The Ancient Greek and Latin Dependency Treebanks

David Bamman and Gregory Crane

Abstract This paper describes the development, composition, and several uses of the Ancient Greek and Latin Dependency Treebanks, large collections of Classical texts in which the syntactic, morphological and lexical information for each word is made explicit. To date, over 200 individuals from around the world have collaborated to annotate over 350,000 words, including the entirety of Homer's *Iliad* and *Odyssey*, Sophocles' *Ajax*, all of the extant works of Hesiod and Aeschylus, and selections from Caesar, Cicero, Jerome, Ovid, Petronius, Propertius, Sallust and Vergil. While perhaps the most straightforward value of such an annotated corpus for Classical philology is the morphosyntactic searching it makes possible, it also enables a large number of downstream tasks as well, such as inducing the syntactic behavior of lexemes and automatically identifying similar passages between texts.

1 Introduction

The definitive Classical reference grammars of the 19th and 20th centuries, such as Herbert Smyth's *Greek Grammar* [1] and Raphael Kühner's *Ausführliche Grammatik der lateinischen Sprache* [2], are monuments of scholarship that distill lifetimes of reading and linguistic observation into succinct aphorisms such as the following:

David Bamman
Perseus Project, Tufts University e-mail: david.bamman@tufts.edu

Gregory Crane
Perseus Project, Tufts University e-mail: gregory.crane@tufts.edu

* Preprint of a chapter to appear in: Caroline Sporleder, Antal van den Bosch and Kalliopi Zervanou (eds.), Language Technology for Cultural Heritage, ser. Foundations of Human Language Processing and Technology (Springer, 2011).

> "Apodotic δέ is very common in Homer and Herodotus, not rare in Attic poetry, but infrequent in Attic prose." (Smyth 2837).

On occasion these works offer a window into the traditional philological practice that lies behind them, as in Kühner's comparison of the Latin *accusativus cum infinitivo* construction with subordinate clauses containing an overt complementizer (e.g., *quod*):

> "So hat nach meiner Zählung bei *doleo* 57 Stellen mit *Acc. c. Inf.* gegen 4 *quod*, bei *miror* 110 gegen 8, bei *glorior* 19 gegen 2, bei *queror* 71 gegen 15, bei *gaudeo* 84 gegen 9 usw." (1914:77)[1]

In its most basic form, classical philology of this sort is by definition a data-driven science: it relies on a fixed dataset (the extant corpus of Ancient Greek and Latin) and builds larger arguments by the simple act of counting. Kühner here publishes his tally of ACI vs. *quod*-clauses in order to advance the argument that the ACI is more frequent in indirect discourse than subordinate clauses are, and one can assume that either such an explicit tally or an implicit one (collected over a lifetime of reading) is what drives Smyth's observations on relative frequency as well.

Where classical philology has so far diverged from data-driven science, however, is in its reliance on the authority of the editor rather than on the data itself. As much as the judgment of Kühner and Smyth may far exceed our own, the cornerstone of the scientific method is the reproducibility of experiments, and as P. Cuzzolin [3] notes about this very passage of Kühner:

> "...it is difficult to say what he meant by the word "Stelle" and impossible to say which texts his counting is based upon."

Ideally, what we want to see is the evidence that drives such linguistic observations – not simply knowing that the ACI is used in some unknown sample of 57 sentences containing *doleo*, but exactly which sentences those are, which textual editions they come from, and how that small sample relates to the corpus at large (if only to measure its significance). While such a work may not have been possible in the print culture of the past, we are at a transformative moment now where we can begin leveraging the scientific method in the service of classical philology.

2 Treebanks

Our work in developing treebanks for Ancient Greek and Latin are our own efforts to help move classical philology into this scientific space. A treebank is a large collection of sentences in which the syntactic relation for every word is made explicit – where a human has encoded an interpretation of the sentence in the form of a linguistic annotation. While much of the research and labor in treebanks over the past

[1] "And so, by my count, with *doleo* there are 57 sentences with the accusative + infinitive against 4 with *quod*, with *miror* 110 against 8, with *glorior* 19 against 2, with *queror* 71 against 15, with *gaudeo* 84 against 9 etc."

twenty years has focused on modern languages such as English [4], Arabic [5, 6] and Czech [7], recent scholarship has seen the rise of a number of treebanks for historical languages as well, including Middle English [8], Early Modern English [9], Old English [10], Medieval Portuguese [11], Ugaritic [12], Latin [13, 14] and several Indo-European translations of the New Testament [15].

Treebanks have been annotated under a variety of grammatical frameworks, with tthe most dominant being the phrase structure grammar employed by the Penn Treebank of English [4] and the dependency grammar in use by the Prague Dependency Treebank of Czech [7]. The defining feature of dependency grammar (Mel'cuk [16], Sgall [17], Tesniere [18]) that distinguishes it from constituent-based formalisms is the absence of non-terminal nodes common in $\overline{\text{X}}$ theory (Chomsky [19]) – while individual words under a phrase structure grammar form part of abstract structures such as NP (noun phrase) and VP (verb phrase), in a dependency grammar they are directly linked to each other via asymmetrical dependency relationships, with each word being the child (or "dependent") of exactly one other word. Dependency grammars deal especially well with languages involving relatively free word order (which in a transformational grammar would otherwise involve a high degree of scrambling), and this flexibility led us to adopt it as the annotation style for our treebanks as well.

Figure 1 represents one such example of a dependency annotation from an elegy of Propertius.

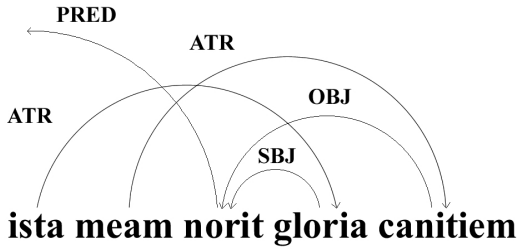


Fig. 1 Dependency graph of the treebank annotation for *ista meam norit gloria canitiem* ("that glory will know my old age"), Propertius I.8.46. Arcs are directed from children to their parents.

Classical texts have been a focused object of study for over two thousand years, with schoolchildren and tenured professors alike scrutinizing their every word; a treebank is simply an effort to capture such analysis in a quantified format that can provide a measurable dataset for reproducing linguistic experiments while also driving a new generation of computational analysis.

3 Building the Ancient Greek and Latin Dependency Treebanks

The Ancient Greek and Latin Dependency Treebanks are the work of over 200 researchers from around the world. The backgrounds of the annotators range from advanced undergraduate students to recent PhDs and professors, with the majority being students in graduate programs in Classics. Annotators undergo an initial training period in which they learn the application of dependency grammar to Greek and Latin texts (as encoded in the guidelines for syntactic annotation[2]) and are actively engaged in new learning afterwards by means of an online forum in which they can ask questions of each other and of project editors; this allows them to be kept current on the most up-to-date codifications to the annotation guidelines while also helping bring new annotators up to speed. In the "standard" model of production, every sentence is annotated by two independent annotators and the differences are then reconciled by a third in order to filter out the biases (and errors) of any single individual.[3] This reconciliation (or "secondary" annotation as it is encoded in the XML release) is undertaken by a more experienced annotator/editor, typically a PhD with specialization in the particular subject area (such as Homer).

As figure 2 illustrates, all annotations are publicly released with the usernames of the primary and secondary annotators (which are then also associated with real names and institutional affiliations). By publicly acknowledging authorship, we are making our first steps toward an ownership model for annotation and hope to provide a means for students, both graduate and undergraduate alike, to engage in the act of scholarly research and in the production of scientific data that can be useful to the wider Classics community.

```
<sentence id="341" document_id="Perseus:text:1999.02.0066" subdoc="book=1:poem=8b" span="ista0:canitiem0">
  <primary>alexlessie</primary>
  <primary>sneil01</primary>
  <secondary>millermo</secondary>
  <word id="1" form="ista" lemma="iste1" postag="p-s---fn-" head="4" relation="ATR"/>
  <word id="2" form="meam" lemma="meus1" postag="a-s---fa-" head="5" relation="ATR"/>
  <word id="3" form="norit" lemma="nosco1" postag="v3srsa---" head="0" relation="PRED"/>
  <word id="4" form="gloria" lemma="gloria1" postag="n-s---fn-" head="3" relation="SBJ"/>
  <word id="5" form="canitiem" lemma="canities1" postag="n-s---fa-" head="3" relation="OBJ"/>
</sentence>
```

Fig. 2 XML fragment from the Latin Dependency Treebank (Propertius I.8.46).

While the goal of the standard method of production is to filter out individual bias, we also want to provide the ability for scholars to publish a record of their own unique interpretation of the text – an interpretation that stands as theirs alone. In this spirit, we have also developed a "scholarly" method of annotation [24] and have published part of our treebank under this model. By publicly releasing data with citable attributions of ownership in this way, we hope to provide a fixed core

[2] The annotation guidelines are available in English (for Greek [20] and Latin [21]) and in Spanish (for Greek [22] and Latin [23]).

[3] The interannotator agreement rate for the Ancient Greek Dependency Treebank measures 87.4% for attachment agreement, 85.3% for label agreement, and 80.6% for labeled attachment [24].

around which other interpretations (by other scholars) can then be layered. Literary works very often license multiple valid syntactic annotations and, for ancient texts especially, scholarly disagreement can be found not only on the level of the correct syntactic parse, but also on the form of the text itself (since we do not have the original text in the author's own hand, but rather a series of copies by medieval scribes). Providing a quantified record of how these multiple interpretations differ can only help drive future research.

4 Ancient Greek Dependency Treebank

The current version of the Ancient Greek Dependency Treebank (v. 1.2) includes the entirety of Homer's *Iliad* and *Odyssey*, Sophocles' *Ajax*, and all of the works of Hesiod and Aeschylus for a total of 309,096 words, as distributed in table 1.

Method	Author	Work	Sentences	Words
Standard	Hesiod	Shield of Heracles	255	3,834
		Theogony	438	8,106
		Works and Days	491	6,941
	Homer	Iliad	8,415	128,102
		Odyssey	6,760	104,467
Scholarly	Aeschylus	Agamemnon	814	9,806
		Eumenides	526	6,380
		Libation Bearers	572	6,563
		Persians	478	6,223
		Prometheus Bound	589	7,045
		Seven Against Thebes	518	6,206
		Suppliants	529	5,949
	Sophocles	Ajax	785	9,474
		Total:	**21,170**	**309,096**

Table 1 Composition of the Ancient Greek Dependency Treebank (version 1.2).

In addition to the index of its syntactic head and the type of relation to it, each word is also annotated 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). All of the files have been freely released under a Creative Commons license.[4]

For the works of Homer and Hesiod, we have followed the standard production method of soliciting annotations from two different annotators and then reconciling the differences between them. Sophocles and Aeschylus, whose textual traditions are much more fragmentary, have presented an ideal case for annotation as scholarly treebanks.

[4] All treebank data can be found at: http://nlp.perseus.tufts.edu/syntax/treebank/.

5 Latin Dependency Treebank

Currently in version 1.5, the Latin Dependency Treebank is comprised of 53,143 words from eight texts, as shown in table 2. As with the Ancient Greek Dependency Treebank, each word is also annotated with the lemma from which it is inflected and its morphological code. All of the texts in this release have been annotated under the standard model of production, with an editor reconciling the differences between two independent annotations.

Method	Author	Work	Sentences	Words
Standard	Caesar	B.G. (Book 2 selections)	71	1,488
	Cicero	In Catilinam 1.1-2.11	327	6,229
	Jerome	Vulgate: Apocalypse	405	8,382
	Ovid	Metamorphoses: Book I	316	4,789
	Petronius	Satyricon 26-78 (Cena Trimalchionis)	1,114	12,474
	Propertius	Elegies: Book I	361	4,857
	Sallust	Catilina	701	12,311
	Vergil	Aeneid (Book 6 selections)	178	2,613
		Total	**3,473**	**53,143**

Table 2 Composition of the Latin Dependency Treebank (version 1.5).

6 The Influence of a Digital Library

The composition of historical treebanks is fundamentally different from that of modern ones. On the one hand, the efficient annotation of Ancient Greek and Latin is hindered by the fact that no native speakers exist and the texts that we have available are typically highly stylized in nature. On the other hand, however, while modern treebanks are generally comprised of newspaper articles,[5] the texts that make up historical treebanks have generally been the focus of scholarly attention for centuries, if not millennia. The Penn-Helsinki Parsed Corpus of Middle English [8], for example, includes Chaucer's 14th-century *Parson's Tale*, while the York Poetry Corpus [26] 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.

In building a workflow for creating treebanks for Ancient Greek and Latin, we attempt to provide as much of this kind of contextualizing information for each sen-

[5] To name just three, the Penn Treebank [4] is comprised of texts from the *Wall Street Journal*; the German TIGER Treebank [25] is built from texts taken from the *Frankfurter Rundschau*; and the Prague Dependency Treebank [7] includes articles from several daily newspapers (*Lidové noviny* and *Mladá fronta Dnes*), a business magazine (*Českomoravský Profit*) and a scientific journal (*Vesmír*).

tence as possible, and embedding our annotation environment within the Perseus Digital Library has been crucial in this respect. Established in 1987 in order to construct a large, heterogeneous collection of textual and visual materials on the archaic and classical Greek world, Perseus today serves as a laboratory for digital library technologies and is also widely used by students, academics and others to access information on the Greco-Roman world [27, 28, 29].

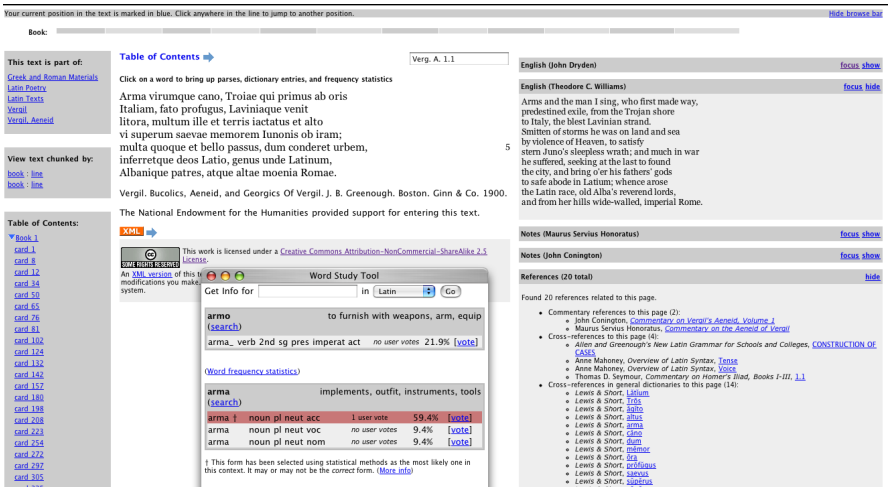


Fig. 3 A screenshot of Vergil's *Aeneid* from the Perseus digital library.

Figure 3 shows a screenshot from this 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.[6]

A cultural heritage digital library has provided a fertile ground for our historical treebanks 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.

6.1 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.

6.1.1 Texts

The Perseus 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 in processing them for annotation, 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 3 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.

6.1.2 Morphological Analysis

As highly inflected languages, Ancient Greek and Latin have 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

[6] 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% [28].

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.

6.1.3 User interface

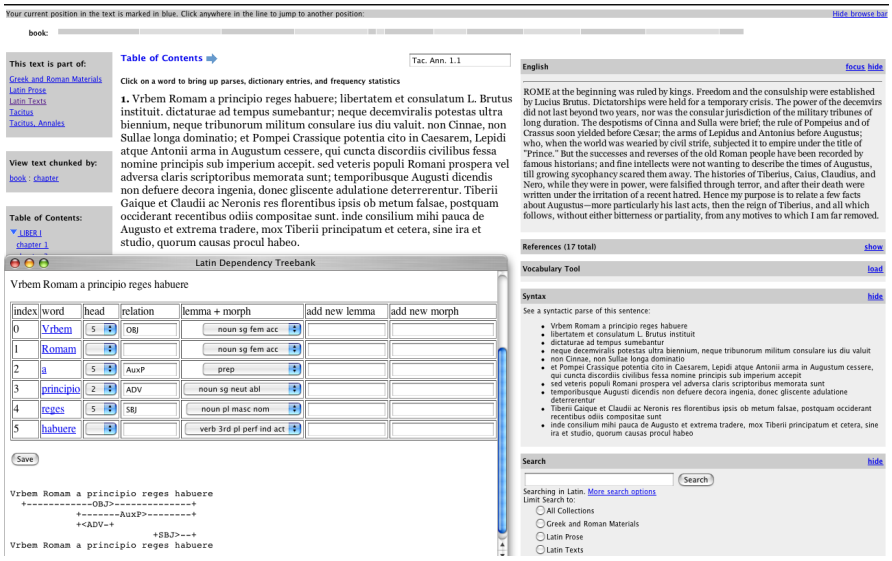


Fig. 4 A screenshot of Tacitus' *Annales* from the Perseus digital library.

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

Our collaboration with the Alpheios Project has also allowed us to integrate a graphical treebank editor into our annotation process to make the construction of trees more intuitive and to provide annotators with greater flexibility as to their preferred input method. Figure 5 shows a tree in the process of being constructed, with a single word (*Romam*) being dragged onto its syntactic head.

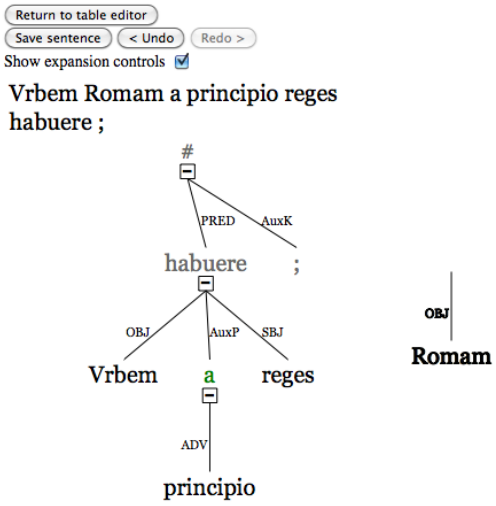


Fig. 5 A screenshot of the first sentence of Tacitus' *Annales* being constructed using the Alpheios graphical editor.

6.2 Reading support

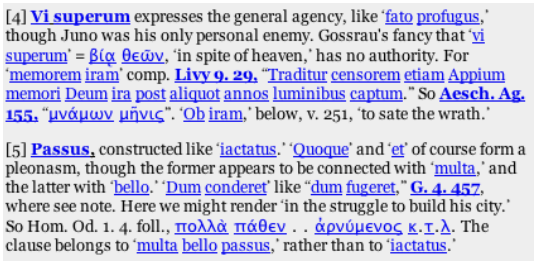
Modern treebanks also differ from historical ones in the fluency of their annotators. The efficient annotation of historical languages is hindered by the fact that no native speakers exist, and this is especially true of Ancient Greek and Latin, both languages with a high degree of flexibility in word order. 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 [30] and the Penn Chinese treebank can report a rate of 240-480 words per hour [31], our annotation speeds are significantly slower, ranging from 97 words per hour to 211, with an average of 124. Our best approach for these languages 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.

6.2.1 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.

6.2.2 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.'

Fig. 6 An excerpt from Conington's commentary on Vergil's *Aeneid* [32], here referring to Book 1, lines 4 and 5.

6.2.3 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 in the Perseus Digital Library 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**.

Fig. 7 Definition of *allabor* (the dictionary entry for *adlabitur*) from Lewis and Short [33].

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.

7 The Impact of Historical Treebanks

The traffic in the Perseus Digital Library currently exceeds 10 million page views by 400,000 distinct users per month. These users are not computational linguists or computer scientists who would typically make use of a treebank; they are a mix of Classical scholars and students. 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. The Ancient Greek and Latin Dependency Treebanks currently yield a powerful

range of search options, including lemmatized and morphosyntactic searching, and have already been valuable for downstream research involving lexicography and identifying textual reuse.

7.1 Lemmatized searching

The ability to conduct a lemma-based textual search has long been a desideratum in Classics,[7] where any given Latin word form, for example, has 3.1 possible analyses on average.[8] 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 easily accessible.

7.2 Morphosyntactic searching

A treebank's major contribution to scholarship is that it encodes an interpretation of the syntax of a sentence, along with a morphological analysis of each word. These two together can be combined into elaborate searches, allowing scholars to find all instances of any particular morphosyntactic construction, such as the different types of subordinate clauses headed by the conjunction *cum* (when *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). This type of searching allows us to gather statistical data on usage

[7] 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.

[8] 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).

while also locating individual examples for further qualitative analysis.[9] Figure 8 displays one tool for such analysis (Annis [35]) with a sample query from the Ancient Greek Dependency Treebank.

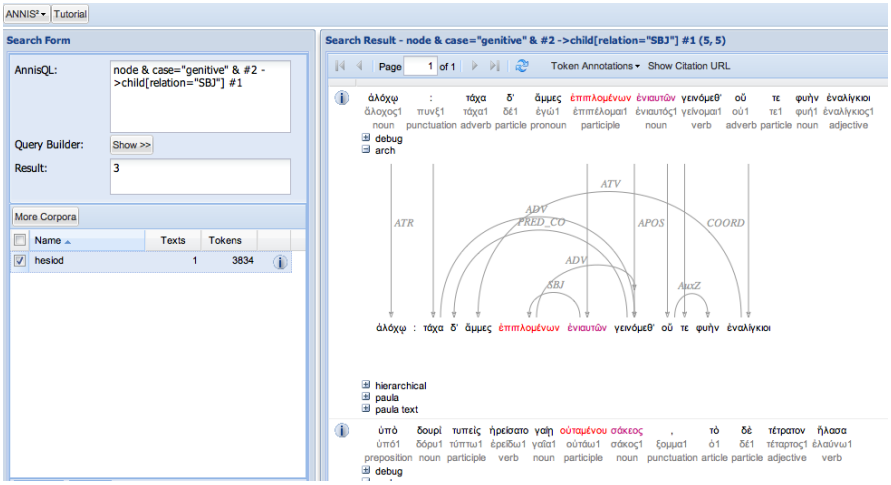


Fig. 8 Morphosyntactic search for genitive absolutes in Hesiod using the Annis search tool [35].

7.3 Lexicography

In addition to driving linguistic research on syntax itself, treebanks have been instrumental for several downstream computational tasks as well. One such task has been automatically inducing lexical information from large corpora in the service of automatically building bilingual dictionaries. Lexical information broadly defines what individual words "mean" and how they interact with others. Lexicographers have been exploiting large, unstructured corpora for this kind of knowledge in the service of dictionary creation since the COBUILD project [36] 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 [37], for instance, is built on a modern German corpus of 1.3 billion words, and we can expect much larger projects

[9] For the importance of a treebank in expediting morphosyntactic research in Latin rhetoric and historical linguistics, see Bamman and Crane [34].

in the future as the web is exploited as a corpus.[10] At the same time, researchers are also subjecting their corpora to more complex automatic processes in order 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 [39] also enable lexicographers to induce information about a word's grammatical behavior as well.

Treebanks have helped drive this work by providing a dataset from which to induce syntactic behavior for individual lexemes [40]. While it is large collections of parallel texts (Latin/English and Greek/English) that provide the basic material for mining the dominant English senses of Greek and Latin words [41], the role of a treebank here is to provide the training material for an automatic parser (such as McDonald et al's MSTParser [42]), which can then provide a syntactic parse for all of the source texts in our comparatively much larger collection. With this syntactic information, we can far better calculate a word's relationships to the other words in a sentence, and more properly delimit what "company" we want to consider when inferring its meaning.

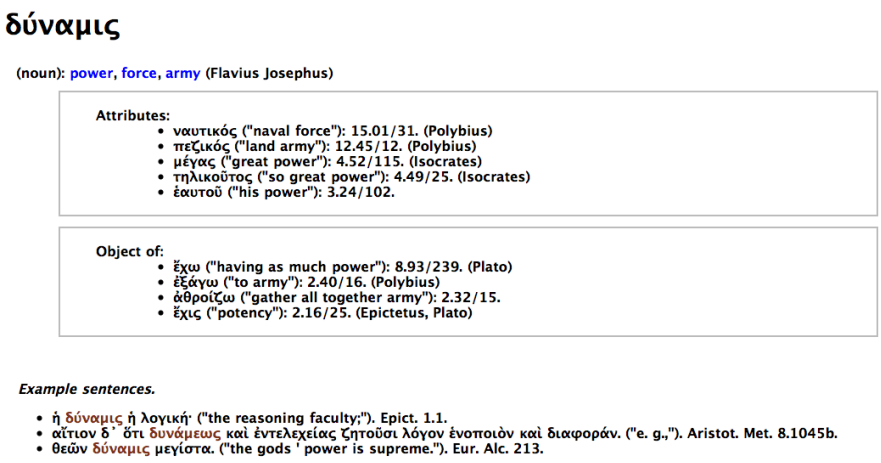


Fig. 9 Automatically derived lexical information for the Greek word δύναμις.

Figure 9 presents one example of such an automatically created lexical entry for the Greek noun δύναμις. While a traditional Greek lexicon such as the LSJ [43] can present much more detailed information about this word, we can here provide a quantitative measure of how frequently each sense appears in our corpus, and for which authors any given sense is dominant. δύναμις in general means "force" or "power" (the two most dominant senses found here), but it also retains a specialized

[10] In 2006, for example, Google released the first version of its Web 1T 5-gram corpus [38], a collection of n-grams (n=1-5) and their frequencies calculated from 1 trillion words of text on the web.

meaning of "military power" as a consequence. Syntactic information lets us specify not just what words it's commonly found with, but exactly *how* those words interact – for example, that when πεζικός ("on foot") modifies it as an attribute, it attains a new meaning of "army." Structural knowledge lets us distinguish between what surrounding words are merely descriptive attributes of a noun in question, and which words require that noun as part of their essential argument structure. While simple collocates induced from unstructured data provide information on what words accompany any individual lexeme, a treebank can specify the exact nature of their interaction on a much more detailed level.

7.4 Discovering textual similarity

Most studies on text reuse focus on identifying either documents that are duplicates or near-duplicates of each other (e.g., web pages) or sentences in one document that have been sampled from another (e.g., in plagiarism detection). These studies generally employ variations of word-level similarity, including relative frequency measures (spotting similarities in the distribution of word patterns between two documents) [44], IR similarity methods based on the TF-IDF scores of individual words [45] and fingerprinting using n-grams [46, 47, 48]. While n-grams are good at approximating syntax in languages with a relatively fixed word order (such as English and German), they are much less effective in languages where the word order is more free, such as Greek and Latin.

Additionally, when attempting to spot some of the more obliques classes of reuse – such as literary allusion – sometimes the strongest similarity can be found at a syntactic level. Consider, for example, the opening lines of the three great epics of Greco-Roman literature, Vergil's *Aeneid* and Homer's *Iliad* and *Odyssey.*

- arma virumque cano ("I sing of arms and the man") [Aen. 1.1]
- ἄνδρα μοι ἔννεπε, μοῦσα ("Tell me of the man, o Muse") [Od. 1.1]
- μῆνιν ἄειδε θεὰ ("Sing, goddess, of the rage") [Il. 1.1]

While there is a semantic similarity in all three examples (all three focus on the act of speaking and in two of the three it is a particular *man* that is spoken about), all three of them are most strongly similar by the explicit form of their structure. Figure 10 illustrates what these three phrases look like when annotated under a dependency grammar. In all cases, the initial phrase (arma/ἄνδρα/μῆνιν) is the direct object of the sentence predicate (cano/ἔννεπε/ἄειδε), wherever that happens to appear in the sentence.[11]

Our work in allusion detection [49] has focused on how to exploit the knowledge encoded in treebanks to automatically discover instances of textual reuse where the derived sentence bears some syntactic similarity to its source. Again, using our Latin

[11] Note that we can also add later epics to this class as well, such as Milton's *Paradise Lost*: "Of man's disobedience, and the fruit of that forbidden tree ... sing, heavenly muse" (1.1-6), where the first syntactic phrase in the sentence is the object of the verb of telling.

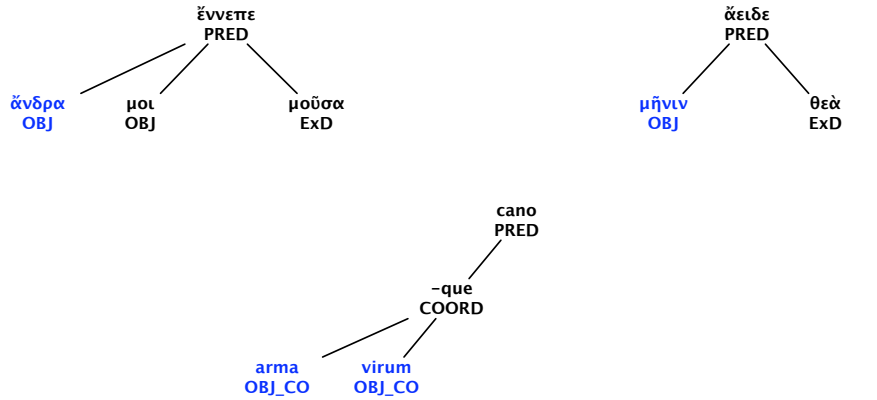


Fig. 10 Syntactic trees of the opening of the *Odyssey*, *Iliad*, and *Aeneid*.

treebank as training data for an automatic parser, we assigned a syntactic structure to all of the sentences in our larger textual collection. From this automatic structure we extracted 12 syntactic features for every word in the sentence – a combination of word-level representation (such as token, lemma or simply the part of speech), the length of the syntactic tree (including either just the parent or the parent and grandparent) and the presence or absence of an edge label (either simply specifying that a structural relation exists between a child and its parent, or also labeling that relationship as, e.g., an attributive one [ATR]). These features were then combined with other standard characteristics (such as word and lemma weights and n-grams) and used to calculate the similarity between two sentences, based on the cosine similarity between the two vectors that they constitute. Since each variable is weighted by TF-IDF, and syntactic features are relatively rare (with corresponding high IDF scores), syntactic features were generally found to be the most informative in establishing similarity. In its ability to generate this structural data, a treebank has enabled us to discover instances of text reuse even when the lexical similarity between two sentences is small and otherwise undetectable.

8 Conclusion

Treebanks already fill a niche in the computational linguistics community by providing valuable datasets for automatic processes such as parsing and grammar induction. Their utility, however, does not end there. The information that treebanks encode is of value to a wide range of potential users, including researchers not only in linguistics but in Classics as well, and we must encourage the use of these resources by making them available to such a diverse community. The treebanks so far are the work of hundreds of individuals who commit their interpretations of Greek and Latin sentences to a format that can be preserved for generations. While

this effort has resulted in the annotation of over 350,000 words of Classical texts, this is still only a small sample of the extant works in the Classical tradition; in the future, we plan to continue encouraging contributions to this ongoing work in order to strengthen, sentence by sentence, the foundation on which data-driven philology can stand.

9 Acknowledgments

Grants from the Alpheios Project ("Building a Greek Treebank"), the National Endowment for the Humanities (PR-50013-08, "The Dynamic Lexicon: Cyberinfrastructure and the Automated Analysis of Historical Languages"), the Andrew W. Mellon Foundation ("The CyberEdition Project: Workflow for Textual Data in Cyberinfrastructure"), the Digital Library Initiative Phrase 2 (IIS-9817484) and the National Science Foundation (BCS-0616521) provided support for this work. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. H.W. Smyth, *Greek Grammar* (Harvard University Press, 1920)
2. R. Kühner, C. Stegmann, *Ausführliche Grammatik der lateinischen Sprache II. Satzlehre. I. Teile Zweite Auflage* (Hahnsche Buchhandlung, Hannover, 1914)
3. P. Cuzzolin, On sentential complementation after *verba affectuum*, in *Linguistic Studies on Latin*, ed. by J. Herman (Benjamins, Amsterdam-Philadelphia, 1991), pp. 167–178
4. M.P. Marcus, B. Santorini, M.A. Marcinkiewicz, Building a large annotated corpus of English: The Penn Treebank, Computational Linguistics **19**(2), 313 (1994)
5. J. Hajič, O. Smrž, P. Zemánek, J. Šnaidauf, E. Beška, Prague Arabic dependency treebank: Development in data and tools, in *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools* (2004)
6. M. Maamouri, A. Bies, T. Buckwalter, W. Mekki, The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus, in *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools* (2004)
7. J. Hajič, Building a syntactically annotated corpus: The Prague Dependency Treebank, in *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, ed. by E. Hajičová (Prague Karolinum, Charles University Press, 1998), pp. 12–19
8. A. Kroch, A. Taylor. Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/ppcme2-release-2/ (2000)
9. A. Kroch, B. Santorini, L. Delfs. Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/ppceme-release-1 (2004)
10. A. Taylor, A. Warner, S. Pintzuk, F. Beths. York-Toronto-Helsinki Parsed Corpus of Old English Prose (2003)
11. V. Rocio, M.A. Alves, J.G. Lopes, M.F. Xavier, G. Vicente, Automated creation of a Medieval Portuguese partial treebank, in *Treebanks: Building and Using Parsed Corpora*, ed. by A. Abeillé (Kluwer Academic Publishers, 2003), pp. 211–227

12. P. Zemánek, A treebank of Ugaritic: Annotating fragmentary attested languages, in *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT2007)* (Bergen, 2007), pp. 213–218
13. D. Bamman, G. 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)* (Association for Computational Linguistics, Prague, 2007), pp. 33–40. URL http://www.aclweb.org/anthology/W/W07/W07-0905
14. M. Passarotti, Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus, in *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche*, ed. by P. Raffaella, F. Diego (Roma, Aracne Editrice, 2007), pp. 187–205
15. D. Haug, M. Jøhndal, Creating a Parallel Treebank of the Old Indo-European Bible Translations, in *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)* (2008)
16. I. Mel'čuk, *Dependency Syntax: Theory and Practice* (University of New York Press, Albany, 1988)
17. P. Sgall, E. Hajičová, J. Panevová, *The Meaning of the Sentence in its Semantic and Pragmatic Aspects* (Dordrecht: Reidel Publishing Company and Prague: Academia, 1986)
18. L. Tesnière, *Éleménts de syntaxe structurale* (Klincksieck, Paris, 1959)
19. N. Chomsky, Remarks on nominalization, in *Reading in English Transformational Grammar*, ed. by R. Jacobs, P. Rosenbaum (Ginn, Waltham, 1970)
20. D. Bamman, G. Crane, Guidelines for the syntactic annotation of Ancient Greek treebanks, version 1.1. Tech. rep., Tufts Digital Library, Medford (2009)
21. D. Bamman, M. Passarotti, G. Crane, S. Raynaud, Guidelines for the syntactic annotation of Latin treebanks, version 1.3. Tech. rep., Tufts Digital Library, Medford (2007)
22. D. Bamman, G. Crane, Pautas para la notación sintáctica del treebank de dependencia para el griego antiguo (1.1), traducción y adaptacón al español de Alejandro Abritta. Tech. rep., Tufts Digital Library, Medford (2010)
23. D. Bamman, M. Passarotti, G. Crane, S. Raynaud, Pautas para la notación sintáctica del treebank de dependencia para el latin (1.3), traducción y adaptacón al español de Alejandro Abritta. Tech. rep., Tufts Digital Library, Medford (2010)
24. D. Bamman, F. Mambrini, G. Crane, An ownership model of annotation: The Ancient Greek Dependency Treebank, in *The Eighth International Workshop on Treebanks and Linguistic Theories* (2009)
25. S. Brants, S. Dipper, S. Hansen, W. Lezius, G. Smith, The TIGER treebank, in *Proceedings of the First Workshop on Treebanks and Linguistic Theories* (Sozopol, 2002), pp. 24–41
26. S. Pintzuk, P. Leendert. York-Helsinki Parsed Corpus of Old English Poetry (2001)
27. G. Crane, From the old to the new: Integrating hypertext into traditional scholarship, in *Hypertext '87: Proceedings of the 1st ACM conference on Hypertext* (ACM Press, 1987), pp. 51–56
28. G. Crane, D. Bamman, L. Cerrato, A. Jones, D.M. Mimno, A. Packel, D. Sculley, G. Weaver, Beyond digital incunabula: Modeling the next generation of digital libraries., in *ECDL 2006* (2006), pp. 353–366
29. G. Crane, New technologies for reading: The lexicon and the digital library, Classical World pp. 471–501 (1998)
30. A. Taylor, M. Marcus, B. Santorini, The Penn Treebank: An overview, in *Treebanks: Building and Using Parsed Corpora*, ed. by A. Abeillé (Kluwer Academic Publishers, 2003), pp. 5–22
31. F.D. Chiou, D. Chiang, M. Palmer, Facilitating treebank annotation using a statistical parser, in *Proceedings of the First International Conference on Human Language Technology Research HLT '01* (2001), pp. 1–4
32. J. Conington (ed.), *P. Vergili Maronis Opera. The Works of Virgil, with Commentary* (Whittaker and Co, London, 1876)
33. C.T. Lewis, C. Short (eds.), *A Latin Dictionary* (Clarendon Press, Oxford, 1879)
34. D. Bamman, G. Crane, The design and use of a Latin dependency treebank, in *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)* (2006), pp. 67–78

35. A. Zeldes, J. Ritz, A. Lüdeling, C. Chiarcos, Annis: A search tool for multi-layer annotated corpora, in *Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009.* (2009)
36. J.M. Sinclair (ed.), *Looking Up: an account of the COBUILD project in lexical computing* (Collins, 1987)
37. A. Klosa, U. Schnörch, P. 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)
38. T. Brants, A. Franz, *Web 1T 5-gram Version 1* (Linguistic Data Consortium, Philadelphia, 2006)
39. A. Kilgarriff, P. Rychlý, P. Smrž, D. Tugwell, The sketch engine, in *Proceedings of the Eleventh EURALEX International Congress* (2004), pp. 105–116
40. D. Bamman, G. Crane, Building a dynamic lexicon from a digital library, in *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* (ACM, New York, NY, USA, 2008), pp. 11–20. DOI http://doi.acm.org/10.1145/1378889.1378892
41. P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, R.L. Mercer, The mathematics of statistical machine translation: parameter estimation, Comput. Linguist. **19**(2), 263 (1993)
42. R. McDonald, F. Pereira, K. Ribarov, J. 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* (2005), pp. 523–530
43. H.G. Liddell, R. Scott, H.S. Jones, R. McKenzie (eds.), *A Greek-English Lexicon, 9th edition* (Oxford University Press, Oxford, 1996)
44. T.C. Hoad, J. Zobel, Methods for identifying versioned and plagiarized documents, J. Am. Soc. Inf. Sci. Technol. **54**(3), 203 (2003). DOI http://dx.doi.org/10.1002/asi.10170
45. D. Metzler, Y. Bernstein, W.B. Croft, A. Moffat, J. Zobel, Similarity measures for tracking information flow, in *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management* (ACM, New York, NY, USA, 2005), pp. 517–524. DOI http://doi.acm.org/10.1145/1099554.1099695
46. S. Brin, J. Davis, H. García-Molina, Copy detection mechanisms for digital documents, SIGMOD Rec. **24**(2), 398 (1995). DOI http://doi.acm.org/10.1145/568271.223855
47. N. Shivakumar, H. Garcia-Molina, SCAM: A copy detection mechanism for digital documents, in *In Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries* (1995)
48. J. Seo, W.B. Croft, Local text reuse detection, in *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (ACM, New York, NY, USA, 2008), pp. 571–578. DOI http://doi.acm.org/10.1145/1390334.1390432
49. D. Bamman, G. Crane, The logic and discovery of textual allusion, in *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)* (Marrakesh, 2008)