Structured Knowledge for Low-Resource Languages: The Latin and Ancient Greek Dependency Treebanks

David Bamman and Gregory Crane

The Perseus Project
Tufts University
Medford, MA

Abstract. We describe here our work in creating treebanks – large collections of syntactically annotated data – for Latin and Ancient Greek. While the treebanks themselves present important datasets for traditional research in philology and linguistics, the layers of structured knowledge they contain (including disambiguated lemma, morphological, and syntactic information for every word) help offset the comparatively small size of extant Greek and Latin texts for text mining applications. We describe two such uses for these Classical treebanks – discovering lexical knowledge from a large corpus with the help of a small treebank, and identifying patterns of text reuse.

1 Introduction

Text mining techniques generally thrive best in high-resource languages for which large corpora have already been developed, such as English and German. The resources available for historical languages like Ancient Greek and Latin, however, pale in comparison to the billion-word raw textual corpora and million-word structured collections available for these modern languages.[1] All historical languages – both potentially high-resource ones like Latin (since it was a lingua franca for millennia) and those much less well attested, like Old Persian and Linear A (only known from a handful of inscriptions) – are finite in scope. While an archaeological find always has the potential to double the size of our extant texts, almost everything that will be said in historical languages has already been said.

This comparative sparseness means that we must develop techniques to get the most out of the texts that we have. What we lack in quantity we can, however, make up for in quality. Greek and Latin have been objects of study for over two thousand years, and have evolved a heavily scrutinized collection of texts with a large body of structured knowledge to contextualize them – textual indices, for instance, disambiguate personal and place names in texts, and individual commentaries make explicit the meaning and structure of the language itself. We are now contributing to this collection of structured knowledge by developing treebanks for Latin and Ancient Greek.

[1] The TLG E Disk [1], for instance, contains a total of 76 million words of Greek from Homer to ca. 600 CE; the Biblioteca Teubneria Latina BTL-1 [2] collection contains 6.6 million words, covering Latin literature up to the second century CE.

Treebanks – large collections of syntactically annotated data – provide fundamental datasets not only for computational tasks like automatic parsing and grammar induction, but also for more traditional lines of research, such as corpus linguistics and classical philology. As a resource in text mining, they have the potential to significantly reduce the amount of noise implicit in discovery by providing an additional layer of information on top of the raw text itself.

We have just released our Latin treebank with 50,000 words, and are in the process now of creating a treebank for Ancient Greek amounting to one million words. Since the word order of both of these languages is relatively free, we have based our annotation style not on the constituent-based grammars of (e.g.) the Penn Treebank of English but rather on the dependency grammar used by the Prague Dependency Treebank of Czech.

In this paper we will present our current work on developing these historical treebanks and the uses to which they can be applied.

2 The Latin Dependency Treebank

A treebank is large collection of sentences that have been syntactically annotated. In building our treebanks for Latin and Ancient Greek, we have relied on the standard production model of soliciting annotations from two independent annotators and then reconciling the differences between them. 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).

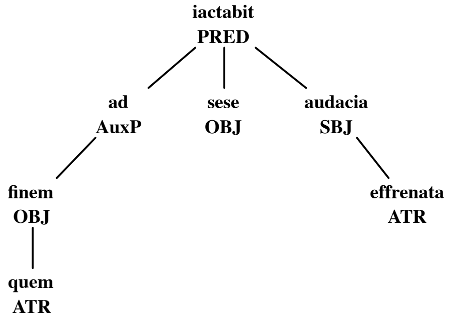
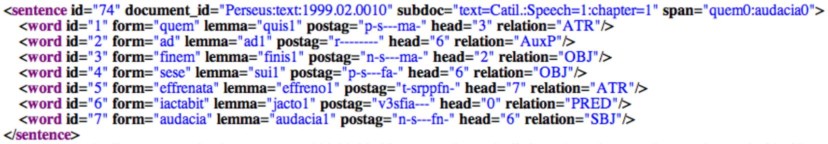


Fig. 1. Dependency graph of the treebank annotation for *quem ad finem sese effrenata iactabit audiacia* ("to what end will your unbridled audacity throw itself?"), Cicero, *In Catiliam* 1.1.

Figures 1 and 2 present two views of a syntactic annotation for a single sentence of Cicero (*quem ad finem sese effrenata iactabit audacia*).[2] 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).



```
<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>
```

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

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) [3] for Czech while tailoring it for Latin via the grammar of Pinkster [4]. 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, Greek 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 [5,6] on the works of Thomas Aquinas and the PROIEL corpus of the New Testament [7]) to create a common set of annotation guidelines to be used as a standard for Latin of any period [8]. This work has also allowed us to share our data as we annotate our respective texts [9]. Now in

Author	Words	Sentences
Caesar	1,488	71
Cicero	6,229	327
Jerome	8,382	405
Ovid	4,789	316
Petronius	12,474	1,114
Propertius	4,857	361
Sallust	12,311	701
Vergil	2,613	178
	53,143	3,473

Table 1. Composition of the Latin Dependency Treebank.

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

version 1.5, the Latin Dependency Treebank is comprised of 53,143 words from eight texts, as shown in table 1. All of the data is freely available.[3]

3 The Ancient Greek Dependency Treebank

Our work in developing a treebank for Latin has prepared us for the more ambitious project of creating a one-million-word treebank for Ancient Greek. Greek presents similar challenges to the representation of syntax as Latin (namely, a relatively free word order). While Greek does have a number of unique features (such as the presence of a definite article and a complex inventory of particles), much of the work that we have developed for Latin has been easily extensible to it (we could, for instance, transfer our representation of the Latin ablative absolute construction – whose annotation evolved significantly over a period of time – to equivalent constructions in Greek, the genitive and accusative absolute). Figure 3 displays a graphical tree for the first line of Hesiod's *Works and Days*.

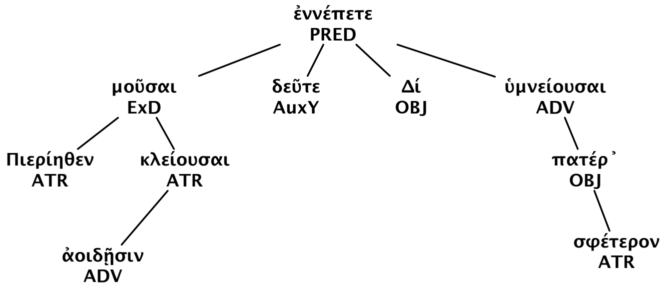


Fig. 3. μοῦσαι Πιερίηϑεν ἀοιδῆσιν κλείουσαι δεῦτε, Δί’ ἐννέπετε, σφέτερον πατέρ’ ὑμνείουσαι ("Muses of Pieria giving glory by song, here!, tell of Zeus, singing of your father") [Hesiod, *Works and Days* 1.1]

While our immediate goals include the annotation of both of Homer's epics, we plan in the near future to expand to Hesiod, Greek drama (including Aeschylus, Sophocles and Euripides), Plato, and other prose authors as well. Table 2 lists the work completed (and available for download) now.

Author	Words	Sentences
Homer, Odyssey	18,790	1,199
Homer, Iliad	3,945	251
	22,735	1,450

Table 2. Composition of the Ancient Greek Dependency Treebank.

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

One direction of research that our work in Greek will take in the future will be in the development of what we can consider “scholarly” treebanks. Classical texts present us with several layers of ambiguity. The first layer is one common to all areas of language (both ancient and modern) – the use of intentional ambiguity. While it is the province of syntactic annotation like this to specifically disambiguate sentences that can potentially have multiple syntactic interpretations – if taken in isolation – but only one specific meaning in a given context, some circumstances license multiple interpretations simultaneously. Jokes and *double entendres* are just two examples where that ambiguity is intentional – indeed, crucial for the intended purpose.

The second area of ambiguity is specific to ancient texts, since over two thousand years separate us now from when the original texts were composed. Diplomatic editions of a work like the *Aeneid* attempt to transmit the text as it appears within one specific manuscript, not necessarily as Vergil first wrote it, since the intervening time introduced variants (due to errors in the act of copying) from one generation of manuscript to the next. The more common “critical” edition of a text attempts to reconstruct what the author originally wrote by comparing several manuscripts, and noting (in an *apparatus criticus*) the letters, words and lines where important manuscripts differ from the “reconstructed” text.

In creating a scholarly treebank, we need to be able to create multiple annotations for a given sentence based on these two areas of ambiguity. We could accomplish this simply by layering additional levels of top of a base annotation, but the ability to encode competing annotations for any given sentence be an important area of research for us in the future.

4 Applications to text mining

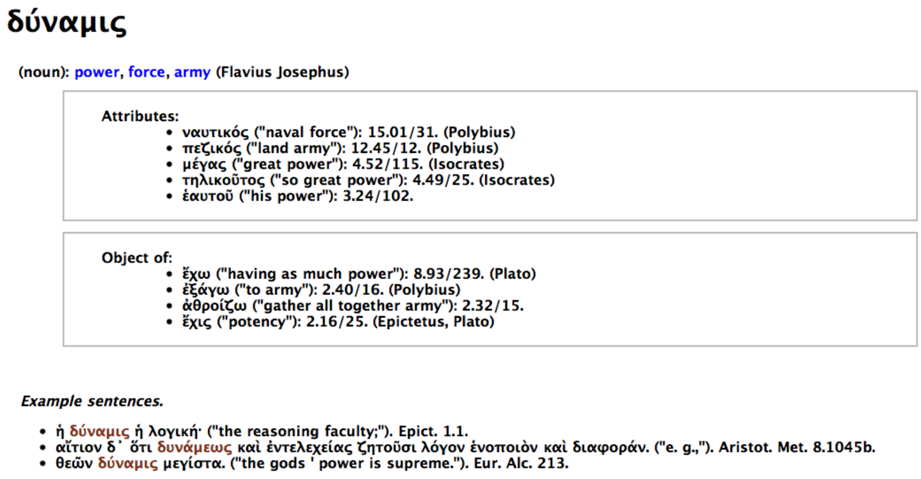
What treebanks offer to text mining applications is an additional layer of human knowledge that can be extracted and used as features. The syntax of a sentence is essentially an encoding of its structure, an abstraction away from the individual words toward the larger propositional knowledge that they touch upon. Because syntax is a structured abstraction, it provides a layer of meaning much broader than that that suggested by individual words alone.

While treebanks on their own are useful for other syntax-centered tasks (such grammar induction or the training of statistical parsers), they also play an important role for many downstream applications well. Two of these tasks are the automatic induction of lexical knowledge and the discovery of text reuse.

4.1 Inducing lexical knowledge

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 [10] 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 [11], for instance, is built on a modern German corpus of 1.3 billion words, and we can expect much larger projects in the future as the web is exploited as a corpus.[4] At the same time, researchers are also subjecting their corpora to more complex automatic processes to extract more knowledge from them. While word frequency and collocation analysis is fundamentally a task of simple counting, projects such as Kilgarriff's Sketch Engine [13] also enable lexicographers to induce information about a word's grammatical behavior as well.



δύναμις

(noun): power, force, army (Flavius Josephus)

Attributes:
- ναυτικός ("naval force"): 15.01/31. (Polybius)
- πεζικός ("land army"): 12.45/12. (Polybius)
- μέγας ("great power"): 4.52/115. (Isocrates)
- τηλικοῦτος ("so great power"): 4.49/25. (Isocrates)
- ἑαυτοῦ ("his power"): 3.24/102.

Object of:
- ἔχω ("having as much power"): 8.93/239. (Plato)
- ἐξάγω ("to army"): 2.40/16. (Polybius)
- ἀθροίζω ("gather all together army"): 2.32/15.
- ἕχις ("potency"): 2.16/25. (Epictetus, Plato)

Example sentences.

- ἡ δύναμις ἡ λογική· ("the reasoning faculty;"). Epict. 1.1.
- αἴτιον δ᾽ ὅτι δυνάμεως καὶ ἐντελεχείας ζητοῦσι λόγον ἑνοποιὸν καὶ διαφοράν. ("e. g.,"). Aristot. Met. 8.1045b.
- θεῶν δύναμις μεγίστα. ("the gods ' power is supreme."). Eur. Alc. 213.

Fig. 4. Automatically derived lexical information for the Greek word δύναμις.

One of our most fruitful areas of research into the intersection between large, unstructured collections (such as unannotated corpora) and small, structured data (such as treebanks) has been in our work in creating dynamic lexica for Greek and Latin [14].

To this end we have used the large collections of parallel texts (Latin/English and Greek/English) in the Perseus Digital Library to mine the dominant senses of a word by inducing its translation equivalents [15] by first aligning them on a chunk level (e.g., *book=1, chapter=1* in both a source text and its translation), then on a sentence level (using Moore's Bilingual Sentence Aligner [16]) and then on the level of individual words (using GIZA++ [17]). The role of a small structured knowledge source like a treebank here is to provide the training material for an automatic parser (such as McDonald et al's MSTParser [18]), which can then provide a syntactic parse for all of the source texts in our comparatively much larger collection. With this syntactic informa-

[4] In 2006, for example, Google released the first version of its Web 1T 5-gram corpus [12], a collection of n-grams (n=1-5) and their frequencies calculated from 1 trillion words of text on the web.

tion, we can much better calculate a word's relationships to the other words in a sentence, and more properly delimit what "company" we want to consider when delimiting its meaning.

Figure 4 presents one example of such an automatically created lexical entry for the Greek noun δύναμις. While a traditional Greek lexicon such as the LSJ [19] can present much more detailed information about this word, we can here provide a quantitative measure of how frequent each sense appears in our corpus, and in what specific authors any given sense is dominant within. A treebank informs this process by allowing us to determine what a word's common attributes are, and what verbs it's commonly the object of. Δύναμις in general means "force" or "power" (the two most dominant senses found here), but it also retains a specialized meaning of "military power" as a consequence. Syntactic information lets us specify not simply with what words it's commonly found with, but exactly *how* those words interact. While simple collocates induced from unstructured data tell you generally what words accompany any individual lexeme, it doesn't provide any information on the nature of that interaction. With a treebank we can distinguish between what surrounding words are *qualities*, for instance, of a noun in question, and which words require that noun as part of their essential argument structure.

4.2 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) [20], IR similarity methods based on the TF-IDF scores of individual words [21] and fingerprinting using n-grams [22,23,24]. While n-grams are good at approximating syntax in languages with a relatively fixed word order (such as English and German), they fail miserably 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 of the three great epics of Greco-Roman literature, Vergil's *Aeneid* and Homer's *Iliad* and *Odyssey.*

(1) arma virumque cano ("I sing of arms and the man") [Aen. 1.1]

(2) ἄνδρα μοι ἔννεπε, μοῦσα ("Tell me of the man, o Muse") [Od. 1.1]

(3) 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 5 illustrates what these three phrases would look like 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.[5]

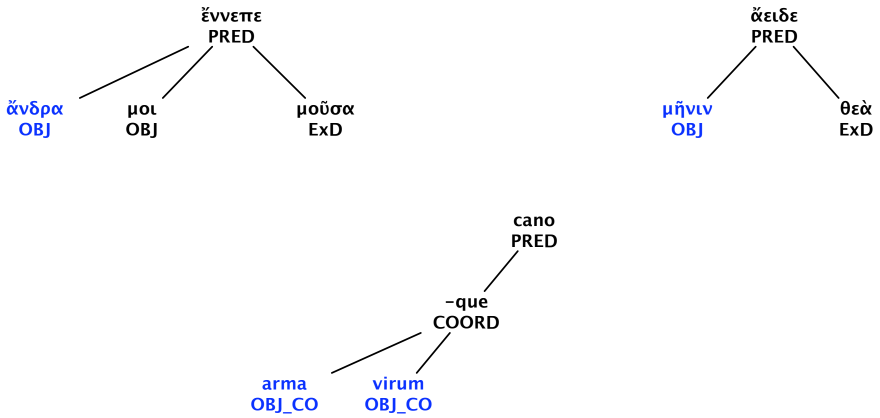


Fig. 5. Syntactic trees of the opening of the *Odyssey*, *Iliad*, and *Aeneid*.

Our work in allusion detection [25] has focussed 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 small 50,000 word Latin 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 (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 found to be generally the most informative in establishing similarity. Incorporating this data lets us discover text reuse even when the lexical similarity between two sentences is small and otherwise undetectable.

[5] 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 sentences is the object of verb of telling.

5 Conclusion

Once we get past the Classical era – past ca. 200 CE – the size of our textual collections written in Latin and Greek begins to approach a level comparable to that of some modern languages.[6] Within that Classical period, however, where Homer, Plato, Cicero, and Vergil – indeed, almost all of our most heavily studied authors – wrote, we are constrained by the humbling size of our extant texts. Even within these constraints, however, we can still mine these texts by elaborating upon the information they contains – making explicit in a layer of annotation what is only implicit in the text itself. The treebanks we are creating for Latin and Ancient Greek are an attempt to add layers of knowledge to these texts. In so doing, we are creating a versatile resource that can be applied not only to traditional tasks but can also form the basis for a wide range of text mining applications as well.

6 Acknowledgments

Grants from the Alpheios Project, the National Endowment for the Humanities (PR-50013-08, “The Dynamic Lexicon: Cyberinfrastructure and the Automated Analysis of Historical Languages”), the Mellon Foundation (“The CyberEdition Project: Workflow for Textual Data in Cyberinfrastructure”) and the National Science Foundation (BCS-0616521, “Automatically Building a Latin Treebank”) provided support for this work.

References

1. Berkowitz, L., Pantelia, M.C., eds.: Thesaurus Linguae Graecae CD ROM #E. CD-ROM. University of California, Irvine (1999)
2. Bibliotheca Teubneriana Latina: BTL-1. K. G. Saur, Stuttgart, Leipzig (1999)
3. Hajič, J.: Building a syntactically annotated corpus: The Prague Dependency Treebank. In Hajičová, E., ed.: Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová. Prague Karolinum, Charles University Press (1998) 12–19
4. Pinkster, H.: Latin Syntax and Semantics. Routledge, London (1990)
5. Busa, R.: 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)
6. Passarotti, M.: Verso il Lessico Tomistico Biculturale. La treebank dell’Index Thomisticus. In Raffaella, P., Diego, F., eds.: Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, Settembre 2006, Roma, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio (2007) 187–205
7. Haug, D., Jøhndal, M.: 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), Marrakesh (2008)

[6] The digital archive of the *Neulateinische Wortliste* [26], for example, contains over 300 million words of Neo-Latin texts.

8. Bamman, D., Passarotti, M., Crane, G., Raynaud, S.: Guidelines for the syntactic annotation of Latin treebanks, version 1.3. Technical report, Tufts Digital Library, Medford (2007)
9. Bamman, D., Passarotti, M., Crane, G., Raynaud, S.: A collaborative model of treebank development. In: Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT2007), Bergen (2007) 1–6
10. Sinclair, J.M., ed.: Looking Up: an account of the COBUILD project in lexical computing. Collins (1987)
11. Klosa, A., Schnörch, U., Storjohann, P.: 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)
12. Brants, T., Franz, A.: Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia (2006)
13. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The sketch engine. In: Proceedings of the Eleventh EURALEX International Congress. (2004) 105–116
14. Bamman, D., Crane, G.: Building a dynamic lexicon from a digital library. In: JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, New York, NY, USA, ACM (2008) 11–20
15. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**(2) (1993) 263–311
16. Moore, R.C.: Fast and accurate sentence alignment of bilingual corpora. In: AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, London, UK, Springer-Verlag (2002) 135–144
17. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51
18. McDonald, R., Pereira, F., Ribarov, K., Hajič, J.: 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) 523–530
19. Liddell, H.G., Scott, R., Jones, H.S., McKenzie, R., eds.: A Greek-English Lexicon, 9th edition. Oxford University Press, Oxford (1996)
20. Hoad, T.C., Zobel, J.: Methods for identifying versioned and plagiarized documents. J. Am. Soc. Inf. Sci. Technol. **54**(3) (2003) 203–215
21. Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, New York, NY, USA, ACM (2005) 517–524
22. Brin, S., Davis, J., García-Molina, H.: Copy detection mechanisms for digital documents. SIGMOD Rec. **24**(2) (1995) 398–409
23. Shivakumar, N., Garcia-Molina, H.: 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)
24. Seo, J., Croft, W.B.: Local text reuse detection. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2008) 571–578
25. Bamman, D., Crane, G.: The logic and discovery of textual allusion. In: Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008), Marrakesh (2008)
26. Ramminger, J.: Neulateinische Wortliste. Ein Wörterbuch des Lateinischen von Petrarca bis 1700. `http://www.neulatein.de` (2003ff.)