##DIGITAL CLASSICS ONLINE##

An Automated Approach to Syntax-based Analysis of Classical Latin

Anjalie Field

Abstract: The goal of this study is to present an automated method for analyzing the style of Latin authors. Many of the common automated methods in stylistic analysis are based on lexical measures, which do not work well with Latin because of the language's high degree of inflection and free word order. In contrast, this study focuses on analysis at a syntax level by examining two constructions, the ablative absolute and the *cum* clause. These constructions are often interchangeable, which suggests an author's choice of construction is typically more stylistic than functional. We first identified these constructions in hand-annotated texts. Next we developed a method for identifying the constructions in unannotated texts, using probabilistic morphological tagging. Our methods identified constructions with enough accuracy to distinguish among different genres and different authors. In particular, we were able to determine which book of Caesar's *Commentarii de Bello Gallico* was not written by Caesar. Furthermore, the usage of ablative absolutes and cum clauses observed in this study is consistent with the usage scholars have observed when analyzing these texts by hand. The proposed methods for an automatic syntax-based analysis are shown to be valuable for the study of classical literature.

1. Introduction

Over the past 50 years, computational methods have become an essential tool for analyzing the style and authorship of texts. Common problems in authorship analysis include plagiarism detection, author profiling, detection of stylistic inconsistencies, and authorship verification.[1] Style and authorship studies are particularly applicable to classical texts, which often involve authorship controversies. Furthermore, since texts from the Greek and Roman era have barely survived, it is in a classicist's best interest to glean as much information as possible from each text. Additionally, classical texts are works that scholars actually care about, as they contribute to scholarship rather than transient interest. Many authorship studies focus on modern documents like newspaper articles, even though most scholars care more about the use of literary devices in Vergil than in the Wall Street Journal. Finally, the volume of electronically available classical texts, especially in Latin, far surpasses the hand-analysis abilities of a small community of scholars.[2]

One of the central ideas behind stylistic and authorship analysis is that certain features of writing are unique to an author, so that even across different genres, texts by the same author will have certain similarities.[3] Most analyses seek to distinguish an author's style by choosing a set of features and using classifiers or distance functions to compare various texts. The traditional

[1] Stamatatos (2009).

[2] Bamman / Crane (2006).

[3] Diederich et al. (2003).

feature sets tend to be at a word or character level, such as the frequency of function words, n-grams of words, characters, or parts of speech, i.e. how often the phrase „if you give,“ the letters „ify,“ or the combination „conjunction pronoun verb“ occur.[4] These traditional feature sets fail to work well with Latin, primarily because Latin is a highly inflected language with a very free word order. In English, function words, such as prepositions, are used to convey the relation among words, but in Latin the relation among words is often expressed by the form of each word rather than by a function word, so measuring the frequency of function words is far less informative.

Furthermore, the prevalence of inflection in Latin makes metrics like counting the most common words difficult. In order to count word frequencies, it is necessary to lemmatize each word in the text, i.e. associate the words „walking“ and „walked“ as forms of the same word, „to walk.“ Because of the many overlapping forms of words, it can be difficult to determine the stem of a word without completely parsing the text. Attempts to parse Latin have been successful on a limited scale: Covington and Koch have both proposed methods for parsing Latin that focus on a small subset of the language, Passarotti reports a high rate of accuracy for dependency parsing medieval Latin, and Koster presents a rule-based method capable of parsing simple sentences.[5] However, there does not yet exist a parser capable of handling the complicated syntax of classical Latin with high accuracy. Because parsing classical Latin remains a non-trivial task, precise lemmatization also remains difficult. Finally, the high degree of inflection in Latin allows for very free word order. Some loose conventions exist, such as placing the subject of a sentence near the beginning and the verb at the end, but in general Latin words can occur almost anywhere in a sentence. Not only does this make parsing more difficult, it also makes metrics like n-gram frequencies less meaningful.

Because of the difficultly of applying lexically based methods to stylistic analysis, one of the goals of this study was to analyze authors’ style at a syntax level. Latin has the ability to express the same idea in many ways by using various types of grammar constructions. Purpose can be expressed using a purpose clause, a relative clause of purpose, a gerund, a gerundive, or a supine.[6] Lexical measures in Latin often fail to represent syntax and grammar, like the difference between a gerund and a supine, which can be strong indicators of style. Additionally, a syntax-based approach has more potential for cross-language analysis than lexically based methods. For example, A.D. Leeman suggests that the Roman historian Sallust was greatly influenced by the Greek historian Thucydides.[7] Jonas Grethlein questions this comparison and instead suggests that Sallust’s writing contains elements of Herodotus.[8] Since many grammar constructions exist in both Latin and Ancient Greek, and some commonalities are easy to identify across languages, a syntax-based approach offers a way to quantify these cross-language comparisons. In general, a syntax-level analysis more closely represents how a classicist might approach reading a text, by paying attention to the use of grammar constructions.

Our paper specifically focuses on two such grammar constructions: the ablative absolute and the *cum* clause. The first part of the study involved developing methods for identifying ablative absolutes and *cum* clauses in texts. Since parsing in Latin is still a difficult problem, we

[4] Mosteller / Wallace (1964), Stamatatos (2009).

[5] Covington (1990), Koch (1994), Passarotti / Dell’Orletta (2010), Koster (2005).

[6] Moreland / Fleischer (1990).

[7] Leeman (1963).

[8] Grethlein (2006).

aimed to show that it is possible to analyze grammar without full scale parsing. Instead, the two constructions were identified by combining part-of-speech tags with a rule-based approach. Furthermore, unlike previous syntax-based analyses, our methods were designed to work on texts without any annotations. The second part of this study involved finding ablative absolutes and *cum* clauses in a variety of texts and looking for patterns in usage across authors and genres.

2. The Ablative Absolute and the *Cum* Clause

The ablative absolute usually consists of a participle, often a perfect passive participle, and a noun in the ablative case. The construction may not have a participle, consisting simply of a noun or an adjective in the ablative, and it may contain additional words, such as objects, adjectives or other qualifiers. It is also grammatically independent from the rest of the sentence, with a different subject and not directly referring to any words in the rest of the sentence. Hence it is „absolute.“ An example is: *his responsis ad Caesarem relatis, iterum ad eum Caesar legatos cum his mandatis mittit,* which translates: “When these answers were reported to Caesar, he sends ambassadors to him a second time with this message” (Caesar, Commentarii de Bello Gallico, 2.5, Translator W. S. Bohn).

The ablative absolute is typically used to provide background or contextual information. It can express time, condition, opposition, cause, or attendant circumstance, and is often best translated with „when“, „since,“ or „although.“[9] The construction is especially common in military and historical accounts, because it allows the author to convey information concisely.[10]

The *cum* clause is also often used adverbially to provide background information. It consists of the conjunction *cum* with a verb in either the indicative or the subjunctive. With the indicative, it is almost always temporal, translated as „when“. With the subjunctive, it can also be causal or concessive, translated as „since“ or „although.“[11] Thus, the *cum* clause and the ablative absolute are both adverbial clauses, used to express contextual information, and in many cases, are somewhat interchangeable.[12] There are other types of phrases used to express contextual information, and there are situations in which ablative absolutes and *cum* clauses are not interchangeable. However, authors generally use these constructions in similar ways.
The similarities between these two constructions suggest the following hypotheses:

1. An author’s decision to use a *cum* clause or an ablative absolute is often more stylistic than functional, so the relative frequency of cum clauses and ablative absolutes in an author’s work is indicative of the author’s style.

2. The relative frequencies of *cum* clauses and ablative absolutes are similar for a given author, with some consistency even across genres, but vary significantly across different authors.

3. Authors writing in the same genre use more similar distributions of ablative absolutes and *cum* clauses than authors writing in different genres.

[9] Bennett (1918).

[10] Leeman (1963); Von Albrecht (1979).

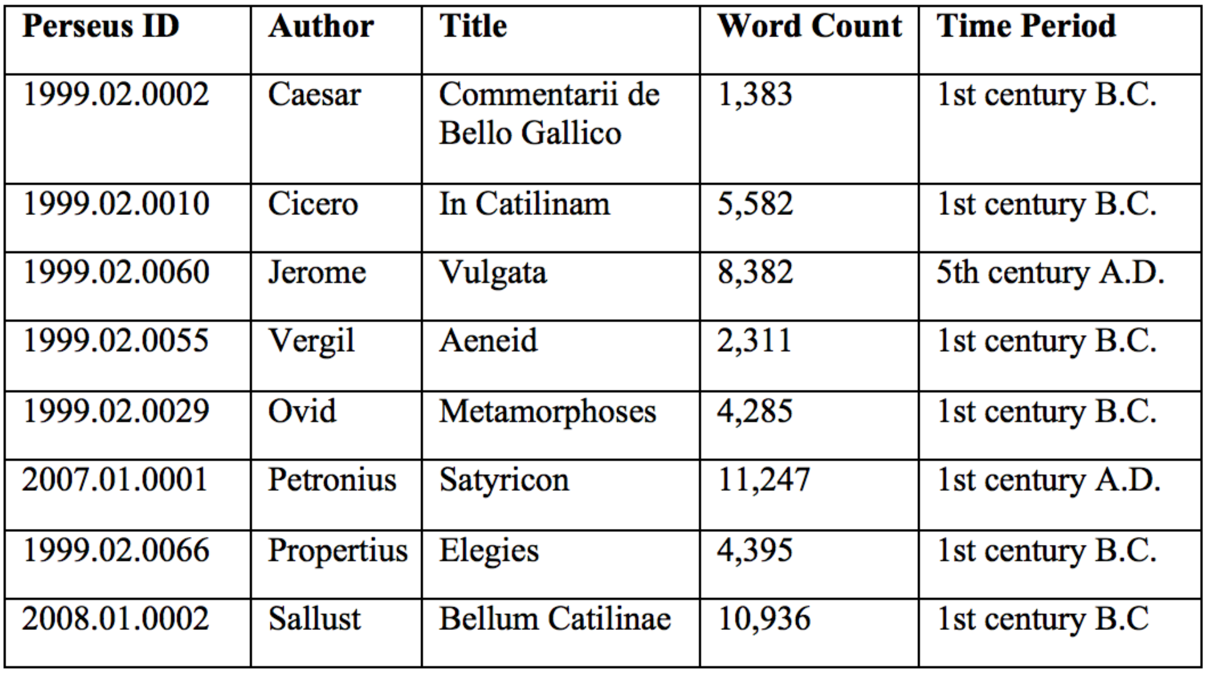
[11] Bennett (1918).

[12] Moreland / Fleischer (1990).

3. Methodology

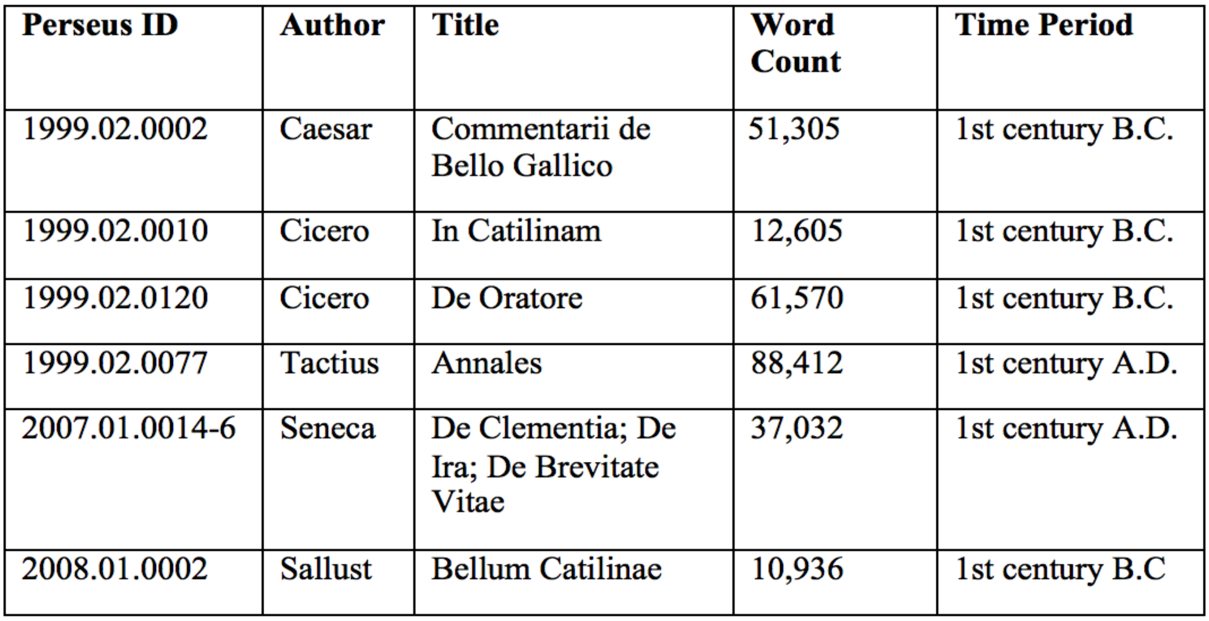
3.1 Description of Data Sets

Parts of this study relied on the use of the Latin Dependency Treebank (LDT), a corpus of syntactically tagged Latin sentences.[13] Table 1 describes the texts included in the annotated data set. The total number of tokens in this data set is 53,143 (48,521 excluding punctuation). Each word in the corpus has been hand-annotated with morphological and part-of-speech information. Each sentence has been further parsed according to a dependency grammar.



Perseus ID	Author	Title	Word Count	Time Period
1999.02.0002	Caesar	Commentarii de Bello Gallico	1,383	1st century B.C.
1999.02.0010	Cicero	In Catilinam	5,582	1st century B.C.
1999.02.0060	Jerome	Vulgata	8,382	5th century A.D.
1999.02.0055	Vergil	Aeneid	2,311	1st century B.C.
1999.02.0029	Ovid	Metamorphoses	4,285	1st century B.C.
2007.01.0001	Petronius	Satyricon	11,247	1st century A.D.
1999.02.0066	Propertius	Elegies	4,395	1st century B.C.
2008.01.0002	Sallust	Bellum Catilinae	10,936	1st century B.C

Table 1: Description of annotated data set



Perseus ID	Author	Title	Word Count	Time Period
1999.02.0002	Caesar	Commentarii de Bello Gallico	51,305	1st century B.C.
1999.02.0010	Cicero	In Catilinam	12,605	1st century B.C.
1999.02.0120	Cicero	De Oratore	61,570	1st century B.C.
1999.02.0077	Tactius	Annales	88,412	1st century A.D.
2007.01.0014-6	Seneca	De Clementia; De Ira; De Brevitate Vitae	37,032	1st century A.D.
2008.01.0002	Sallust	Bellum Catilinae	10,936	1st century B.C

Table 2: Description of unannotated data set

[13] Bamman et al. (2007), http://nlp.perseus.tufts.edu/syntax/treebank/.

For this study, we also compiled a data set of unannotated texts, which are summarized in Table 2. The data set consists of classical prose, and we specifically chose it to represent a range of authors and genres. Notably, the data set includes two different works by Cicero and works by several historians. All texts were obtained from the Perseus Project.[14]

3.2 Identification of Ablative Absolutes and *Cum* Clauses in Hand-Annotated Data

We developed rules for identifying ablative absolutes and *cum* clauses in the annotated texts that target how the texts were annotated. *Cum* clauses are relatively straightforward constructions, consisting of simply the word *cum* functioning as a conjunction with a finite verb in either the indicative or the subjunctive. However, the word *cum* can be used as either a conjunction or a preposition in Latin. In the hand-annotated data, a *cum* signifying a *cum* clause can be easily distinguished from cum the preposition, because words are tagged with their part-of-speech. Thus, according to our rules, any instance of the word *cum*, where *cum* was tagged as a conjunction, was counted as a *cum* clause. The number of *cum* clauses in each text was counted using a python script.

In contrast, the ablative absolute is a more ambiguous construction that can take various forms. While the basic form involves a noun and a participle in the ablative case, it is also possible to omit the participle and have simply a noun, usually with an adjective, in the ablative case. Furthermore, the participle can govern other objects or qualifiers, such as adjectives or prepositional phrases. Since this study aimed to count the number of ablative absolutes in a text, we focused on identifying the nouns and participles that signify an ablative absolute, without considering other words that might be part of the construction. The annotation guidelines describe how ablative absolutes were annotated: “the noun should be annotated as the subject of the participle, with the participle (as the head of the ablative absolute phrase) depending on the main verb as an adverbial.”[15] However, the actual annotations are not so clear-cut. Cases where ablative absolutes have multiple participles and nouns are annotated differently. For example, participles do not always depend on the predicate in the sentence; they can instead depend on other words, like conjunctions.

Because the defined description of the annotation of ablative absolutes is too simplistic, we tested various restrictions to determine a set of rules that most accurately finds ablative absolutes. The testing started with a very broad definition, namely flagging any clause that contains an ablative participle, and then we added in more constraints to eliminate false positives. These trials focused on finding the most accurate system for identifying ablative absolutes with only a few rules, rather than trying to cover all possible combinations of conjunctions, subordinate clauses, and commas.
The final criteria we used to identify ablative absolutes are:

1. The phrase must contain a participle in the ablative case with an adverbial relation
2. The phrase must contain a noun with a subject relation
3. The noun must meet one of the following:

[14] http://www.perseus.tufts.edu.

[15] Bamman (2007).

- Depend on the participle
- Depend on a conjunction that depends on the participle (indicates two nouns, 1 participle)
- Depend on conjunction that participle also depends on (indicates 2 participles, 1 noun)

This classification excludes some constructions, notably, ablative absolutes containing no participle. However, this restriction significantly reduces the number of false positives, and it simply focuses the study on a more specific construction: an ablative absolute containing a participle.

Our search counts the number of ablative absolutes in a text by the number of participle-noun pairs, thus an ablative absolute with 1 noun and 2 particles would count as 1 ablative absolute. We implemented these rules by using a python script to search through each text.

3.3 Identification of Ablative Absolutes and *Cum* Clauses in Unannotated Data

The purpose of this study was to automate the analysis of large corpora, rather than just small hand-annotated corpora. This goal necessitated methods for identifying constructions in unannotated texts.

In order to better handle ambiguous word forms, TreeTagger was used to assign part-of-speech tags and case tags to each word in the text. This program uses decisions trees to conduct probabilistic tagging.[16] The hand-annotated data in the LDT was used as training data, since using these data resulted in higher accuracy than the provided training files. In processing a text, we first tagged the entire text for part-of-speech and for case. Then, we divided the text into clauses according to all punctuation markers, including periods, commas, semicolons, parentheses, brackets, and quotation marks. Finally, the text was searched for *cum* clauses and ablative absolutes.

Like the search for *cum* clauses in the hand-annotated texts, the search for *cum* clauses in the unannotated texts used the part-of-speech tags assigned by TreeTagger to distinguish between *cum* the preposition and *cum* the conjunction. Any clause containing the word *cum* that was tagged as a conjunction by TreeTagger was considered to be a *cum* clause.
Ablative absolutes were identified as:

1. The clause contains a word tagged as a participle and tagged as the ablative case
2. The clause contains a noun that could match in gender, number, and case with the participle

The phrasing “could match” refers to the ambiguity of Latin word forms. More specifically, for a clause containing a participle, rule 2 above is satisfied if any other word in the clause can be interpreted as noun matching in gender, number, and case with the participle, even if the word can be interpreted in a different way. Thus if a clause contains a masculine participle and a noun that could be masculine or feminine, rule 2 would be satisfied. Lemmatization of words, or the identification of their possible forms was performed using the Morpheus Engine developed by the Perseus Project. When queried, Morpheus provides all possible forms of the given word. Texts were pre-processed by querying Morpheus for all words in the text and sto-

[16] Schmid (1994).

ring lemmatization information in a local database, which was then used to lemmatize words in a text while searching for ablative absolutes.

We automated tagging with TreeTagger and implementation of the rules using python scripts. Each punctuation-separated clause was counted at most once, even if it contained multiple ablative absolutes.

4. Results

4.1 Identification of Syntactic Constructions in Hand-Annotated Texts

Figures 1 and 2 show the frequencies of ablative absolutes and *cum* clauses for each author in the hand-annotated corpus. For texts where fewer than 20 clauses were identified as containing the given syntactic construction, each clause was hand-checked to determine if it contained an ablative absolute or *cum* clause. For texts where more than 20 clauses were identified (Ovid, Petronius, Propertius, Sallust for ablative absolutes; Cicero, Jerome, Petronius for *cum* clauses), 1 in 5 constructions was hand-checked.

For ablative absolutes, 1 false positive was identified in Cicero, and for *cum* clauses, 1 false positive was identified in Petronius, which appears to be the result of an incorrect tag. The lack of false positives suggests that these constructions were found with high precision. Hand-checking the found constructions does not ensure that the search method had a high recall rate, but the search for *cum* clauses was very straightforward with little room for error. Additionally, the use of broad search criteria when identifying ablative absolutes, which relied primarily on the definition of the construction, suggests that this method was very inclusive with few false negatives.

While *cum* clauses are unmistakable, it can be ambiguous whether or not a construction is strictly an ablative absolute. Our definition of an ablative absolute is very broad. For example, one construction identified in Ovid was: “sic aquilam penna fugiunt trepidante columbae” (Ovid, *Metamorphoses*, 1.506), which translates “thus doves flee from an eagle with a trembling wing.” The phrase „trembling wing“ was considered an ablative absolute. However, this phrase could also be taken as a simple ablative of description, depicting what the doves look like, or even as an ablative of means, explaining how the doves flee. The phrase „ablative absolute“ describes a particular usage of the ablative case, but usages of the ablative are not always easy to classify. Thus, this study more generally focused on ablative participles that are used adverbially in a sentence.

Figure 1 shows that Caesar uses ablative absolutes most frequently, while Jerome uses no ablative absolutes and Cicero uses very few. In contrast, Figure 2 shows that Cicero uses *cum* clauses very frequently. The lowest usage rates of *cum* clauses are found in Sallust and in Vergil.

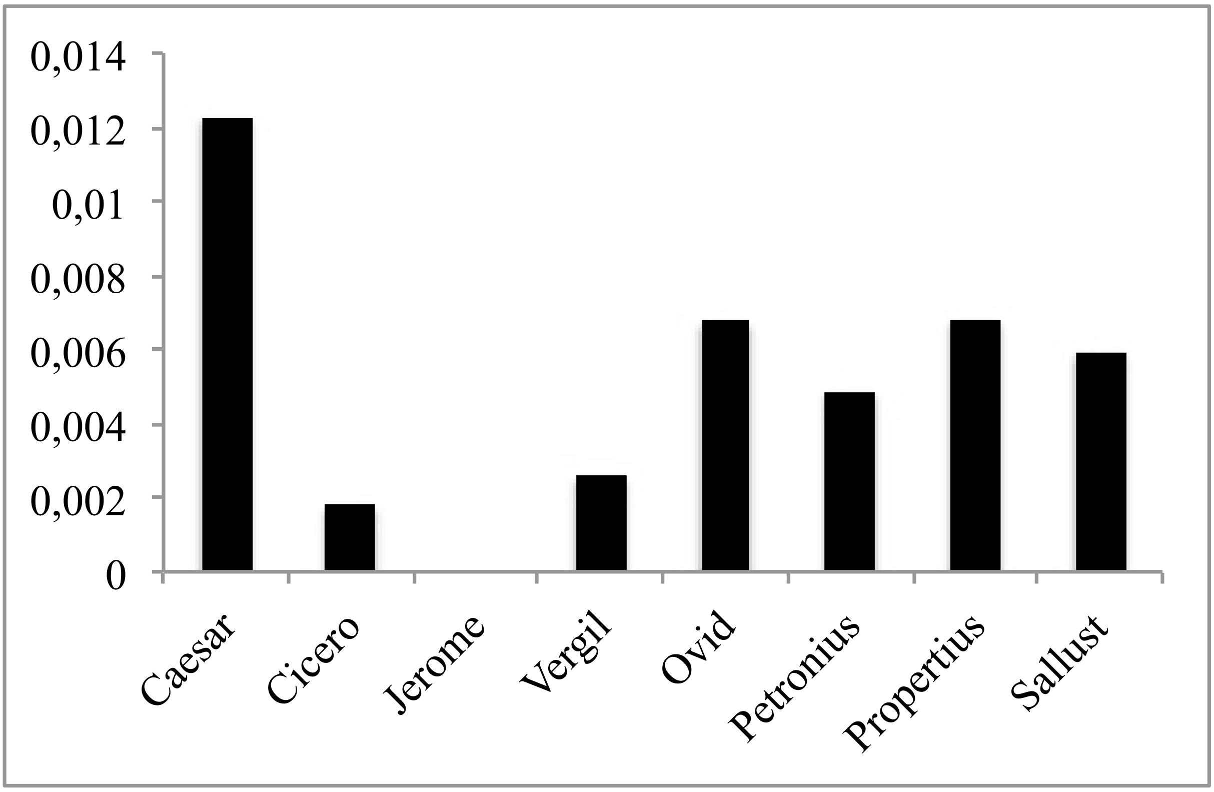


Figure 1: Rates of ablative absolutes in hand-annotated texts, expressed as the number of ablative absolutes found divided by the number of words in each text

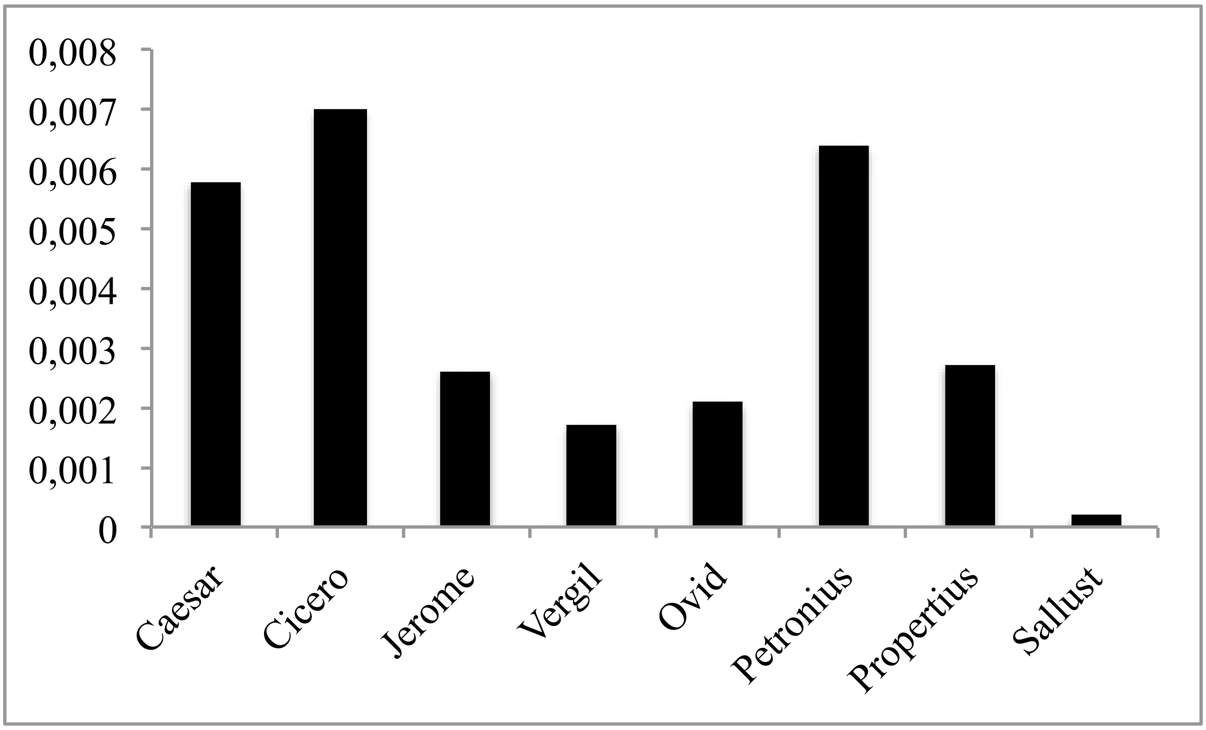


Figure 2: Rates of *cum* clauses in hand-annotated texts, expressed as the number of *cum* clauses found in each text divided by the number of words, excluding punctuation, in the text

4.2 Identification of Syntactic Constructions in Unannotated Texts

To identify the target structures in unannotated texts, we used TreeTagger to tag the texts followed by the application of rules to identify constructions. In order to measure the accuracy of this method, it was tested on the hand-annotated data, but ignoring annotations. The sentences in the hand-annotated text were randomly divided into 10 equal sections. For each of the 10 sections, 1 section was used as test data and the remaining 9 sections were used as training data. The accuracy rates are calculated as: total number of correct tags / total number of tags across all 10 tests. Figure 3 shows the accuracy rates for using TreeTagger on the hand-annotated data. In using TreeTagger for identifying ablative absolutes and *cum* clauses, only part-of-speech (95.5% accuracy) and case tags (84.0%) were considered.

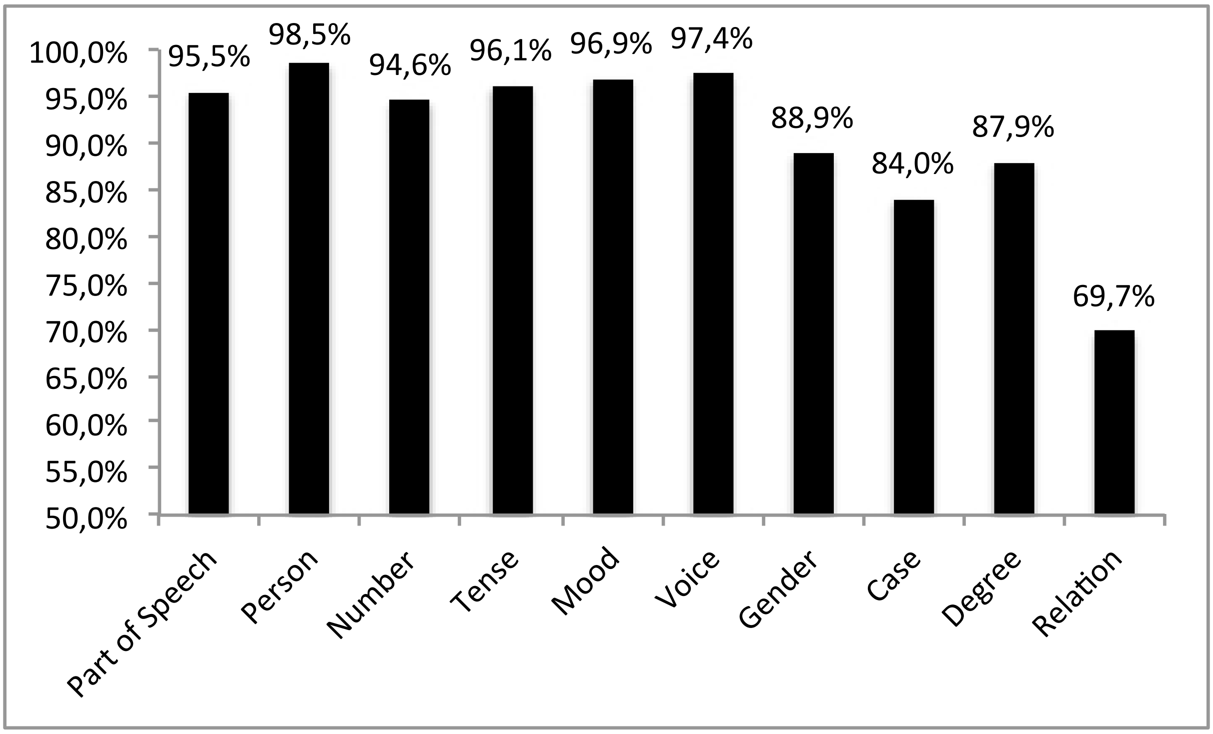
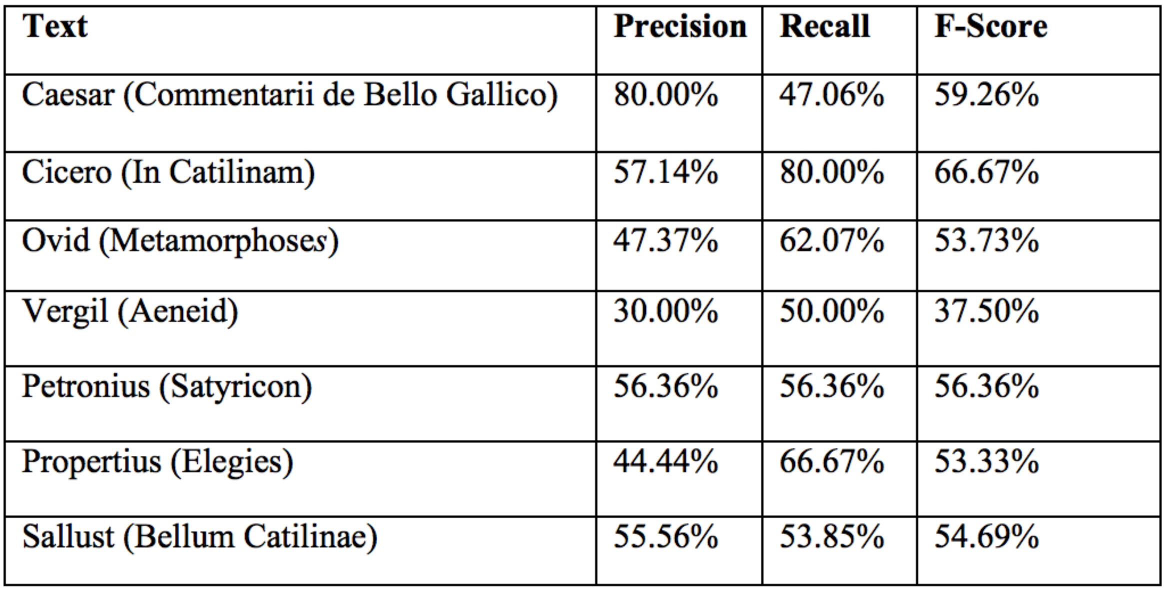


Figure 3: Accuracy of TreeTagger in tagging morphological information in the hand-annotated texts

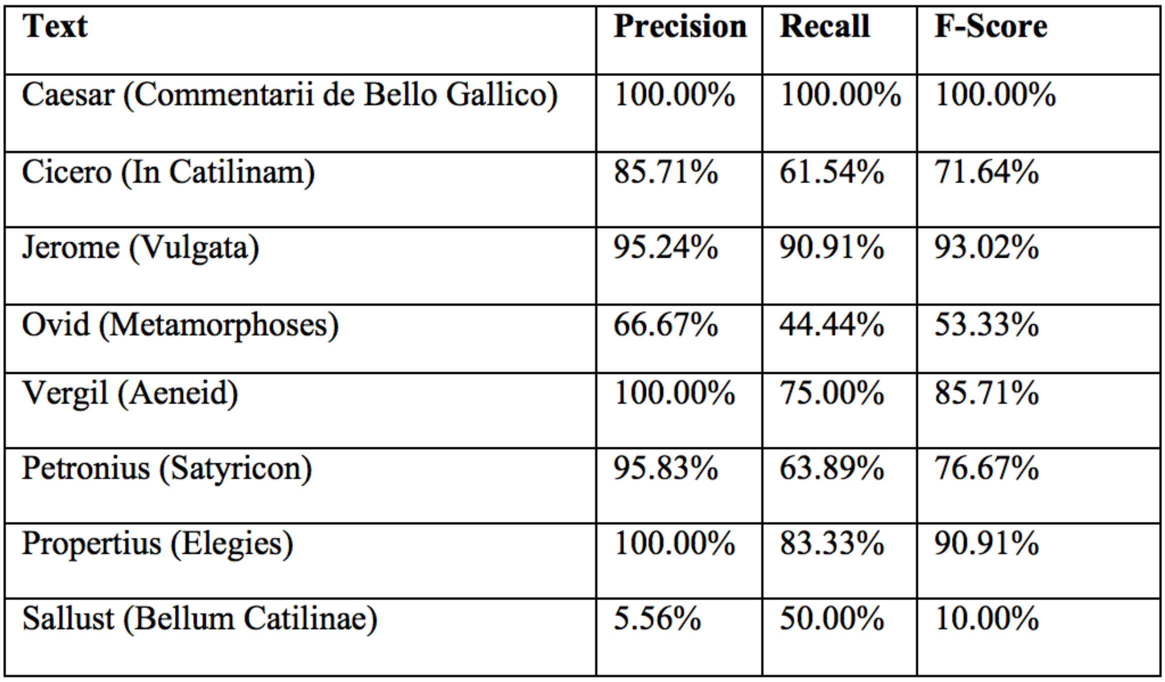
Tables 3 and 4 show the accuracy rates for identifying syntactic constructions in the hand-annotated data, ignoring annotations. Precision was measured as (number of clauses found with and without using annotations) / (number of clauses found without using annotations). Recall was measured as (number of clauses found with and without using annotations) / (number of clauses found using annotations). Thus, precision measures how many of the clauses identified actually contained the correct constructions, while recall measures how many of the clauses containing the correct constructions were identified. F-Score, essentially a weighted average of precision and recall, is defined as: 2 * (precision * recall) / (precision + recall).

The lowest accuracy rates occur for identifying ablative absolutes in Ovid and Vergil. However, the search was able to identify *cum* clauses in Caesar with 100% accuracy. The size of the data set for Vergil is very small, with only 6 ablative absolutes and 4 *cum* clauses. Similarly, the analyzed section of Sallust only contains 2 *cum* clauses. A larger data set would likely result in more reliable estimates of precision and accuracy. Because the data set was small, the variations in the accuracy could be exaggerated. Jerome is omitted from Table 3, since this text contains no ablative absolutes.



Text	Precision	Recall	F-Score
Caesar (Commentarii de Bello Gallico)	80.00%	47.06%	59.26%
Cicero (In Catilinam)	57.14%	80.00%	66.67%
Ovid (Metamorphose*s*)	47.37%	62.07%	53.73%
Vergil (Aeneid)	30.00%	50.00%	37.50%
Petronius (Satyricon)	56.36%	56.36%	56.36%
Propertius (Elegies)	44.44%	66.67%	53.33%
Sallust (Bellum Catilinae)	55.56%	53.85%	54.69%

Table 3: Accuracy of identifying ablative absolutes in unannotated texts



Text	Precision	Recall	F-Score
Caesar (Commentarii de Bello Gallico)	100.00%	100.00%	100.00%
Cicero (In Catilinam)	85.71%	61.54%	71.64%
Jerome (Vulgata)	95.24%	90.91%	93.02%
Ovid (Metamorphoses)	66.67%	44.44%	53.33%
Vergil (Aeneid)	100.00%	75.00%	85.71%
Petronius (Satyricon)	95.83%	63.89%	76.67%
Propertius (Elegies)	100.00%	83.33%	90.91%
Sallust (Bellum Catilinae)	5.56%	50.00%	10.00%

Table 4: Accuracy of identifying *cum* clauses in unannotated texts

Figure 4 shows the results of counting *cum* clause and ablative absolute frequencies in a variety of texts. Frequencies were calculated as: (the number of constructions identified / the number of words in the text segment). Each point represents the frequencies in a text segment. For Caesar, Tacitus, and Cicero, each point represents a book of the specified work. For Seneca, each point represents a complete essay or a fragment of an essay (*De Brevitate Vitae*, *De Ira*, and *De Clementia*). For Sallust, the *Bellum Catilinae* was divided into 3 segments of equal length.

These frequencies are normalized for text length, but the length of each text fragment varied by author. Chapters of Cicero's *De Oratore* contain as many as 26,865 words, while fragments of

Tacitus's *Annales* contain as few as 526 (Book 5) words. In general, the texts group by author. The Cicero texts, both from *In Catilinam* and *De Oratore* are clustered in the same group, as are all of the chapters of Tacitus's *Annales*. The sections of Sallust are more varied in their frequencies of ablative absolutes, but all have very few cum clauses. Similarly, the books of Caesar's *Commentarii de Bello Gallico* vary in their frequencies of ablative absolutes, but are more consistent in their frequencies of *cum clauses*, with the exception of Book 8 (Figure 4b). Book 8 is an outlier in this data set, falling 1.83 IQRs beyond the 3rd quartile. Figure 4c, a more detailed view of the books of Tacitus's *Annales*, suggests these books are more similar than the books of Caesar's *Commentarii de Bello Gallico*.

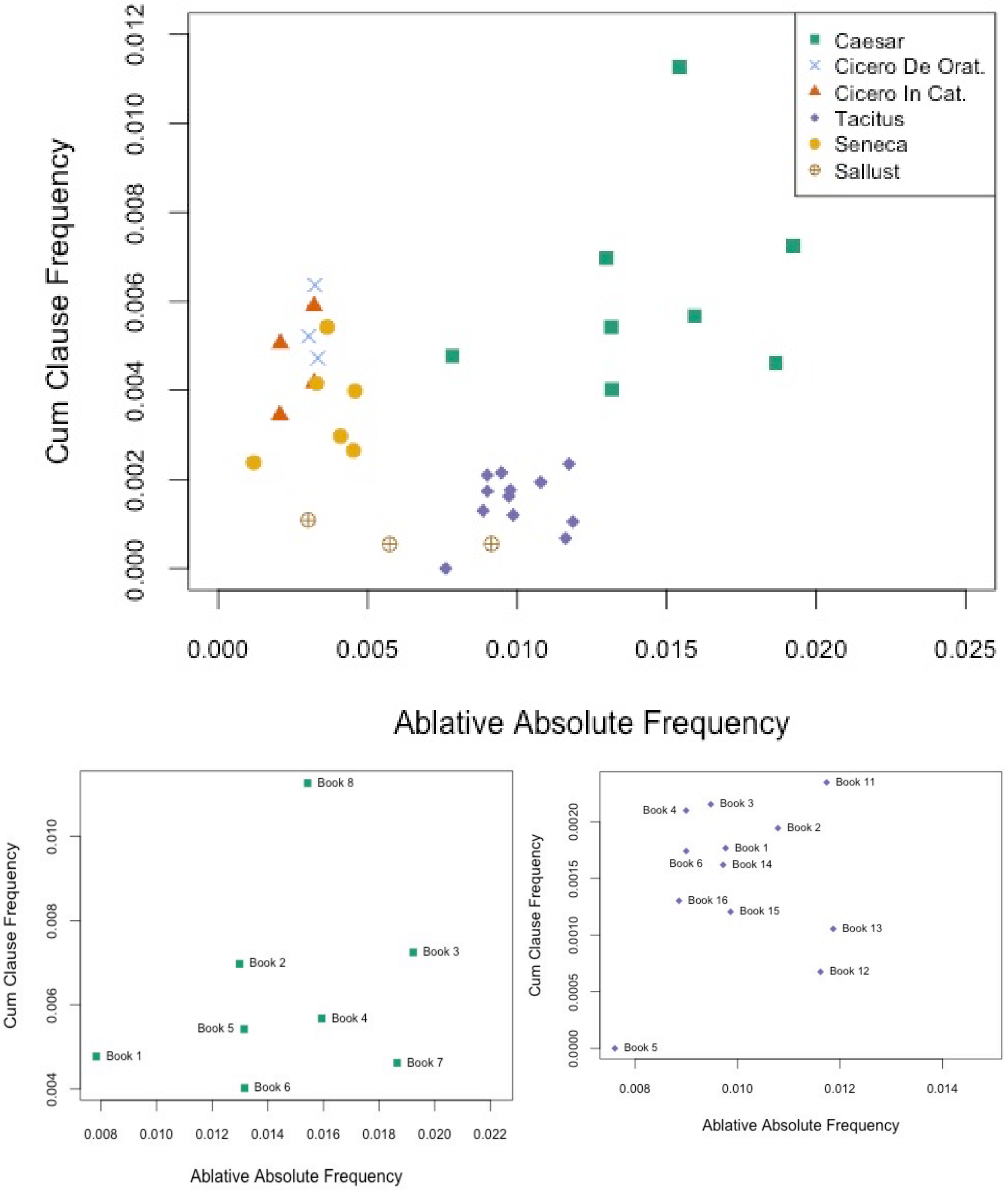


Figure 4: A comparison of *cum* clause frequencies and ablative absolute frequencies for a variety of authors; top (a), frequency rates in all authors; bottom left (b), frequency rates in Caesar; bottom right (c), frequency rates in Tacitus

Figure 5 offers a different perspective on the relationship between ablative absolutes and *cum* clauses in Caesar and Tacitus. We show what percentage of the clauses identified as either ablative absolutes or *cum* clauses were identified as ablative absolutes (i.e. number ablative absolutes identified / total number of clauses identified * 100). Figure 5a (Caesar) shows an increase in the percentage of ablative absolutes across Books 1 to 7, with a sharp decrease in Book 8. Figure 5b (Tacitus) shows no clear progression in the usage of ablative absolutes vs. *cum* clauses.

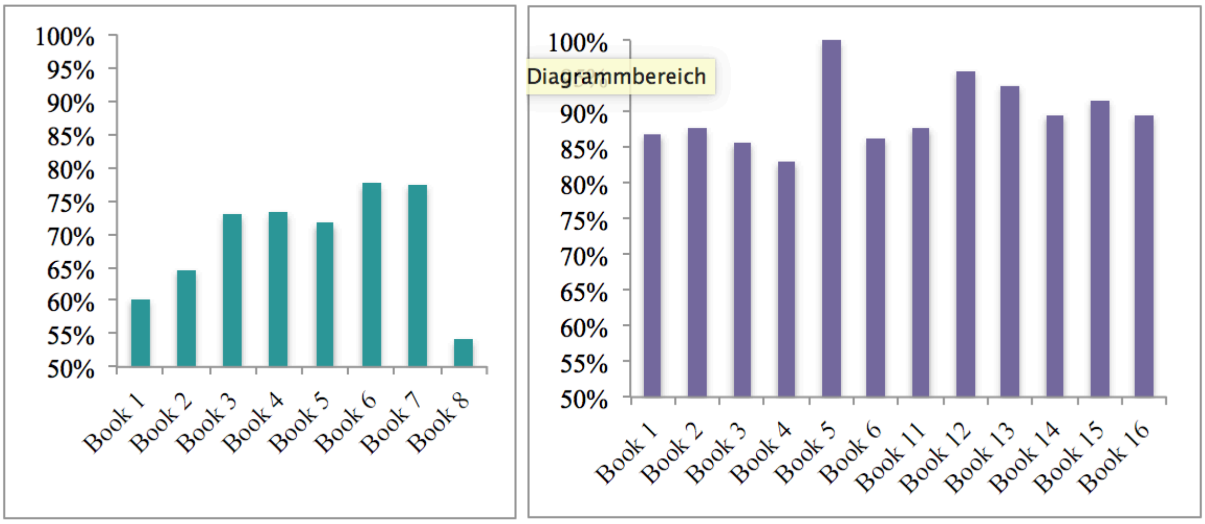


Figure 5: Percentage of ablative absolutes out of identified clauses; left, in Caesar (a); right, in Tacitus (b)

5. Discussion

In the following sections we discuss our main results.

5.1 Differentiation of Book 8 of *Commentarii de Bello Gallico*

Figures 1, 2, and 4 all demonstrate how our method for identifying syntactic constructions can distinguish between texts. In particular, Figures 4b and 4c suggest that our method identifies stylistic differences, rather than just genre or content differences. One of the classic problems in stylistic analysis, and especially in authorship attribution, is the tendency to extract features representative of the text's content, rather than of the author's style.[17] However, examining books within the same works, namely Caesar's *Commentarii de Bello Gallico* and Tacitus's *Annales*, minimizes any content differences.

Figures 4b and 5a show a clear difference between Book 8 of Caesar's *Commentarii de Bello Gallico* and the rest of the work, specifically because Book 8 contains a much higher frequency of *cum* clauses than any other book. This distinction is somewhat expected, as Book 8 was not written by Caesar. Instead, it is attributed to one of his officers, Aulus Hirtius. By counting syntactic constructions, it is possible to determine that Book 8 of the Commentarii de Bello Gallico is quite different from the other books, and even without further information, this difference raises the question of authorship. These results suggest that Aulus Hirtius has

[17] Gamon (2004).

a unique style distinguishable from Caesar and possibly distinguishable from other historians of his time. Other scholars have also observed a difference between Hirtius's section of the work and Caesar's sections. Kathryn Welch notes, "Book 8 reveals proportionally more about the legates," and calls it "arguably the most boring book in the Caesarian Corpus."[18] Closer examination of Book 8 and comparison with works of disputed authorship, like the *Alexandrine, African*, and *Spanish Wars*, could help determine whether or not these disputed works were written by Aulus Hirtius, as some scholars believe.[19] In analyzing the style and language of *Bellum Alexandrinum*, Gaertner and Hausburg observe a generally heterogeneous style and also mention the usage of subordinate conjunctions, including the use of cum + subjunctive instead of *post(ea)quam* in certain sections.[20]

Although Book 8 offers the clearest example of how our syntax-based analysis can be applied to open questions of authorship and style, the trends in other books also reflect previously observed stylistic differences. Much of the scholarship on the style of Caesar's *Commentarii* focuses on the literary nature of the work. While *commentarii* in general were though to follow the same style as *annales*, consisting of a plain recording of events so that other historians could use them as a basis for more ornate works, Caesar's works show more attention to language and style than is thought to be typical of the genre.[21] In particular, some research has shown that the *Commentarii de Bello Gallico* becomes more literary over course of the work, straying further and further from the expected style. First, most scholars agree that Book 1 is considerably different from the rest of the work. J. J. Schlicher describes it as: "a book of argument as much as it is a book of war and conquest," and Kathryn Welch notes, "the legates have little or no role in its action".[22] Figures 4b and 5a show that Book 1 contains ablative absolutes the least frequently, and even strays close to the styles of Seneca and Cicero. The ablative absolute is a very succinct construction, useful primarily for narrating events concisely. The highly rhetorical nature of Book 1 can explain why the usage of ablative absolutes is so low, especially as compared with other books.

Schlicher further analyzes the development of Caesar's style, claiming that it becomes more periodic over the course of *Commentarii de Bello Gallico* as Caesar uses participles in place of subordinate clauses. Gotoff also observes how Caesar uses the ablative absolute to facilitate a periodic style, though argues that Caesar's style is reasonably consistent across the work.[23] In contrast, Eden and later Kraus agree with Schlicher, commenting on the heterogeneous style of the *Commentarii de Bello Gallico* and its inability to fit into a traditional genre.[24] Our results show the same progression of style observed by Schlicher and Eden. Specifically, Schlicher counts occurrences of subordinate clauses (temporal or circumstantial), ablative absolutes, and participial phrases, and reports the percentage of each of these 3 constructions out of the total counted clauses, comparable to Figure 5a. His counts show the percentage of ablative absolutes increases from Books 1 to 2 and Books 2 to 3, drops slightly from Books 3 to 4, drops more significantly from Books 4 to 5, and then increases for Books 6 and 7. As his analysis focuses on Caesar, he does not report counts for Book 8. Although we focus only on the ablative abso-

[18] Welch (1998).

[19] Daly (1951).

[20] Gaertner / Hausburg (2013).

[21] Eden (1962).

[22] Welch (1998); Schlicher (1936).

[23] Gotoff (1984).

[24] Eden (1962); Kraus (2005).

lute and one type of subordinate clause, the *cum clause*, we see almost the exact same trend in Figure 5a. Furthermore, we see a large drop in the percentage of ablative absolutes from Book 7 to Book 8, showing how Hirtius's book does not fit the progression of Caesar's style. Overall, our automated method was able to identify the same increase in the usage of ablative absolutes over subordinate clauses as Schlicher's hand analysis.

5.2 Consistency of Tacitus's Syntax in *Annales*

While Figures 4 and 5 show a range in Caesar's usage of ablative absolutes and *cum* clauses, Tacitus's style remains fairly consistent across *Annales*. As with Caesar, scholars have debated the development of Tacitus's style over time and within *Annales*. Most agree that his style from his earlier works, like *Histories*, to his later works, like *Annales*, becomes more compressed, stronger, and more „Tacitean", but some authors claim that the final books of *Annales* regress to a more „Ciceroian" style.[25] F.R.D. Goodyear questions this claim, suggesting that Books 13–16 might have some unusual vocabulary, but that Tacitus's overall style and syntax remain relatively consistent. Goodyear examines some specific markers, focusing on lexical measures like frequencies of certain adjectives and prepositions, but he suggests that a closer examination of the ablative absolute might offer further insight into the consistency of Tacitus's style.[26] Figure 4c reveals little difference between the final books of Tacitus's *Annales* (Books 13–16) and the rest of the work. While Book 5 has an unusually low number of *cum* clauses, and Books 11–13 have higher numbers of ablative absolutes, Books 13-16 form no group distinguishable from the rest of the work. The unusual syntax in Book 5 likely occurs because this book has only survived as a fragment and contains less than 600 words. Thus, this fragment is too small to accurately demonstrate Tacitus's style. Overall, these data support Goodyear's claim, that there is evidence of continuous style between the final books of the *Annales* and the rest of the work.

5.3 Syntax Usage Varies Across Genres

Although, the use of syntactic constructions varies enough within *Commentarii de Bello Gallico* to distinguish features of different books, the variation greatly increases across authors and genres. The hand-annotated data, which can be considered highly accurate, shows a high frequency of ablative absolutes in the works of Caesar and much lower frequencies in the works of Vergil and Cicero. The ablative absolute is commonly associated with the style of military reports. Adams finds ablative absolutes in Plautus's parodies of such reports and notes their frequency in texts that summarize military events.[27] Caesar's *Commentarii de Bello Gallico* falls into this category, as it relates the events of the Gallic Wars, essentially a military history. Although Caesar's style varies within the work, the frequent use of the ablative absolute demonstrates Caesar's overall adherence to the normal style of military descriptions, rather than a more rhetorical or poetic style, as in Cicero's *In Catilinam* or Vergil's *Aeneid*, where ablative absolutes are scarce. While the ablative absolute's ability to convey information concisely

[25] Löfstedt (1948).

[26] Goodyear (1968).

[27] Adams (2005).

makes it useful for military reports and historical accounts, such utilitarian language does not belong in poetry or stylized prose.

Analysis of the unannotated data presented in Figure 4 confirms the same trend; specifically Caesar and Tacitus use ablative absolutes more frequently than Cicero and Seneca. Authors like Seneca and Cicero prefer to take more time to express ideas, especially in orations, since the speaker wants to give the listener time to process information. A.D. Leeman has observed the different usage of ablative absolutes in Caesar and Cicero, and he estimates that Caesar uses about 10 times as many ablative absolutes as Cicero.[28] While the ratio in these data is closer to 5:1, the difference in usage is still clear. By counting the frequencies of ablative absolutes in a range of texts, we are able to systematically observe the usage patterns identified by other scholars and to generalize them across authors.

5.4 Variation of Syntactic Constructions Among Historians

Furthermore, our method highlights deviations from these trends. Although the frequency of ablative absolutes generally distinguishes between the plainer style of military and historical accounts and the more ornate style of philosophy and orations, Sallust, a historian, defies the pattern by using far fewer ablative absolutes than either Tacitus or Caesar. The deviation that our method detects coincides with theories about Sallust's motivations and style. Sallust wrote *Bellum Catilinae* earlier in his life than when most historians begin writing, and some have speculated that he had ulterior motives in writing the work, more than just recording history for future generations. Scholars have observed some peculiarities in his style, including his tendency to use the historical infinitive where most authors would use the imperfect tense.[29] More generally, his style is also thought to be especially poetic and paratactic.[30] Parataxis is a writing form that uses short parallel sentences, rather than nested clauses and subordination. Our results reveal Sallust's tendency to use few *cum* clauses and few ablative absolutes, which could reflect a more general avoidance of subordinate clauses as a result of a paratactic style.

When comparing Sallust with Tacitus and Caesar, who both use ablative absolutes frequently, the difference between Caesar and Sallust is not wholly unexpected. Previous scholars have observed the high frequency of ablative absolutes in Caesar as compared to Sallust.[31] However, the difference between Sallust and Tacitus is less expected, since Tacitus's style is often thought to be Sallustian.[32] The lack of similarity demonstrates that ablative absolutes and *cum* clauses are a very small subset of an author's style. Although Sallust and Tacitus differ in the use these particular constructions, they may have similarities in other aspects of their styles, such as vocabulary choice or other syntax.

We can further compare Caesar and Tacitus. There is conflicting literature on how much these authors differ in their use of ablative absolutes. Leeman observes more ablative absolutes in Caesar than in Tacitus.[33] In contrast, J.N. Adams claims that the descriptions of battles in

[28] Leeman (1963).

[29] Von Albrecht (1979).

[30] Leeman (1963).

[31] Von Albrecht (1979); Leeman (1963).

[32] Goodyear (1968).

[33] Leeman (1963).

Tacitus use the language of military reports, as indicated by the frequency of ablative absolutes, and even compares Tacitus's militaristic style in battle scenes to Caesar's works.[34] From Figure 4, the high frequency of ablative absolutes we find in Tacitus lends support to Adams's claim. Nevertheless, a clear distinction occurs between Tacitus and Caesar, since Caesar uses *cum* clauses much more frequently than Tacitus.

5.6 Stylistic Implications of Varied Accuracy Rates

Although our method was able to identify syntactic constructions with enough accuracy to detect patterns in usage, accuracy remains a limiting factor in analyzing the unannotated data. However, although the accuracy of identifying constructions was very low for some authors, these accuracy rates are also a reflection on the author's style. For example, the imperfect recall of *cum* clauses in Cicero in Table 4 largely occurs because of 1 particular sentence: *cum arma, cum securis, cum fascis, cum tubas, cum signa militaria, cum aquilam illam argenteam…scirem esse praemissam,* which translates: "When I knew that arms, that the axes, the fasces, and trumpets, and military standards and that silver eagle...had been sent on?" (Cicero, In Catilinam 2.6, Translator C. D. Yonge).

This sentence contains a series of 6 *cum*-noun pairings. Unsurprisingly, TreeTagger tags all 6 of these *cums* as prepositions, which seems natural, since they are all in self-contained clauses followed by nouns. However, on closer inspection, this sentence actually consists of a series of parallel clauses, in which Cicero repeats the conjunction *cum* with each noun in the sentence and omits a verb. Because applying TreeTagger to this sentence results in all 6 *cums* tagged as prepositions instead of conjunctions, this one construction greatly contributes to the error rate of identifying *cum* clauses in Cicero. The repeated conjunction translates awkwardly into English, but this type of construction is not uncommon in Latin. Cicero in particular uses such repetition frequently, and similar constructions occur throughout *In Catilinam*. The first section alone contains two examples, where Cicero repeats the word *nihil* and then the word *quid* (Cicero, In Catilinam 1.1). In this way, the accuracy of identifying syntactic constructions reflects Cicero's style just as much as the actual construction identified.

The relationship between style and accuracy of identifying syntactic constructions becomes more apparent by looking at a different stylistic element: non-projectivity. The non-projectivity rate refers to how often constituents of a phrase are broken up by other constituents. For example, Vergil writes, *Troiae qui primus ab oris* (Vergil, Aeneid, 1.1), breaking up the phrase "Trojan shore" by separating the words *Troiae* and *oris*. High rates of non-projectivity can make text analyses, such as parsing, more difficult.[35] Bamman and Crane calculate the non-projectivity rates for some of the text segments in this data set, displayed in Table 5. For reference, the non-projectivity rate in Swedish is approximately 0.94% and in Czech is approximately 1.81%.[36]

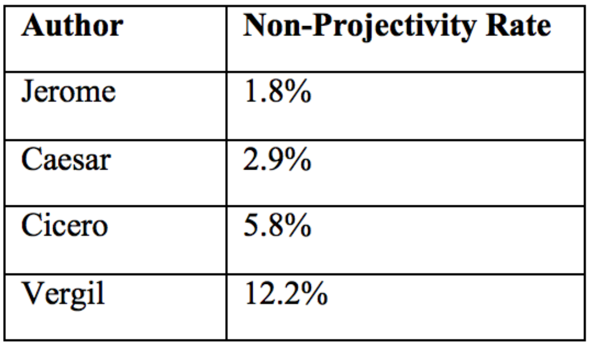
Both Jerome and Caesar write in fairly straightforward prose. They have low non-projectivity rates, and syntactic analysis was very accurate in both authors. In contrast, Cicero's *In Catilinam* has a high rate of non-projectivity, reflecting the deeply stylized nature of Roman oratory. Similarly, the poet Vergil also has a high rate of non-projectivity. It seems logical that poetry inherently involves manipulation of word order, because poets must arrange their verse to fit

[34] Adams (1973).

[35] Nivre / Nilsson (2005).

[36] Bamman / Crane (2006).

the meter. High rates of non-projectivity have been observed in Ancient Greek poetry as well, which suggests that discontinuous constituents could be an expected feature of poetry in a free word order language.[37] The accuracy rates for identifying syntactic constructions in both Cicero and Vergil are lower than the accuracy rates for Jerome and Caesar, suggesting that non-projectivity could influence the accuracy of these methods. Specifically, a low accuracy implies high non-projectivity.



Author	Non-Projectivity Rate
Jerome	1.8%
Caesar	2.9%
Cicero	5.8%
Vergil	12.2%

Table 5: Non-projectivity rates in segments of hand-annotated texts

Although the similarities between constructions identified in the annotated data and the unannotated suggest our method does reflect true syntax usage in unannotated texts, accurate identification of syntactic constructions is not strictly necessary for distinguishing between the styles of various authors. Even if the texts written by Cicero seem to have a low rate of ablative absolutes simply because identifying ablative absolutes in Cicero is difficult, the fact the finding ablative absolutes in Cicero is difficult reflects unique elements about Cicero's style. How well syntactic analysis (specifically parsing) works on different texts has been used as a feature in authorship attribution studies.[38]

6. Related Work

The main advantage of our method for syntactic analysis is its applicability to unannotated texts without automated parsing. The concept of a syntax-based method for stylistic analysis is not new. Baayen et al. developed a method for authorship attribution that focused on syntactic rewrite rules and resulted in higher accuracy than word-based methods. However, their method was only tested on annotated texts.[39] Similarly, Bamman, Passarotti and Crane analyzed Latin syntax change over time, specifically in the shift from an *Accusativus cum Infinitivo* (ACI) construction to *quia/quod* clauses. The study was able to identify a shift in usage by examining two sets of hand-annotated data: the Latin Dependency TreeBank (LDT), consisting of classical Latin, and Index Thomisticus (IT-TB) consisting of the works of Thomas Aquinas, written about 13 centuries later.[40] That study was very similar to this one in that it focused on counting specific grammar constructions in hand-annotated data and comparing their frequencies across different time periods. However, because our method was not limited to annotated data, we were able to examine a broader range of texts, comparing constructions across different authors and genres, not just different time periods.

Other studies have examined unannotated texts but require automated parsing. Stamatatos et al. propose a computer-based method for authorship attribution that uses low-level markers, like sentence boundaries, and syntactic-level markers, like noun phrase counts. Although, this method was able to distinguish between authors of Modern Greek, which is also a highly in-

[37] Mambrini / Passarotti (2013).

[38] Stamatatos et. al. (2001).

[39] Baaeyn et al. (1996).

[40] Bamman et al. (2008).

flected language with variable word order, it requires constituent parsing.[41] Not only is Latin parsing difficult, studies involving Latin parsing typically focus on dependency parsers, in which words are linked to their immediate head, over constituent parsers, in which words are grouped in phrasal categories. Bamman and Crane used a set of 30K hand-annotated words to train a Latin dependency parser and Lee, Naradowsky, and Smith used an expanded version of this data set (53K) to train a combined approach to morphological and syntactic tagging.[42] Neither method achieved an accuracy rate greater than 65%.

However, Bamman and Crane were still able to use the tags generated by their parser to extract valuable information about selectional preferences. Breaking down the accuracy of Bamman and Crane's parser, their precision rates ranged from 34% to 68%, and their recall rates ranged from 27% to 71% for tagging relationships between words. These accuracy rates are comparable with the precision and recall rates of our method for syntactic construction identification (Tables 3 and 4).[43]

7. Conclusions

Comparison between constructions in the hand-annotated data and in the unannotated data suggests that the methods proposed in this study are accurate enough to facilitate a syntax-based analysis of classical Latin. Furthermore, the distribution of ablative absolutes and *cum* clauses identified in various authors is generally consistent with past analyses of classical Latin. This consistency confirms some the observations of scholars who examined these texts, including the frequent use of ablative absolutes in history and military accounts and the infrequent use of ablative absolutes in more ornate prose.

A more in depth analysis of specific authors, Caesar in particular, also reflects observations about these authors and contributes evidence to open debates about their style, such as how Caesar's style changes throughout the *Commentarii de Bello Gallico*. Similarly, the analysis of Caesar's *Commentarii de Bello Gallico* demonstrates that our methods can help resolve questions of authorship attribution. Our methods were able to distinguish between Book 8 of *Commentarii de Bello Gallico*, which was not written by Caesar, and the rest of the work. Overall, the consistency with manual research affirms the usefulness of these methods, indicating that they are accurate enough to contribute to the study of classical literature.

More generally, this study demonstrates that an automated syntax-based analysis of Latin is both useful and possible. Analysis of specific constructions can distinguish between the style of different authors. Furthermore, unlike traditional lexically based measures, such as word-frequencies or n-gram frequencies, this sort of analysis can target constructions that classicists are interested in studying. Automatic identification of syntax can be applied to existing literature to help answer questions that classicists have been asking for centuries.

[41] Stamatatos et al. (2001).

[42] Bamman / Crane (2008), Lee et al. (2011).

[43] Bamman / Crane (2008).

8. References

Adams (1973): J.N. Adams, "The Vocabulary of the Speeches in Tacitus' Historical Works", Bulletin of the Institute of Classical Studies 20, 124–144.

Adams (2005): J.N. Adams, "The Bellum Africum" in: Adams, J. N., Lapidge, M., and Reinhardt, T. (Hrsg.), Aspects of the Language of Latin Prose, Oxford/New York, 73–96.

Albrecht (1979): M. V. Albrecht, Masters of Roman Prose: from Cato to Apuleius, Francis Cairns.

Baaeyn et al. (1996): H. Baayen / H. Van Halteren / F. Tweedie, "Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution", Literary and Linguistic Computing 11, 121-132.

Bamman / Crane (2006): D. Bamman / G. Crane, "The Design and Use of a Latin Dependency Treebank", in: Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT, 2006), Prague, 67–78.

Bamman / Crane (2008): D. Bamman / G. Crane, "Building a Dynamic Lexicon from a Digital Library", in: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (ACM, 2008), New York, 67–78.

Bamman et al. (2007): D. Bamman / M. Passarotti / G. Crane / S. Raynaud, "Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3)", Technical report. Medford: Tufts Digital Library.

Bamman et al. (2008): D. Bamman / M. Passarotti / G. Crane, "A Case Study in Treebank Collaboration and Comparison: Accusativus cum Infinitivo and Subordination in Latin", The Prague Bulletin of Mathematical Linguistics 90, 109–122.

Bennett (1918): C.E. Bennett, A New Latin Grammar, Allyn and Bacon.
G. Celano / G. Crane / G. Almas et al, "The Ancient Greek and Latin Dependency Treebanks", https://perseusdl.github.io/treebank_data/, (Accessed 2015).

Covington (1990): M. Covington, "A Dependency Parser for Variable-Word-Order Languages", Technical Report AI-1990-01, Artificial Intelligence Programs, The University of Georgia Athens.

Daly (1951): L. Daly, "Aulus Hirtius and the Corpus Caesarianum", The Classical Weekly 44, 113–117.

Diederich et al. (2003): J. Diederich / J. Kindermann / E. Leopold / G. Paass, "Authorship Attribution with Support Vector Machines", Applied Intelligence 19, 109–123.

Eden (1962): P. T. Eden, "Caesar's Style: Inheritance versus Intelligence, Glotta 40, no. ½, 74–117.

Gaertner / Hausburg (2013): J. Gaertner / B. Hausburg, Caesar and the Bellum Alexandrinum Göttingen.

Gamon (2004): M. Gamon, “Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features”, in: Proceedings of the 20th Annual Conference on Computational Linguistics (ACL, 2004), Morristown, NJ, 611–617.

Goodyear 1968: F.R.D. Goodyear, “Development of Language and Style in the Annals of Tacitus”, The Journal of Roman Studies 58, 22–31.

Gotoff (1984): H.C. Gotoff, „Towards a practical criticism of Caesar’s prose style“, Illinois Classical Studies 9.1, 1–18.

Grethlein (2006): J. Grethlein, “The Unthucydidean Voice of Sallust”, Transactions of the American Philological Association 136, 299–327.

Koch (1994): U. Koch, “The Enhancement of a Dependency Parser for Latin”, Research Report AI-1993-03, Artificial Intelligence Programs, The University of Georgia Athens.

Koster (2005): C. Koster, “Constructing a parser for Latin”, Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Berlin – Heidelberg, 48–59.

Kraus (2005): C. Kraus, “Hair, Hegemony, and Historiography: Caesar’s Style and its Earliest Critics”, Proceedings-British Academy, Vol. 129. Oxford University Press Inc.

Lee et al. (2011): J. Lee / J. Naradowsky / D. Smith, “A Discriminative Model for Joint Morphological Disambiguation and Dependency Parsing”, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (ACL, 2011), 885–894.

Leeman (1963): A.D. Leeman, Oratonis Ratio: The Stylistic Theories and Practice of the Roman Orators Historians and Philosophers, A.M. Hakkert.

Löfstedt (1948): E. Löfstedt, “On the Style of Tacitus”, The Journal of Roman Studies 38, 1–8.

Mambrini / Passarotti (2013): F. Mambrini / M. Passarotti, “Non-projectivity in the Ancient Greek Dependency Treebank”, in: Proceedings of the Second International Conference on Dependency Linguistics (DepLing, 2013), Prague, 177–186.

Mosteller / Wallace (1964): F. Mosteller / D. L. Wallace, Inference and Disputed Authorship: The Federalist, Addison-Wesley.

Moreland / Fleischer (1990): F.L. Moreland / R.M. Fleischer, Latin: An Intensive Course, University of California Press.

Nivre / Nilsson (2005): J. Nivre / J. Nilsson, “Pseudo-projective Dependency Parsing”, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL, 2005), Ann Arbor, Michigan, 99–106.

Passarotti / Dell’Orletta (2010): M. Passarotti / F. Dell’Orletta, “Improvements in Parsing the Index Thomisticus Treebank. Revision, Combination and a Feature Model for Medieval Latin”, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC, 2010), Valletta, Malta, 1964–1971.

"The Perseus Digital Library", http://www.perseus.tufts.edu/hopper/ (Accessed 2015).

Schlicher (1936): J.J. Schlicher, „The development of Caesar's narrative style", Classical Philology 31.3, 212–224.

Schmid (1994): H. Schmid, "Probabilistic Part-of-speech Tagging Using Decision Trees", in: Proceedings of the International Conference on New Methods in Language Processing, 44–49.

H. Schmid, "TreeTagger - A Part-of-speech Tagger for Many Languages", http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Accessed 2015).

Stamatatos et al. (2001): E. Stamatatos / N. Fakotakis / G. Kokkinakis, "Computer-based Authorship Attribution without Lexical Measures", Computers and the Humanities 35, 193–214.

Stamatatos (2009): E. Stamatatos, "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for Information Science and Technology 60, 538–556.

Welch (1998): K. Welch, „Caesar and his officers in the Gallic War commentaries", Julius Caesar as Artful Reporter, 85–110.

Autorenkontakt[44]

Anjalie Field
Princeton University
Department of Computer Science

Email: aefield@alumni.princeton.edu

[44] Die Rechte für Inhalt, Texte, Graphiken und Abbildungen liegen, wenn nicht anders vermerkt, bei den Autoren.