

The Treebanked Conspiracy. Actors and Actions in *Bellum Catilinae*

Marco Passarotti

CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 - 20123 Milan, Italy
marco.passarotti@unicatt.it

Berta González Saavedra

Dep. de Filología Griega y Lingüística Indoeuropea
Universidad Complutense de Madrid
Pl. M. Pelayo - 28040 Madrid, Spain
bergonza@ucm.es

Abstract

In the context of the *Index Thomisticus* Treebank project, we have recently enhanced the entire text of *Bellum Catilinae* by Sallust with a layer of semantic annotation. By exploiting the results of semantic role labeling, ellipsis resolution and coreference analysis, this paper presents a study of the main Actors and Actions (and their relations) in *Bellum Catilinae*.

1 Introduction

The large majority of the currently available treebanks includes data taken from contemporary books, magazines, journals and, mostly, newspapers. Such data are used for different purposes in both theoretical and computational linguistics, the most widespread being supporting and evaluating theoretical assumptions with empirical evidence and providing data for various tasks in stochastic NLP, like inducing grammars and training/testing tools.

Across the last decade, a small, but ever growing, bunch of dependency treebanks for ancient languages was built. In this respect, the main treebanks now available are those for Latin and Ancient Greek, with The Ancient Greek and Latin Dependency Treebank (AGLDT) (Bamman and Crane, 2011), the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2011) and the PROIEL corpus (Haug and Jøhndal, 2008).

Treebanks for ancient languages tend to include literary, historical, philosophical and/or documentary texts. This makes the very use of such resources different from that of treebanks for modern languages. Indeed, instead of exploiting data to draw linguistic generalizations, users of such treebanks are more interested in the linguistic features of the texts themselves available in the corpus. For instance, there is more interest and scientific motivation in exploiting the treebanked texts of Sophocles to study their specific syntactic characteristics than in using the evidence provided by such texts as sufficiently representative of Ancient Greek, which they are not.

Not only the use of data is different, but also users are. Indeed, it is quite uncommon that scholars from literature, philosophy or history make use of linguistic resources like treebanks for modern languages in their research work. Instead, they represent some of the typical users of treebanks for ancient languages as well as of diachronic treebanks. Such resources become even more useful for this kind of users from the Humanities when they are enhanced also with a semantic layer of annotation, on top of the syntactic one. This is due to the large interest of such scholars in semantic interpretation of texts through syntax.

In this area, the *Index Thomisticus* Treebank project has recently enhanced a selection of texts taken from the IT-TB and the AGLDT with semantic annotation. This paper describes the dependency-based annotation style applied on these data and presents a use case of exploitation of them for literary analysis purposes. In particular, the analysis focuses on the main Actors and Actions in Sallust's *Bellum Catilinae*.¹ The work is performed by using the results of semantic role labeling, coreference analysis and ellipsis resolution applied on the source data.

¹Written probably between 43 and 40 BCE, *Bellum Catilinae* tells the story of the so called second Catilinarian conspiracy (63 BCE), a plot, devised by Catiline and a group of aristocrats and veterans, to overthrow the Roman Republic. The text of *Bellum Catilinae* available from the AGLDT is the one edited by Ahlberg (1919). It includes 10,936 words and 701 sentences. In this paper, English translations of *Bellum Catilinae* are taken from Ramsey (2014).

2 Data

In the context of the *Index Thomisticus* Treebank project hosted at the CIRCSE research centre of the Università Cattolica del Sacro Cuore in Milan, Italy (<http://itreebank.marginalia.it/>), we have added a new layer of semantic annotation on top of a selection of syntactically annotated data taken from the IT-TB and the Latin portion of the AGLDT (González Saavedra and Passarotti, 2014).

In particular, around 2,000 sentences (approx. 27,000 words) were annotated out of *Summa contra Gentiles* of Thomas Aquinas (IT-TB). The entire *Bellum Catilinae* of Sallust (BC) and small excerpts of 100 sentences each from texts of Caesar and Cicero were annotated from the AGLDT.

2.1 Annotation Style

The style of the semantic layer of annotation used in the IT-TB project is based on Functional Generative Description (FGD) (Sgall et al., 1986), a dependency-based theoretical framework developed in Prague and intensively applied and tested while building the Prague Dependency Treebank of Czech (PDT) (Hajič et al., 2000).

The PDT is a dependency-based treebank with a three-layer structure. The (so ordered) layers are a “morphological layer” (morphological tagging and lemmatization), an “analytical” layer (annotation of surface syntax) and a “tectogrammatical” layer (annotation of underlying syntax). Both the analytical and the tectogrammatical layers describe the sentence structure with dependency tree-graphs, respectively named analytical tree structures (ATs) and tectogrammatical tree structures (TGTS).

In ATs every word and punctuation mark of the sentence is represented by a node of a rooted dependency tree. The edges of the tree correspond to dependency relations that are labelled with (surface) syntactic functions called “analytical functions” (like Subject, Object etc.).

TGTS describe the underlying structure of the sentence, conceived as the semantically relevant counterpart of the grammatical means of expression (described by ATs). The nodes of TGTS include autosemantic words only (represented by “tectogrammatical lemmas”: “t-lemmas”), while function words and punctuation marks collapse into the nodes for autosemantic words. Semantic role labeling is performed by assigning to nodes semantic role tags called “functors”. These are divided into two classes according to valency: (a) arguments, called “inner participants”, i.e. obligatory complementations of verbs, nouns, adjectives and adverbs: Actor,² Patient, Addressee, Effect and Origin; (b) adjuncts, called “free modifications”: different kinds of adverbials, like Place, Time, Manner etc.

Also coreference analysis and ellipsis resolution are performed at the tectogrammatical layer and are represented in TGTS through arrows (coreference) and newly added nodes (ellipsis). In particular, there are two kinds of coreference: (a) “grammatical coreference”, in which it is possible to pinpoint the coreferred expression on the basis of grammatical rules (mostly with relative pronouns) and (b) “textual coreference”, realized not only by grammatical means, but also via context (mostly with personal pronouns).

2.2 From ATs to TGTS

The workflow for tectogrammatical annotation in the IT-TB is based on TGTS automatically converted from ATs.³ The TGTS that result from the conversion are then checked and refined manually by two annotators. The conversion is performed by adapting to Latin a number of ATs-to-TGTS conversion modules provided by the NLP framework *Treex* (Žabokrtský, 2011).⁴

For instance, Figure 1 shows the ATs for the sentence “cum [with] eo [him] se [himself] consul[em] [consul] initium [beginning] agundi [of acting] facturum [would have made]” (BC 21.4) (“[Catiline promised that] as consul with him, he would launch his undertaking”), which presents a case of predicate

²The definition of Actor in the PDT is semantically quite underspecified, as it refers to “the human or non-human originator of the event, the bearer of the event or a quality/property, the experiencer or possessor” (Mikulová et al., 2006, page 461).

³The guidelines for analytical annotation of the IT-TB (as well as of the Latin portion of the AGLDT) are those of Bamman et al. (2007). The guidelines for tectogrammatical annotation are those of the PDT (Mikulová et al., 2006), with a few modifications for representing Latin-specific constructions.

⁴See González Saavedra and Passarotti (2014) for details on ATs-to-TGTS conversion in the IT-TB and, especially, for an evaluation of the accuracy of the conversion process.

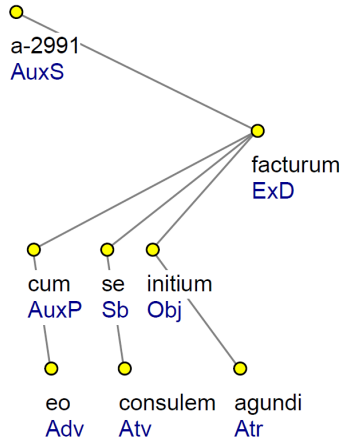


Figure 1: ATS of BC 21.4.

ellipsis. The sentence is an objective subordinate clause lacking the predicate of its governing clause (“[Catiline promised that]”). In ATSs, this is represented by assigning the analytical function ExD (External Dependency) to the main predicate of the sentence. In the ATS of Figure 1, the node for *facturum* is assigned ExD, because here *facturum* depends on a node that is missing and, thus, it is “external” to the current tree.

Figure 2 shows the TGTS for this sentence. The TGTS in Figure 2 resolves the ellipsis of the main clause. Three sentences before this one in the text, Sallust writes “Catiline polliceri” (“Catiline promised [to men]”). The sentence in BC 21.4 still depends on this clause. Once resolved the ellipsis of *polliceor*, the TGTS must represent its arguments. Among these, both the Actor and the Addressee result from ellipsis resolution: Catiline is the Actor and the men (*homo*) are the Addressee. The Patient of *polliceor*, instead, is represented by the entire objective subordinate clause of BC 21.4. In this clause, the Actor is again Catiline, as it is represented by the textual coreference of the node depending on *facio* which is assigned t-lemma #PersPron.⁵ this node is not newly added because it is textually represented by the reflexive pronoun *se*. The Patient of *facio* is *initium*, which is specified by a restrictor (RSTR; the verb *ago*) governing a newly added node for a generic Actor (#Gen). Such Actor is assigned when its denotation cannot be retrieved contextually, which mostly happens when impersonal clauses are concerned, like in this case (literally: “the beginning of acting”).

The prepositional phrase “cum eo” (“with him”) is represented in the TGTS of Figure 2 by the node for *is* (form *eo*), while that for the preposition *cum* collapses. The personal pronoun *is* is linked with a previous occurrence of the proper name *Antonius* via a textual coreference and it is assigned functor ACMP, which is used for the adjuncts that express manner by specifying a circumstance (an object, person, event) that accompanies (or fails to accompany) the event or entity modified by the adjunct.

In TGTSs, predicative complements (functor: COMPL) are adjuncts with a dual semantic dependency relation. They simultaneously modify a noun and a verb. The dependency on the verb is represented by means of an edge. In Figure 2, this is the edge that connects *facio* with *consul*. The dependency on the noun is represented by means of a specific complement reference, which is graphically represented by a green arrow (going from *consul* to #PersPron in Figure 2).

3 Results and Discussion

One of the added values of tectogrammatical annotation is that it provides information that, although it is accessible to readers, is missing in texts. Looking at the example sentence discussed in the previous section, we see that there is no explicit occurrence of Catiline playing the role of Actor of a verb. Instead, if we exploit tectogrammatical annotation, we can retrieve that actually that sentence says that Catiline performs two different Actions (namely, *polliceor* and *facio*).

⁵#PersPron is the t-lemma assigned to nodes representing possessive and personal pronouns (including reflexives).

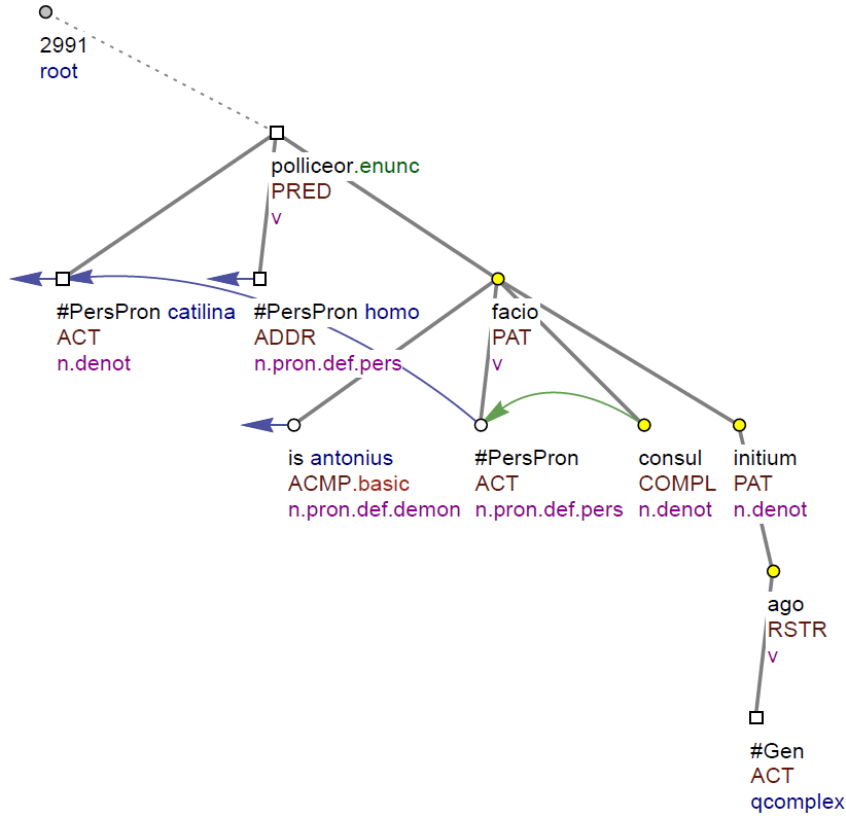


Figure 2: TGTS of BC 21.4.

Tectogrammatical annotation puts us in the condition to answer the basic research question of the work described in this paper: “who does what in *Bellum Catilinae*?”. In other words, what we look for are all the couples Actor-Action in BC regardless of the fact that they do explicitly occur in the text.⁶

3.1 Querying the Data

All data can be freely downloaded from the website of the IT-TB project. The treebanks can be queried through an implementation of the PML-TQ search engine (Prague Markup Language Tree Query) (Štěpánek and Pajas, 2010). We ran a bunch of queries in order to retrieve all the couples Actor-Action in BC. The basic query just searches for all the Actors of a verb:

```
t-node $n0 := [ gram/sempos = 'v' ,
  echild t-node $n1 := [ functor = 'ACT' ] ] ;
```

This query searches for all the nodes of a TGTS (t-node, named \$n0) that are assigned PoS verb (gram/sempos = 'v') and govern either directly or indirectly (echild) a node (\$n1) with functor ACT (functor = 'ACT').⁷ The query does not limit the output to nodes with an explicit textual correspondence, but includes also those newly added in TGTSs, as result of ellipsis resolution.

The output resulting from the query above needs further refinement, as it features several cases of both relative and personal pronouns whose denotation is resolved in TGTSs by coreference analysis. For instance, three Actor-Action couples result from the TGTS of Figure 2: #PersPron-polliceor, #PersPron-facio and #Gen-ago. While #Gen is a generic argument whose denotation cannot be retrieved contextually, both the #PersPron nodes are assigned a textual coreference in the TGTS, thus enabling to replace them with the t-lemma they are coreferent with.

⁶In this work, we consider Actions as represented by verbs only. Deverbal nominalizations are thus excluded.

⁷Direct or indirect government is set in order to retrieve Actors occurring in coordinated constructions (headed by the coordinating element).

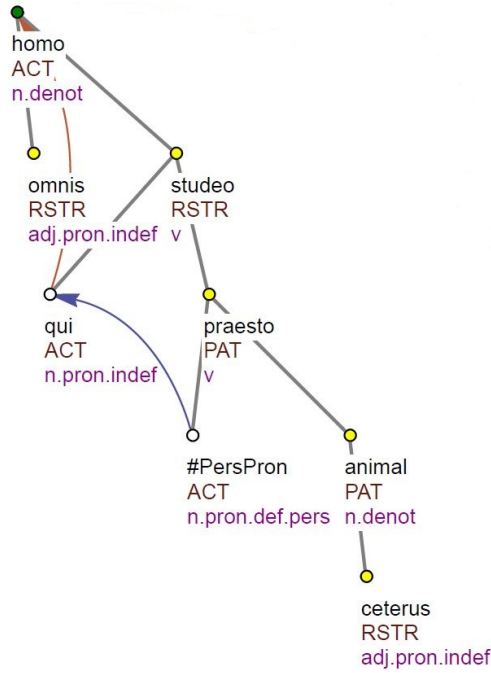


Figure 3: TGTS of BC 1.4 (part).

We ran a number of queries to replace in the output of the basic query all coreferred `#PersPron` t-lemmas with those of the nodes they are linked with via textual coreference. Then we did the same for all coreferred t-lemmas of relative pronouns, which are linked to their antecedent via grammatical coreference.

Not only such queries must consider both direct and indirect linking, as well as textual and grammatical coreference, but they also have to address mixed indirect coreferences. For instance, this is the case of the first noun phrase in the first sentence of BC: “Omnis [all] homines [men], qui [who] sese [themselves] student [be eager] praestare [to stand out] ceteris [others] animalibus [animals] [...]” (BC 1.1) (“All humans who are keen to surpass other animals [...]”). Figure 3 shows the portion of the TGTS for the first sentence of BC concerning this phrase.

From Figure 3, one can see that the denotation (*homo*) of the `#PersPron` node playing the role of Actor of *praesto* is retrieved (a) indirectly, by passing through the node for *qui*, and (b) in mixed fashion, i.e. via a textual coreference (from `#PersPron` to *qui*) plus a grammatical coreference (from *qui* to *homo*).

A model of such kind of complex queries is the following:

```
t-node $n0 := [ functor = 'ACT' ,
               eparent t-node $n2 := [ gram/sempos = 'v' ] ,
               coref.text.rf t-node $n1 := [ coref.gram.rf t-node $n3 := [ ] ] ] ;
```

The t-node named `$n0` is an Actor that depends either directly, or indirectly (`eparent`) on t-node `$n2`, which is a verb. `$n0` has a textual coreference with `$n1`, which in turn has a grammatical coreference with `$n3`.⁸

3.2 Actors and Actions

Tables 1 and 2 report respectively the main Actions and the main Actors in BC. These are defined as the Actions performed by the highest number of different Actors and, conversely, as the Actors that perform the highest number of different Actions.⁹

⁸The longest coreference chain we found in BC includes 5 textual coreferences.

⁹The absence of verbs like *possum* (“can”) and *volo, velle* (“to want”) in Table 1 is due to the treatment of modal predicates in TGTS (see Mikulová et al., 2006, pages 318-320). Not coreferred Actors are excluded from Table 2. These are the generic Actor (`#Gen`) and those pronouns that do not undergo coreference analysis in TGTSs, i.e. indefinite and interrogative pronouns (like *alius* and *quis*), as well as both explicit and generated personal pronouns of first and second person.

Action	Actors	Occ.	Generated
sum	179	268	38
habeo	43	84	10
facio	39	87	4
convenio	20	8	2
dico	18	41	9
do	18	22	3
hortor	16	11	2
venio	14	11	0
coepio	13	18	7
puto	13	10	0
peto	13	12	0
cognosco	13	20	0

Table 1: Main Actions.

Action	Actors	Occ.	Generated
catilina	133	61	6
cicero	33	18	0
homo	32	40	3
res	24	147	4
petreius	20	3	0
lentulus	20	27	6
consul	20	32	0
caesar	20	13	0
populus	19	18	0
curius	19	5	0
vulturcius	18	10	0
vir	18	16	0
animus	18	59	2

Table 2: Main Actors.

Beside Actions and the number of their different Actors, Table 1 reports also the total number of occurrences of each Action and, among these, the number of generated occurrences (resulting from ellipsis resolution). The case of *convenio* (“to come together”) is worth noting, as it turns out that it has 20 different Actors for just 8 occurrences (2 of which are generated). This happens because in some of its occurrences *convenio* has more than one Actor, like for instance in the sentence “eo [there] convenere [to come together] senatorii [senatorial] ordinis [order] P. Lentulus Sura , P. Autronius , L. Cassius Longinus , C. Cethegus , P. et Ser . Sullae Ser. filii , L. Vargunteius , Q. Annius , M. Porcius Laeca , L. Bestia , Q. Curius” (BC 17.3) (“There were present from the senatorial order...”).

Not surprisingly, Catiline is the star of BC, being the Actor of 133 different Actions (i.e. verbs) in 61 occurrences (6 out of which are generated). Traditionally, together with Catiline, the three other main characters of BC are considered to be Caesar, Cato and Cicero, who give the main speeches reported in the text. If we look at the Actions each of them performs and focus on those that Catiline only performs (i.e. those not shared with the others), we can see which Actions are peculiar of Catiline. These are represented by the verbs *dimitto* (“to send out”) and *paro* (“to prepare”).

Interestingly enough, *dimitto* and *paro* not only correspond to the Actions performed by Catiline only (and not also by Caesar, Cato or Cicero), but they are also those Actions that Catiline most frequently performs (6 times), just after *facio* (“to make”) (10) and *habeo* (“to have”) (7), and more than *sum* (“to be”) (5) and *video* (“to see”) (5). If for *dimitto* this result is biased by a case of ellipsis resolution applied on a multiple coordination in one sentence (BC 27.1), *paro* offers a wider range of occurrences. By exploiting semantic role labeling, we can know what Catiline prepares in BC. The most frequent Patients of the occurrences of *paro* in BC with Catiline as Actor are the following: *arma* (“implements of war”, “weapons”), *incendium* (“burning”), *insidiae* (“trap”) and *interficio* (“to destroy”). Indeed, Catiline is a bad guy in BC.

Given that Catiline plays the role of Actor in BC more than three times more than Cicero, one can expect that most of the Actions performed by Cicero are common with Catiline and that these Actions are more frequently performed by Catiline than Cicero. Actually, there are some deviations from such trend. The most clear example is the verb *refero* (“to bear back”, “to report”), whose Actor is Cicero in two occurrences while Catiline does never perform it. Moreover, there are three verbs that feature Cicero as Actor more than once and more than Catiline. These are *cognosco* (“to know”) and *praecipio* (“to take in advance”, “to warn”). Both these verbs have Cicero as Actor twice and Catiline once. Finally, the Action most frequently performed by Cicero (3) is represented by the verb *iubeo* (“to give an order”, “to command”). Also Catiline is Actor of *iubeo*, but only in two occurrences.

In order to understand if the Actors reported in Table 2 can be properly organized into homogeneous groups defined by the Actions of them, we performed a clustering analysis of the results.

3.3 Clustering the Actors

Clustering is the process of organizing objects (“observations”) into groups (“clusters”) whose members are similar in some way. One of the trickiest issues in clustering is to define what ‘similarity’ means and to find a clustering algorithm that computes efficiently the degree of similarity between two objects that are being compared.

Hierarchical clustering is a specific method of cluster analysis that seeks to build a hierarchy of clusters. Hierarchical clustering can be performed by following two main strategies: (a) agglomerative (bottom-up): each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy; (b) divisive (top-down): all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In this work, we apply hierarchical agglomerative clustering to compute the degree of similarity/dissimilarity between the Actors reported in Table 2. Such degree is obtained by comparing Actors by the Actions they perform. First, we compute the amount of shared and non-shared Actions between the members of all the possible couples of Actors. Then, we compare the distribution of shared and not shared Actions by their relative frequency.¹⁰ As for the distance measure, the analysis is run on document-term matrices by using the cosine distance¹¹

$$d(i; i') = 1 - \cos\{(x_{i1}, x_{i2}, \dots, x_{ik}), (x_{i'1}, x_{i'2}, \dots, x_{i'k})\}.$$

The arguments of the *cosine* function in the preceding relationship are two rows, i and i' , in a document-term matrix; x_{ij} and $x_{i'j}$ provide the number of occurrences of verb j ($j=1, \dots, k$) in the two sets of Actions corresponding to rows i and i' (“profiles”). Zero distance between two sets (cosine = 1) holds when two sets with the same profile are concerned (i.e. they have the same relative conditional distributions of terms). In the opposite case, if two sets do not share any word, the corresponding profiles have maximum distance (cosine = 0).

As for clustering, we run a “complete” linkage agglomeration method. While building clusters by agglomeration, at each stage the distance (similarity) between clusters is determined by the distance (similarity) between the two elements, one from each cluster, that are most distant. Thus, complete linkage ensures that all items in a cluster are within some maximum distance (or minimum similarity) to each other.

Roughly speaking, according to our clustering method, Actors that share a high number of Actions with similar distribution are considered to have a high degree of similarity and, thus, fall into the same or related clusters. Figure 4 plots the results and includes three main clusters.

Moving from top to bottom, the first cluster includes the two most similar Actors according to the Actions they perform. These are *cicero* and *consul* (“consul”). This happens although BC includes several occurrences of *consul* that are not referred to Cicero. Actually, Marcus Tullius Cicero is the consul par excellence in Roman political history and he was the only consul among the Actors considered here, as Caesar would become consul for the first time in 59 BCE, four years after the facts told in BC. The second most similar couple of Actors is the one including *catilina* and *lentulus* (similar at height 0.76). Catiline was the one who devised the conspiracy narrated in BC. Publius Cornelius Lentulus was one of the main conspirators. In particular, he took the place of Catiline as chief of the conspirators in Rome, when Catiline had to leave the city after the famous second speech of Cicero *In Catilinam*. The two characters are, thus, strictly related. In the same larger cluster are *curius* and *populus* (“people”). Quintus Curius was another conspirator, although his role was actually ambivalent. Being a friend of Catiline, he took part in the conspiracy, but at the same time it was because of him that it was foiled. According to Sallust, Curius, to boast with his mistress Fulvia, told her the details of the conspiracy, which she informed Cicero about. Moreover, Curius accused Caesar of being a conspirator. Such an undefined role is played also by “the people”. In those passages where Sallust talks about “the Roman people” (“*populus romanus*”), these are mostly positively depicted. Conversely, there are also places in

¹⁰All the experiments were performed with the R statistical software (R Development Core Team, 2012). More details on the clustering method used here are in Passarotti and Cantaluppi (2016).

¹¹A document-term matrix is a mathematical matrix that holds frequencies of distinct terms for each document. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

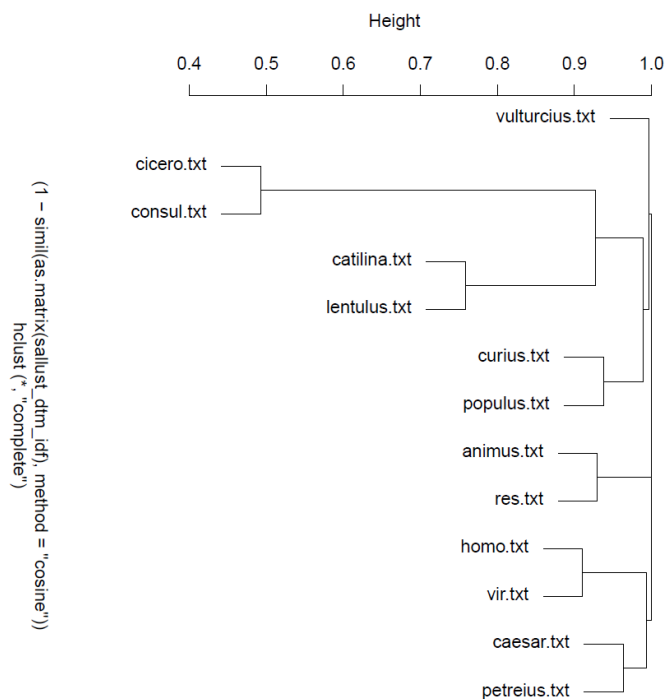


Figure 4: Clustering the Actors.

BC where the people act badly. Finally, Titus Vulturcius, a conspirator playing a subordinate role in the plot, falls into the same cluster, standing quite apart from the others.

The second cluster includes just two lemmas: *animus* (“soul”) and *res* (“thing”). These are the only not human Actors, among the ones considered here.

The third cluster features two couples of Actors. The first includes lemmas *homo* (“human being”, “man”) and *vir* (“adult male”, “man”), which are semantically strictly related, standing in hypernym/hyponym relation. The second couple is formed by *petreius* and *caesar*. Marcus Petreius plays a positive role in BC, having led the senatorial forces in the victory over Catiline in Pistoia. It is worth noting that such a positive character in the plot gets clustered together with Caesar. The future dictator Gaius Iulius Caesar hoped for the success of the second conspiracy of Catiline, just like he did for the first. However, Sallust’s intent is to lift Caesar of any suspicion of a possible link with Catiline. He emphasizes the Caesar’s concern for legality, depicting him (together with Cato) as the faithful guardian of “*mos maiorum*”, the core, unwritten code of Roman traditionalism. Putting Caesar under such a positive light is strictly connected to the fact that, while BC was being written, Caesar was deified by decree of the Roman Senate (on 1st January 42 BCE), after his assassin on the Ides of March 44 BCE.

4 Conclusion

The work described in this paper represents a case study showing how much useful a treebank enhanced with semantic annotation can be for literary studies. In this respect, there is still much to do. On one side, still too few literary texts provided with such annotation layer are currently available. On the other, the use of linguistic resources like treebanks remains dramatically confined in the area of computational and theoretical linguistics, not impacting other communities which might largely benefit from such resources.

To overcome the former, one desideratum is building NLP tools able to provide good accuracy rates of semantic annotation across different domains. As for the latter, developers of treebanks based on literary data and/or texts written in ancient languages must more and more get in touch with different kinds of domain experts from the Humanities, like philologists, historical linguists, philosophers, historians and scholars in literature. Indeed, across the last few years, this looks like a growing trend, with several events and special issues of scientific journals dedicated to different topics in computational linguistics and the Humanities. We hope that this is just the beginning of a fruitful joint work.

References

- Axel W. Ahlberg. 1919. *C. Sallusti Crispi. Catiline, Iugurtha, Orationes Et Epistulae Excerptae De Historiis*. Teubner, Leipzig.
- David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*. Springer, pages 79–98.
- David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. *Guidelines for the Syntactic Annotation of Latin Treebanks*. Tufts University Digital Library, Boston, MA.
- Berta González Saavedra and Marco Passarotti. 2014. Challenges in enhancing the Index Thomisticus treebank with semantic and pragmatic annotation. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT-13)*. Department of Linguistics, University of Tübingen, pages 265–270.
- Jan Hajič, Alena Böhmová, Eva Hajičová, and Barbora Vidová Hladká. 2000. The Prague dependency treebank: A three-level annotation scenario. In *Treebanks: Building and Using Parsed Corpora*. Kluwer, pages 103–127.
- Dag Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008)*. ELRA, pages 27–34.
- Marie Mikulová et al. 2006. *Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank*. Institute of Formal and Applied Linguistics, Prague, Czech Republic.
- Marco Passarotti. 2011. Language resources. The state of the art of Latin and the Index Thomisticus treebank project. In *Corpus ancients et Bases de données*. Presses universitaires de Nancy, pages 301–320.
- Marco Passarotti and Gabriele Cantaluppi. 2016. A statistical investigation into the corpus of Seneca. In *Latinitatis Rationes. Descriptive and Historical Accounts for the Latin Language*. De Gruyter, pages 684–706.
- R Development Core Team. 2012. *R: A language and environment for statistical computing*. Foundation for Statistical Computing, Vienna, Austria.
- John T. Ramsey. 2014. *Sallust. The war with Catiline. The war with Jugurtha*. Harvard University Press, The Loeb Classical Library 116, Cambridge, MA.
- Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, NL.
- Jan Štěpánek and Petr Pajas. 2010. Querying diverse treebanks in a uniform way. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*. ELRA, pages 1828–1835.
- Zdeněk Žabokrtský. 2011. Treex – an open-source framework for natural language processing. In *Information Technologies Applications and Theory*. Univerzita Pavla Jozefa Šafárika v Košiciach, pages 7–14.