucial decision points in the annotation process, and represents judgments by authorities in the field with expertise in that particular text.

[4] **Vi superum** expresses the general agency, like 'fato profugus,' though Juno was his only personal enemy. Gossrau's fancy that 'vi superum' = βίᾳ θεῶν, 'in spite of heaven,' has no authority. For 'memorem iram' comp. **Livy 9. 29.** "Traditur censorem etiam Appium memori Deum ira post aliquot annos luminibus captum." So **Aesch. Ag. 155.** "μνάμων μῆνις". 'Ob iram,' below, v. 251, 'to sate the wrath.'

[5] **Passus,** constructed like 'iactatus.' 'Quoque' and 'et' of course form a pleonasm, though the former appears to be connected with 'multa,' and the latter with 'bello.' 'Dum conderet' like "dum fugeret," **G. 4. 457**, where see note. Here we might render 'in the struggle to build his city.' So Hom. Od. 1. 4. foll., πολλὰ πάθεν . . ἀρνύμενος κ.τ.λ. The clause belongs to 'multa bello passus,' rather than to 'iactatus.'

Figure 3: An excerpt from Conington's commentary on Vergil's *Aeneid* (Conington, 1876), here referring to Book 1, lines 4 and 5.

Machine-Readable Dictionaries. In addition to providing lists of stems for morphological analyzers, machine-readable dictionaries also provide valuable reading support for the process of lemma selection. Every available morphological analysis for a word is paired with the word stem (a lemma) from which it is derived, but analyses are often ambiguous between different lemmas. The extremely common form *est*, for example, is a third person singular present indicative active verb, but can be inflected from two different lemmas: the verb *sum* (to be) and the verb *edo* (to eat). In this case, we can use the text already tagged to suggest a more probable form (*sum* appears much more frequently and is therefore the likelier candidate), but in less dominant cases, we can use the dictionary: since the word stems involved in morphological analysis have been derived from the dictionary lemmas, we can map each analysis to a dictionary definition, so that, for instance, if an annotator is unfamiliar with the distinction between the lemmas *occido1* (to strike down) and *occido2* (to fall), their respective definitions can clarify it.

Machine-readable dictionaries, however, are also a valuable annotation resource in that they often provide exemplary syntactic information as part of their definitions. Consider, for example, the following line from Book 6, line 2 of Vergil's *Aeneid*: *et tandem Euboicis Cumarum adlabitur oris* ("and at last it glides to the Euboean shores of Cumae"). The noun *oris* (shores) here is technically ambiguous, and can be derived from a single lemma (*ora*) as a noun in either the dative or ablative case. The dictionary definition of *allabor* (to glide), however, disambiguates this for us, since it notes that the verb is often constructed with either the dative or the accusative case.

al-lābor (**adl-**), lapsus, 3, v. dep.,

I. *to glide to* or *toward* something, *to come to, to fly, fall, flow, slide*, and the like; constr. with dat. or acc. (**poet.**—oftenest in Verg.— "or in more elevated prose): viro adlapsa sagitta est," **Verg. A. 12, 319**: "fama adlabitur aurīs," *id. ib. 9, 474*: Curetum adlabimur oris, *we land upon*, etc., *id. ib. 3, 131*; cf. *id. ib. 3, 569*: "mare crescenti adlabitur aestu," *rolls up with increasing wave, id. ib. 10, 292*: "adlapsus genibus," *falling down at his knees, Sen. Hippol. 666*.—In prose: umor adlapsus extrinsecus, * *Cic. Div. 2, 27, 58*: "angues duo ex occulto adlapsi," **Liv. 25, 16**.

Figure 4: Definition of *allabor* (the dictionary entry for *adlabitur*) from Lewis and Short (1879).

Every word in our digital library is linked to a list of its possible morphological analyses, and each of those analyses is linked to its respective dictionary entry. The place of a treebank in a digital library allows for this tight level of integration.

5 The Impact of a Historical Treebank

The traffic in our library currently exceeds 10 million page views by 400,000 distinct users per month (as approximated by unique IP addresses). These users are not computational linguists or computer scientists who would typically make use of a treebank; they are a mix of Classical scholars, students, and amateurs. These different audiences have equally different uses for a large corpus of syntactically annotated sentences: for one group it can provide additional reading support, and for the other a scholarly resource to be queried.

5.1 Treebank as Reading Support

Our digital library is predominantly a reading environment: source texts in Greek and Latin are presented with attendant materials to help facilitate their understanding. The broadest of these materials are translations, which present sentence-level equivalents of the original; commentaries provide a more detailed analysis of individual words and phrases. A treebank has the potential to be a valuable contextual resource by providing syntactic information for every word in a sentence, not simply those chosen by a commentator for discussion.

5.2 Treebank as a Scholarly Resource

For Classical scholars, a treebank can also be used as a scholarly resource. Not all Classicists are programmers, however, and many of those who would like to use such a resource would profit little from an XML source file. We have already released version 1.3 of the Latin Dependency Treebank in its XML source, but we also plan to incorporate it into the digital library as an object to be queried. This will yield a powerful range of search options, including lemmatized and morpho-syntactic searching, and will be especially valuable for research involving lexicography and semantic classification.

Lemmatized searching. The ability to conduct a lemma-based textual search has long been a desideratum in Classics,[4] where any given Latin word form has 3.1 possible analyses on average.[5] Locating all inflections of *edo* (to eat) in the texts of Caesar, for example, would involve two things:

1. Searching for all possible inflections of the root word. This amounts to 202 different word forms attested in our texts (including compounds with enclitics).

2. Eliminating all results that are homonyms derived from a different lemma. Since several inflections of *edo* are homonyms with inflections of the far more common *sum* (to be), many of the found results will be false positives and have to be discarded.

This is a laborious process and, as such, is rarely undertaken by Classical scholars: the lack of such a resource has constrained the set of questions we can ask about a text. Since a treebank encodes each word's lemma in addition to its morphological and syntactic analysis, this information is now free for the taking.

Morpho-syntactic searching. A treebank's major contribution to scholarship is that it encodes the syntax of a sentence, along with a morphological analysis of each word. These two together can be combined into elaborate searches. Treebanks allow scholars to find all instances of any particular construction. For example:

- When the conjunction *cum* is the head of a subordinate clause whose verb is indicative, it is often recognized as a temporal clause, qualifying the time of the main clause's action;

- When that verb is subjunctive, however, the clause retains a different meaning, as either circumstantial, causal, or adversative.

These different clause types can be found by querying the treebank: in the first case, by searching for indicative verbs that syntactically depend on *cum*; in the second, for subjunctive verbs that depend on it. In version 1.3 of the Latin Dependency Treebank, *cum* is the head of a subordinate clause 38 times: in 7 of these clauses an indicative verb depends on it, while in 31 of them a subjunctive one does. This type of searching allows us to gather statistical data while also locating all instances for further qualitative analysis.[6]

Lexicography. Searching for a combination of lemma and morpho-syntactic information can yield powerful results, which we can illustrate with a question from Latin lexicography: how does the meaning of a word change across authors and over time? If we take a single verb – *libero* (to free, liberate) – we can chart its use in various authors by asking a more specific question: what do different Latin authors want to be liberated from? We can imagine that an orator of the republic has little need to speak of liberation from eternal death, while an apostolic father is just as unlikely to speak of being freed from another's monetary debt.

[4] Both the Perseus Project and the Thesaurus Linguae Graecae (http://www.tlg.uci.edu) allow users to search for all inflected forms of a lemma in their texts, but neither filters results that are homonyms derived from different lemmas.

[5] Based on the average number of lemma + morphology combinations for all unique word tokens in our 3.4 million word corpus. The word form *amor*, for example, has 3 analyses: as a first-person singular present indicative passive verb derived from the lemma *amo* (to love) and as either a nominative or vocative masculine singular noun derived from *amor* (love).

[6] For the importance of a treebank in expediting morpho-syntactic research in Latin rhetoric and historical linguistics, see Bamman and Crane (2006).

We can answer this more general question by transforming it into a syntactic one: what are the most common complements of the lemma *libero* that are expressed in oblique cases (e.g., ablative, genitive, etc.) or as prepositional phrases? In a small test of 100 instances of the lemma in Cicero and Jerome, we find an interesting answer, presented in Table 2.

Cicero		Jerome	
periculo	14	manu	22
metu	8	morte	3
cura	6	ore	3
aere	3	latronibus	2
scelere	3	inimico	2
suspicione	3	bello	2

Table 2: Count of objects *liberated from* in Cicero and Jerome that occur with frequency greater than 1 in a corpus of 100 sentences from each author containing any inflected form of the verb *libero*.

The most common entities that Cicero speaks of being liberated from clearly reflect the cares of an orator of the republic: *periculo* (danger), *metu* (fear), *cura* (care), and *aere* (debt). Jerome, however, uses *libero* to speak of liberation from a very different set of things: his actors speak of deliverance from *manu* (e.g., the hand of the Egyptians), from *ore* (e.g., the mouth of the lion) and from *morte* (death). A treebank encoded with lemma and morpho-syntactic information lets us quantify these typical arguments and thereby identify the use of the word at any given time.

Named entity labeling. Our treebank's place in a digital library also means that complex searches can draw on the resources that already lie therein. Two of our major reference works include Smith's *Dictionary of Greek and Roman Geography* (1854), which contains 11,564 place names, and Smith's *Dictionary of Greek and Roman Biography and Mythology* (1873), which contains 20,336 personal names. By mapping the lemmas in our treebank to the entries in these dictionaries, we can determine each lemma's broad semantic class. After supplementing the Classical Dictionary with names from the Vulgate, we find that the most common people in the treebank are *Iesus*, *Aeneas*, *Caesar*, *Catilina*, *Satanas*, *Sibylla*, *Phoebus*, *Misenus* and *Iohannes*; the most common place names are *Gallia*, *Babylon*, *Troia*, *Hierusalem*, *Avernus* and *Sardis*.

One use of such classification is to search for verbs that are typically found with sentient agents. We can find this by simply searching the treebank for all active verbs with subjects known to be people (i.e., subjects whose lemmas can be mapped to an entry in Smith's *Dictionary*). An excerpt of the list that results is given in Table 3.

mitto	to send
iubeo	to order
duco	to lead
impono	to place
amo	to love
incipio	to begin
condo	to hide

Table 3: Common verbs with people as subjects in the Latin Dependency Treebank 1.3.

Aside from its intrinsic value of providing a catalogue of such verbs, a list like this is also useful for classifying common nouns: if a verb is frequently found with a person as its subject, all of its subjects in general will likely be sentient as well. Table 4 presents a complete list of subjects of the active voice of the verb *mitto* (to send) as attested in our treebank.

angelus	angel
Caesar	Caesar
deus	God
diabolus	devil
Remi	Gallic tribe
serpens	serpent
ficus	fig tree

Table 4: Subjects of active *mitto* in the Latin Dependency Treebank 1.3.

Only two of these subjects are proper names (*Caesar* and *Remi*) that can be found in Smith's *Dictionary*, but almost all of these nouns clearly belong to the same semantic class – *angelus*, *deus*, *diabolus* and *serpens* (at least in this text) are entities with cognition.

Inducing semantic relationships of this sort is the typical domain of clustering techniques such as latent semantic analysis (Deerwester et al., 1990), but those methods generally work best on large corpora. By embedding this syntactic resource in a digital library and linking it to external resources such as reference works, we can find similar semantic relationships with a much smaller corpus.

6 Conclusion

Treebanks already fill a niche in the NLP community by providing valuable datasets for automatic processes such as parsing and grammar induction. Their utility, however, does not end there. The linguistic information that treebanks encode is of value to a wide range of potential users, including professional scholars, students and amateurs, and we must encourage the use of these resources by making them available to such a diverse community. The digital library described in this paper has proved to be crucial for the development and deployment of our treebank: since the natural intuitions of native speakers are hard to come by for historical languagues, it is all the more important to leverage the cultural heritage resources we already have.

7 Acknowledgments

Grants from the Digital Library Initiative Phrase 2 (IIS-9817484) and the National Science Foundation (BCS-0616521) provided support for this work.

References

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

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

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–41, Sozopol.

Fu-Dong Chiou, David Chiang, and Martha Palmer. 2001. Facilitating treebank annotation using a statistical parser. In *Proceedings of the First International Conference on Human Language Technology Research HLT '01*, pages 1–4.

John Conington, editor. 1876. *P. Vergili Maronis Opera. The Works of Virgil, with Commentary*. Whittaker and Co, London.

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

Gregory Crane, David Bamman, Lisa Cerrato, Alison Jones, David M. Mimno, Adrian Packel, David Sculley, and Gabriel Weaver. 2006. Beyond digital incunabula: Modeling the next generation of digital libraries. In *ECDL 2006*, pages 353–366.

Gregory Crane. 1987. 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.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Jan Hajič. 1998. 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.

A. Kroch and A. Taylor. 2000. Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/ppcme2-release-2/.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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

Susan Pintzuk and Plug Leendert. 2001. York-Helsinki Parsed Corpus of Old English Poetry.

William Smith. 1854. *A Dictionary of Greek and Roman Geography*. Walton and Maberly, London.

William Smith. 1873. *A Dictionary of Greek and Roman Biography and Mythology*. Spottiswoode, London.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: An overview. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 5–22. Kluwer Academic Publishers.