PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 111 OCTOBER 2018 5–28

Using Tectogrammatical Annotation for Studying Actors and Actions in Sallust's Bellum Catilinae

Berta González Saavedra,[a] Marco Passarotti[b]

[a] Dep. de Filología Clásica, Universidad Autónoma de Madrid, Spain
[b] CIRCSE Research Centre. Università Cattolica del Sacro Cuore, Milan, Italy

Abstract

In the context of the *Index Thomisticus* Treebank project, we have enhanced the full text of *Bellum Catilinae* by Sallust with semantic annotation. The annotation style resembles the one used for the so called "tectogrammatical" layer of the Prague Dependency Treebank. By exploiting the results of semantic role labeling, ellipsis resolution and coreference analysis, this paper presents a network-based study of the main Actors and Actions (and their relations) in *Bellum Catilinae*.

1. Introduction

Since the second half of the nineties, the research area dealing with enhancing linguistic data with syntactic annotation ("treebanking") has faced a turn from constituency-based to dependency-based annotation schemata. The result is the current availability of several dependency treebanks for quite a number of languages. Most of these are now part of Universal Dependencies (`http://universaldependencies.org/`), an ever growing collection of dependency treebanks for several different languages following a cross-linguistically consistent annotation schema, which is in the process of becoming the standard de facto in the field.

The large majority of the currently available treebanks includes data taken from contemporary books, magazines, journals and, mostly, newspapers. Such data are

This paper is an extended version of the work presented by Passarotti and González Saavedra (2017) at the Sixteenth Edition of the International Workshop on Treebanks and Linguistic Theories (TLT-16), 23-24 January 2017, Prague, Czech Republic.

 Corresponding author: marco.passarotti@unicatt.it
Cite as: Berta González Saavedra, Marco Passarotti. Using Tectogrammatical Annotation for Studying Actors and Actions in Sallust's Bellum Catilinae. The Prague Bulletin of Mathematical Linguistics No. 111, 2018, pp. 5–28.
doi: 10.2478/pralin-2018-0005.

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 Natural Language Processing (NLP), like inducing grammars and training/testing tools.

Throughout the last decade, a small but constantly 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). Moreover, dependency treebanks are available also for other ancient languages, like Gothic and Old Church Slavonic (part of PROIEL), and Hittite (Inglese, 2015). Such linguistic resources for ancient languages support studies in historical linguistics together with a number of treebanks that include texts representing different diachronic phases of modern languages. Examples are the York-Toronto-Helsinki corpus (Taylor, 2007) and the Penn Corpora of Historical English (Taylor and Kroch, 1994; Kroch et al., 2004), the MCVF corpus for French (Martineau, 2008), the Tromsø Old Russian and OCS Treebank (Eckhoff and Berdicevskis, 2015) and RRuDi for Russian (Meyer, 2011), and the Mercurius Treebank for Early High New German (Demske, 2007).

Unlike those for modern languages, treebanks for ancient languages tend to include literary, historical, philosophical and/or documentary texts. This makes the very use of such resources peculiar. Indeed, instead of exploiting data to draw (cross-)linguistic generalizations, the 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, by using the results of semantic role labeling, coreference analysis and ellipsis resolution applied on the source data, the analysis focuses on the main Actors and Actions in Sallust's *Bellum Catilinae*.

Written 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.[1] One of the masterpieces of the Latin literature, *Bellum Catilinae* has been object of several and exhaustive studies, especially by historians of the Roman republican period and scholars in Latin Literature and Linguistics. From a historical perspective, particular attention has been paid to the intention of Sallust when writing his book (Conley, 1981) as well as to the amount of historical incongruences in his text (Syme, 1964). In addition, multiple contributions focus on various aspects of the language of Sallust (Batstone, 2010; Schröder, 2015; Tannenbaum, 2005).

The figure of Catiline has always fascinated the general public, particularly because of the complexities of his character as organizer of the conspiracy. Thus, a significant number of works deal with Sallust's depiction of Catiline, most of the times comparing it to the one provided by Cicero in his *In Catilinam*, which was no doubt the most important source for Sallust while writing *Bellum Catilinae*.[2]

Precisely because of such kind of portrayals, throughout the centuries the image of Catiline was deformed to the point that modern scholars often considered him an evil character (Earl, 1958) as well as the personification of ambition and greed (McConaghy, 1974). Contrary to these approaches, Ann Thomas Wilkins (1994) proposed in the nineties to read Sallust's treatment of the character of Catiline in a more complex way, according to which the author would be using the conspiracy of Catiline with the clear intention to show the decadence of Rome. In the view of Wilkins, the purposes of Sallust lay on the structure of the work, as she creates an antithesis between the first part of the account -where Catiline's conspiracy is presented from the perspective of Roman oligarchy– and the second part –where the distinction between the revolutionaries and the members of the establishment is already blurred. To this aim, Wilkins focuses on the distribution of the book, considering narrative periods, discourses, moral digressions and, most of all, how the descriptive words used for and by the main characters become common for both sides.

Moving from such a linguistic-based approach, we believe that analyzing the main Actors and Actions in *Bellum Catilinae* through the (deep) semantic annotation of the entire text of Sallust can provide a strong empirical support helping historians and Literature scholars to shed some further light on Sallust's portrayal of Catiline.

---

[1] 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] On this question, see Broughton (1936) and Waters (1970). An interesting perspective is provided by Syme (1964; page 73), who defends that Cicero is not the only author Sallust used for the compilation of *Bellum Catilinae*.

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 (ATSs) and tectogrammatical tree structures (TGTSs).

In ATSs 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.).

TGTSs describe the underlying structure of the sentence, conceived as the semantically relevant counterpart of the grammatical means of expression (described by ATSs). The nodes of TGTSs 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,[3] Patient, Ad-

[3]The 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).

dressee, Effect and Origin; (b) adjuncts, called "free modifications": different kinds of adverbials, like Place, Time, Manner etc.[4]

Also coreference analysis and ellipsis resolution are performed at the tectogrammatical layer and are represented in TGTSs 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 Analytical to Tectogrammatical Layer

2.2.1. Converting from ATSs to TGTSs in the *Index Thomisticus* Treebank Project

The workflow for tectogrammatical annotation in the IT-TB project is based on TGTSs automatically converted from ATSs.[5] The TGTSs 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).[6]

Relying on ATSs, the basic functions of these modules are the following:

a. to collapse ATSs nodes of function words and punctuation marks, as they no longer receive a node for themselves in TGTSs, but collapse into the nodes for autosemantic words;
b. to assign "grammatemes", i.e. semantic counterparts of morphological categories (for instance, pluralia tantum are tagged with the number grammateme "singular");
c. to resolve grammatical coreferences;
d. to assign semantic roles.

Tasks (a) and (b) are quite simple and the application of the modules that are responsible for them results in good accuracy on average. Collapsing nodes for function

[4]The organization of functors into inner participants and free modifications is further exploited by linking textual tectogrammatical annotation with fundamental lexical information provided by a valency lexicon that features the valency frame(s) for all those verbs, nouns, adjectives and adverbs capable of valency that occur in the treebank. The valency lexicon of Latin, called *Latin Vallex* (Passarotti et al., 2016), was built in corpus-driven fashion, by adding to the lexicon all the valency-capable words that annotators progressively got through. A similar approach to build a valency lexicon based on treebank annotation is that of *PDT-Vallex* for Czech (Urešová, 2009).

[5]The 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, 2006), with a few modifications for representing Latin-specific constructions (`http://itreebank.marginalia.it/doc/Guidelines_tectogrammatical_Latin.pdf`).

[6]See 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.

words and punctuations relies on the structure of the ATSs given in input. In this respect, Latin does not feature any specific property requiring for modifications of the ATS-to-TGTS conversion procedure available in *Treex* and already applied to other languages. Assigning grammates is a task strictly related with the lexical properties of the nodes in TGTSs. Thus, we are in the process of populating the modules that assign grammatemes with lists of words (lemmas) that are regularly assigned the same grammatemes.

The automatic processing of task (c) results from the application of a number of modules aimed to resolve only the grammatical coreference that shows the simplest possible construction occurring in ATSs, i.e. the one featuring an occurrence of a relative pronoun directly depending on the main predicate of the relative clause. However, this construction is highly frequent for relative clauses. For instance, among the 326 occurrences of the relative pronoun *qui* in the portion of the IT-TB featuring tectogrammatical annotation, 176 present this construction and are correctly assigned their grammatical coreference by the conversion modules. The remaining 150 occurrences either lack grammatical coreference or do occur in more complex constructions.

In order to assign semantic roles automatically (task (d)), we rely both on analytical functions and on lexical properties of the ATSs nodes. For instance, all the nodes with analytical function Sb (Subject) that depend on an active verb are assigned functor ACT (Actor), and all the main predicates of subordinate clauses introduced by the conjunction *si* 'if' are assigned functor COND (Condition).

2.2.2. Examples of ATSs and TGTSs from *Bellum Catilinae*

In this section we report a number of examples of ATSs and TGTSs from BC.

Figure 1 shows the ATS for the sentence "Sed [but] maxume [most of all] adulescentium [of the young] familiaritatem [intimacy] adpetebat [sought]" (BC 14.5) ("But most of all [Catiline] sought the intimacy of young men").

The ATS in Figure 1 features as many nodes as the words of the sentence (5) plus the root node, which reports the ID of the sentence in the Latin portion of the AGLDT ("a-" here means "analytical") and it is assigned by default the analytical function AuxS (Sen-

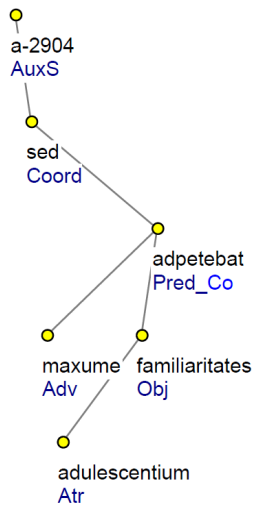

Figure 1. ATS of the sentence "Sed maxume adulescentium familiaritatem adpetebat" (BC 14.5).

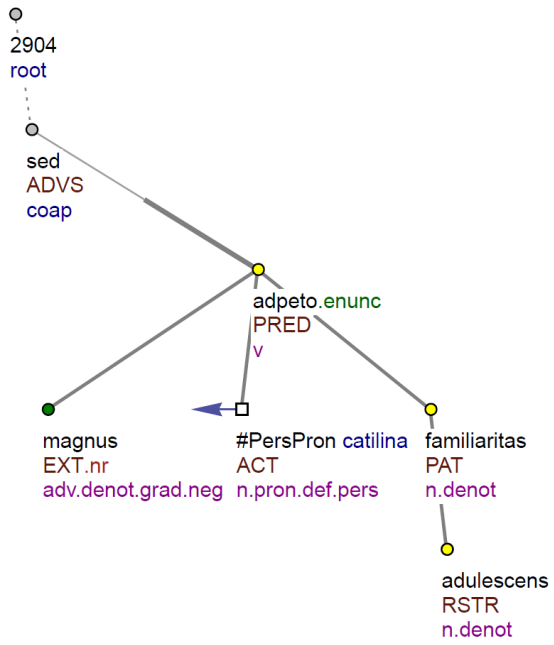

Figure 2. TGTS of the sentence "Sed maxume adulescentium familiaritatem adpetebat" (BC 14.5).

tence). Nodes are arranged from left to right according to the order of the words in the sentence. Each node is assigned an analytical function.[7]

The TGTS shown in Figure 2 features all the nodes of the corresponding ATS plus one. This newly added square node results from both ellipsis resolution and coreference analysis.

As for the former, the square node fills the position for a missing argument of the verb *adpeto*. Here "missing" means that the argument is not explicitly represented by either a lexical item or a phrase in the text. In this sentence, the verb *adpeto* is considered a word with two arguments, which are represented respectively by an Actor (ACT: missing) and a Patient (PAT: *familiaritas*).

As for the latter, the newly added node for the missing Actor is assigned t-lemma #PersPron,[8] which means that the node represents the missing occurrence of a per-

[7]Coord: coordination. Pred_Co: coordinated main predicate. Adv: adverbial modifier (adjunct). Obj: direct or indirect object (argument). Atr: attributive.

[8]#PersPron is the t-lemma assigned to nodes representing possessive and personal pronouns (including reflexives). Sets of different morphological lemmas can be grouped under the same t-lemma in TGTSs. This

sonal pronoun (like *is* 'he'), which is permitted by the pro-drop nature of Latin. The node is linked via a textual coreference to the last previous occurrence of the lemma *catilina*, which represents its denotation.

In Figure 2, nodes are arranged from left to right reflecting information structure according to Topic-Focus Articulation, moving from Topic (left) to Focus (right).[9] Each node is assigned a functor and a so called "semantic part of speech". The occurrence of the lemma *magnus* (form *maxume*) represents an EXT (Extent), i.e. an adjunct that expresses manner by specifying extent or intensity of the event or a circumstance. The semantic part of speech for this occurrence is that for gradable adverbs that can be negated. *Familiaritas* is a denominating semantic noun (n.denot) further specified by another noun acting as a restrictor of its head in the TGTS (functor: RSTR). Finally, *sed* is a an adversative (ADVS) coordinating connective (coap).

The main predicate of the sentence is assigned the so called "sentential modality", which consists in speech act annotation. In the TGTS shown in Figure 2, the sentence is an "enunciation" (enunc).

Figure 3 shows the ATS for the sentence "Sed [but] iuventutem [the young], quam [whom], ut [as] supra [above] diximus [we said], illexerat [he had ensnared], multis [many] modis [ways] mala [bad] facinora [crimes] edocebat [he taught]" (BC 16.1) ("The young men whom he had ensnared, as I have mentioned above, were instructed by him in wicked deeds of many forms").

The only analytical functions in Figure 3 that do not occur also in Figure 1 are AuxX (assigned to punctuation marks) and AuxC (for subordinating conjunctions). Figure 4 shows the corresponding TGTS.

In this sentence, *catilina* is Actor of two verbs: *illicio* and *edoceo*. In both cases, pronoun dropping and ellipsis resolution is performed. The Actor of

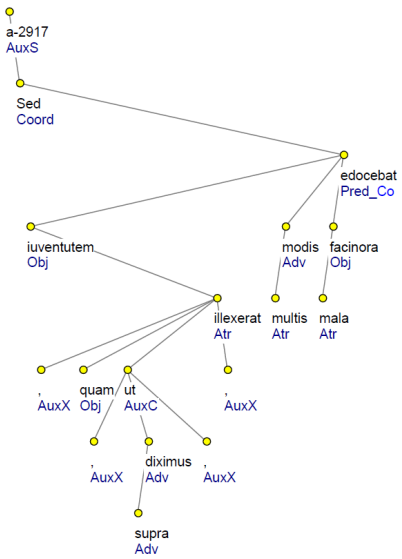

Figure 3. ATS of the sentence "Sed iuventutem, quam, ut supra diximus, illexerat, multis modis mala facinora edocebat" (BC 16.1).

is the case, for instance, of morphological lemmas *aliquis* 'someone', *quis* 'who?' 'which?', *quisquis* 'whoever' and *unusquisque* 'each', which are all assigned t-lemma *quis*.

[9]For details about Topic-Focus Articulation, see Mikulová et al. (2006; pages 1118-1188)

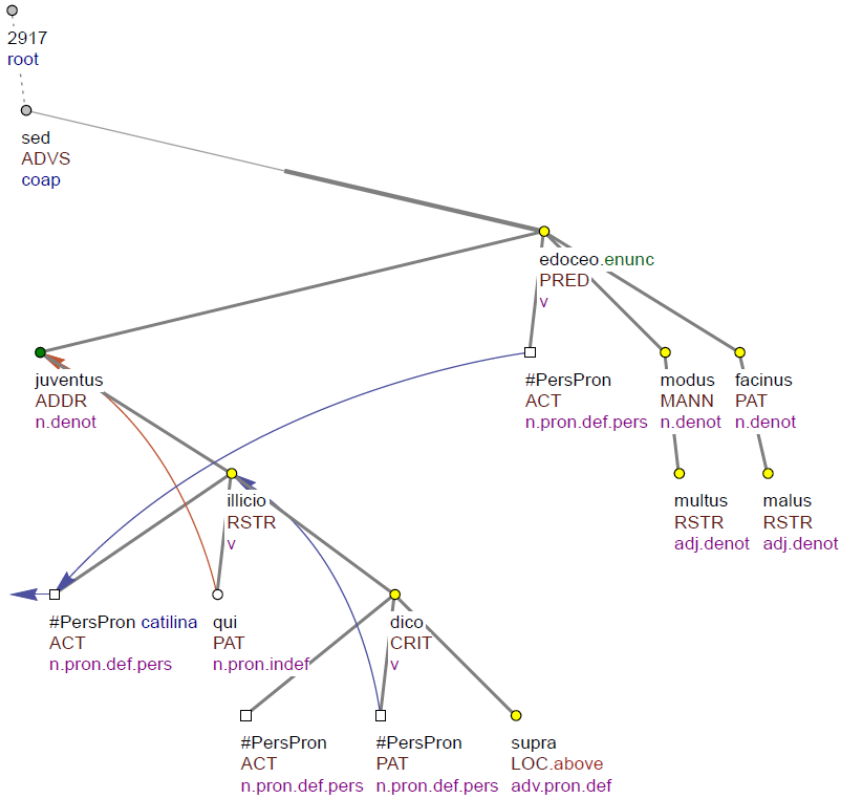

Figure 4. TGTS of the sentence "Sed iuventutem, quam, ut supra diximus, illexerat, multis modis mala facinora edocebat" (BC 16.1).

edoceo is linked via a textual coreference to that of *illicio*, which is in turn textually coreferred to the previous occurrence of *catilina*. Equally, the newly added node standing for the Patient of the verb *dico* is linked to *illicio*, because what "we have said above" is that "he had ensnared the young men". Figure 4 shows also a grammatical coreference holding between the relative pronoun *qui* (Patient of *illicio*) and the noun *juventus* (Addresse of *edoceo*). As for the functors, LOC is assigned to Locatives answering the question "where?" and it is here further specified by the subfunctor "above" (*supra*). MANN is a functor for such an adjunct that expresses manner (*modis*). Finally, it is worth noting that the TGTS of Figure 4 does not include the node for the function word *ut*, which collapses into that for the autosemantic word *dico*.

Figure 5 shows the ATS for the sentence "cum [with] eo [him] se [himself] consulem [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 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 5, 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.[10]

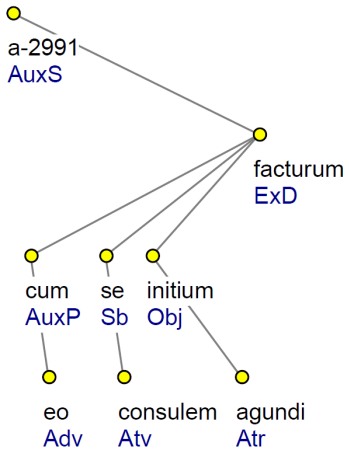

Figure 5. ATS of the sentence "cum eo se consulem initium agundi facturum" (BC 21.4).

Figure 6 shows the TGTS for this sentence. The TGTS 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 Addresse result from ellipsis resolution: Catiline is the Actor and the men (*homo*) are the Addresse. 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: 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 General 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 (literaly: "the beginning of acting").

The prepositional phrase "cum eo" ("with him") is represented in the TGTS of Figure 6 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,

[10]The analytical function Atv is assigned to verbal attributes, i.e. (predicative) complements not participating in government (*consulem*). AuxP is used for prepositions (*cum*).

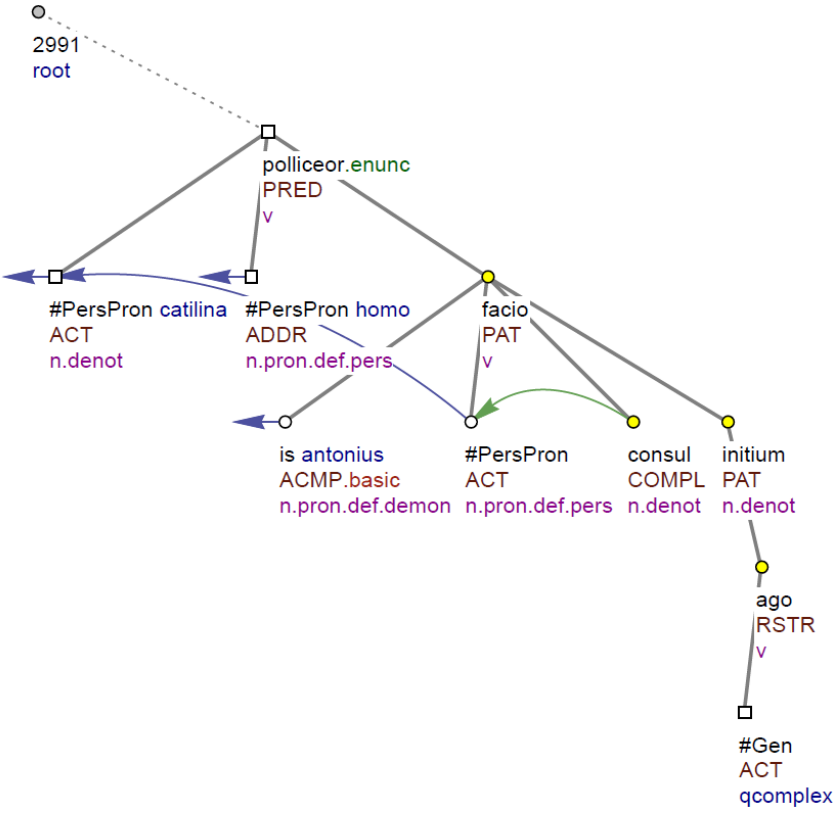

Figure 6. TGTS of the sentence "cum eo se consulem initium agundi facturum" (BC 21.4).

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 6, 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 an arrow (going from *consul* to #PersPron in Figure 6).

3. Methodology

One of the added values of tectogrammatical annotation is that it provides information that, although it is accessible to readers, is explicitly missing in the text. For instance, looking at the example sentence whose ATS and TGTS are shown in Figures 5 and 6 respectively, 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 carries (implicit) information about the fact that Catiline performs two different Actions (namely, *polliceor* and *facio*).

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.[11]

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 `$n0`) that are assigned PoS verb (`gram/sempos = v'`) and govern either directly or indirectly (`echild`) a t-node (`$n1`) with functor ACT (`functor = 'ACT'`).[12] 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 6: #PersPron-*polliceor*, #PersPron-*facio* and #Gen-*ago*. While #Gen is a General 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. In particular, the newly added

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

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

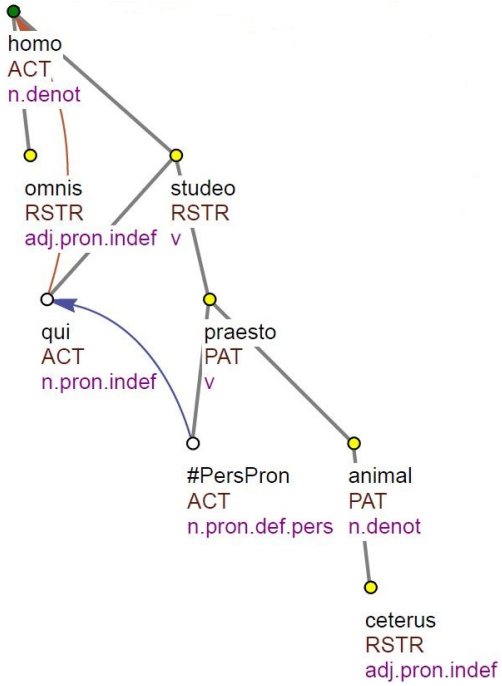

Figure 7. TGTS of the phrase "Omnis homines, qui sese student praestare ceteris animalibus [...]" (BC 1.1).

#PersPron node depending on *polliceor* is directly linked via textual coreference with its antecedent (*catilina*), while the #PersPron node depending on *facio* shows an indirect linking with its antecedent by passing through the other #PersPron node.

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 with 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 7 shows the portion of the TGTS for the first sentence of BC concerning this phrase.

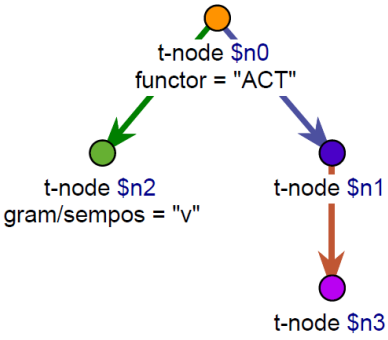

Figure 8. A graphical query in PML-TQ.

From Figure 7, 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 (graphically represented in Figure 8):

```
                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 `$n3`. In the TGTS of Figure 7, `$n2` is the node for *praesto*, `$n0` is the #PersPron node depending on *praesto*, `$n1` is *qui* and `$n3` is *homo*. By just printing in the output of the query the t-lemma for node `$n3`, it is possible to replace #PersPron with *homo* in the list of the Actor-Action couples.[13]

3.2. Networking the Data

Once built the list of all the Actor-Action couples and having enhanced each couple with its frequency of occurrence in the TGTSs of BC, we induced automatically a network from the list.

In order to build the network out of the tectogrammatical annotation of BC, we applied the method developed by Ferrer i Cancho et al. (2004). According to this

[13]The longest coreference chain we found in BC includes 5 textual coreferences.

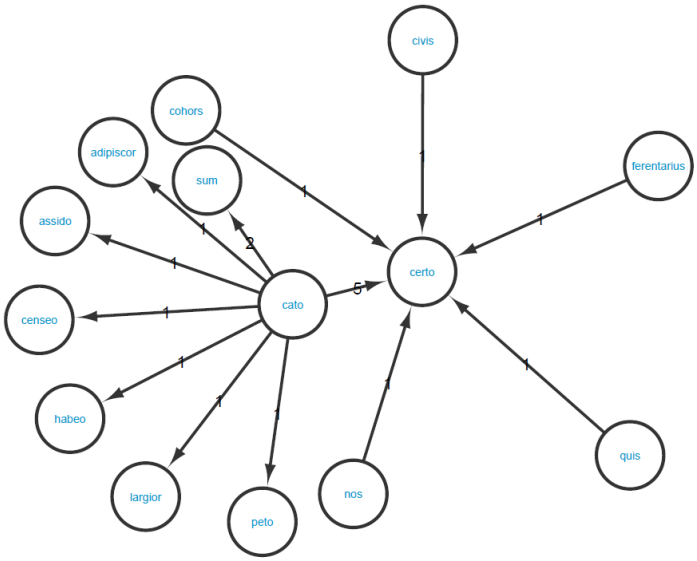

Figure 9. The tecto-based subnetwork for cato *and* certo.

method, a dependency relation appearing in the source treebank is converted into an edge in the network and two vertices are linked in the network if they appear at least once in a dependency relation in the treebank. The edges are directed according to the direction of the dependency relation in the treebank. In the case of our network, the vertices are Actors and Actions, and the edges are the dependency relations holding between Actors and Actions. The edges are directed from the Actors to the Actions.

The network is built by accumulating sentence structures from the treebank. The treebank is parsed sentence by sentence and new vertices are added to the network. When a vertex is already present in the network, more links are added to it.

The result is a "tecto-based" network containing all the dependencies between Actors and Actions in the input treebank. Edges are weighted by frequency, i.e. each connection between two vertices is enhanced with the number of its occurrences in the source TGTSs.

Figure 9 shows the portion of the tecto-based network concerning the Actor *cato* and the Action *certo* 'to fight'. The node for *cato* is connected to all the Actions that *cato* performs in BC via outgoing edges enhanced with the frequency of the connection they represent; for instance, from Figure 9 one can understand that *cato* performs the Action represented by *certo* five times. Conversely, the node for *certo* is connected to all the Actors that perform such Action via ingoing edges.

The full tecto-based network of BC is shown in Figure 10. The nodes of this network represent all the Actors and the Actions of BC, while its edges are all the dependency relations holding between them in the source TGTSs.

In the following, we first use some topological properties of the tecto-based network of BC to study Actors and Actions in BC. Then, we run a clustering analysis of its vertices with the highest out-degree (i.e. the Actors reported in Table 1) to understand if they can be properly organized into homogeneous groups defined by the set of the vertices they are connected to via outgoing edges (i.e. the Actions they perform).

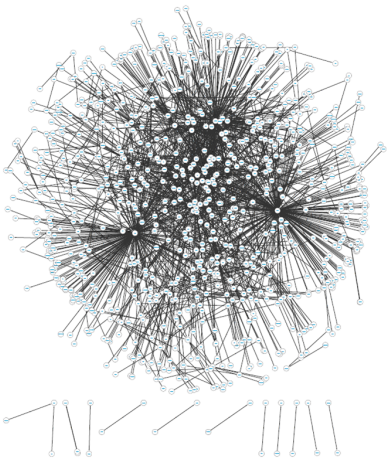

Figure 10. The tecto-based network of Bellum Catilinae.

4. Results and Discussion

4.1. Actors and Actions

Table 1 reports the main Actions and the main Actors in BC. These are defined as the vertices in the tecto-based network with the highest out-degree (Actors) and in-degree (Actions) respectively.[14] In other words, this means that the main Actors are those that perform the highest number of different Actions and, conversely, the main Actions are those performed by the highest number of different Actors.[15]

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.

[14]In a network, the degree of a vertex *s* is the number of its edges, i.e. different relations holding between *s* and other vertices in the network. In a directed network (like the tecto-based network here concerned), the degree results from the sum of the out-degree, which labels the number of edges that are directed from the vertex, and of the in-degree, which labels the number of edges that are directed to the vertex.

[15]The 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, 2006, pp. 318–320). Not coreferred Actors are excluded from Table 1. These are the General 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

Actor	Actions	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 1. Main Actions (left) and Actors (right).

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,[16] *paro* offers a wider range of occurrences. By exploiting

[16] "Igitur [Catilina] C. Manlium Faesulas atque in eam partem Etruriae [dimisit], Septimium quendam Camertem in agrum Picenum [dimisit], C. Iulium in Apuliam dimisit, praeterea alium alio [dimisit], quem ubique opportunum sibi fore credebat" (BC 27.1) ("He, therefore, dispatched Gaius Manlius to Faesulae and that region of Etruria, a certain Septimius of Camerinum to the Picene district, and Gaius Julius to Apulia; others too to other places, wherever he believed that each would be serviceable to him"). The three occurrences of *dimisit* put in square brackets are generated in the TGTS of this sentence via ellipsis resolution. *Catilina* is generated as well, playing the role of Actor of all the generated occurrences of *dimisit*.

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'. Such Patients of *paro* show the complexity of the character of Catiline, who is depicted somewhere negatively (mostly in the first half of BC) and somewhere else positively. In fact, while looking at the Patients of *paro*, we see that Catiline is not only someone who prepares malitious acts (*insidias parare*), but he also encourages the revolutionaries to the arms (*arma parare*), which is presented by Sallust under a positive light, as Wilkins (1994) points out (page 51).

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 by 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.

4.2. Clustering the Actors

Clustering is a technique that deals with finding a structure in a collection of data. In particular, clustering is the process of organizing objects (called "observations") into groups ("clusters") whose members are similar in some way. One of 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 1. 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 Ac-

tions by their relative frequency.[17] As for the distance measure, the analysis is run on document-term matrices by using the cosine distance[18]

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

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, ..., 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 11 plots the results.

Looking at Figure 11, it turns out that there are 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*. 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. This is further confirmed by the following words of Cato's speech, which closely connect the decision to be taken by the Senate about Lentulus with that about the army of Catiline: "Qua re quom de P. Lentulo ceterisque statuetis, pro certo habetote vos simul de exercitu Catilinae et de omnibus coniuratis decernere" (BC 52.17)

[17] All the experiments were performed with the R statistical software (R Development Core Team, 2012). More details about the clustering method used here can be found in Passarotti and Cantaluppi (2016).

[18] A 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.

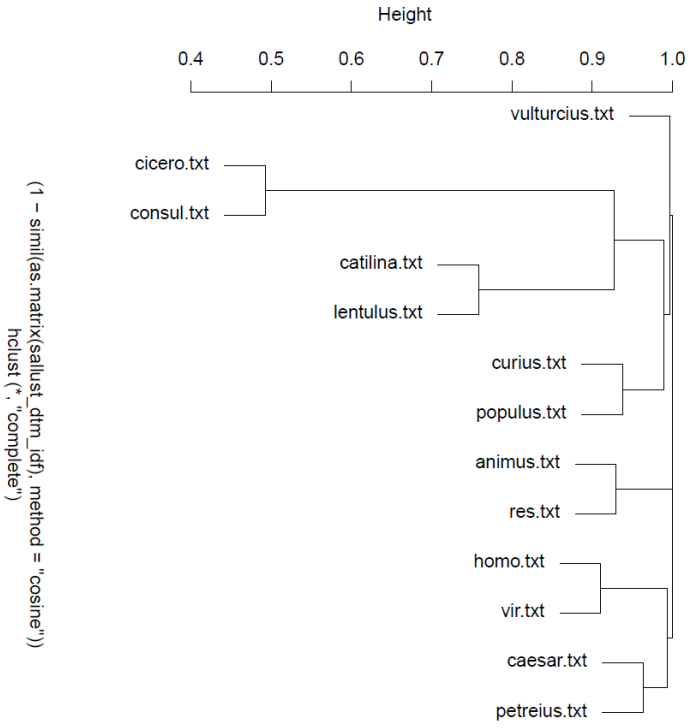

Figure 11. Clustering the Actors.

("Be assured, then, that when you decide the fate of Publius Lentulus and the rest, you will at the same time be passing judgment on Catiline's army and all the conspirators"). In this respect, the destinies of Lentulus and Catiline are not only linked to each other, but they are also strictly bound to the outcome of the conspiracy. Indeed, as Wilkins (1994; page 95) points out, Lentulus's execution on one side and the death of Catiline on the other represent respectively the first and the last step in the failure of the conspiracy.

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

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.

5. Conclusion and Future Work

In the context of the work presented in this paper, there are two main open issues to address in the near future. First, we must enhance data with coreference analysis of either explicit or generated first and second personal pronouns. This is needed because BC features speeches given by different characters, which makes the reference of such pronouns change across the text. Second, our work must consider also Actions represented by (deverbal) nouns together with their either explicit or generated Actors, which tend to occur as subjective genitives.

As far as the interpretation of some specific aspects of BC is concerned, we have shown that using tectogrammatical annotation for studying the Actors and the Actions of the plot can clarify to some extent how Sallust conceived Catiline's portrayal. In this respect, future work must focus specifically on the Actions performed by Catiline, by taking into consideration the structure of the book in order to determine the evolution of the character throughout the chapters. Also, performing the tectogrammatical annotation of Cicero's *In Catilinam* would help to compare the two portrayals of Catiline.

More generally speaking, the work described here 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.

Acknowledgements

This research is supported by the Italian Ministry of Education, University and Research (MIUR), FIR-2013 project "Developing and Integrating Advanced Language Resources for Latin" (ID: RBFR13EWQN).

Bibliography

Ahlberg, Axel W. *C. Sallusti Crispi. Catiline, Iugurtha, Orationes Et Epistulae Excerptae De Historiis*. Teubner, Leipzig, 1919.

Bamman, David and Gregory Crane. The Ancient Greek and Latin Dependency Treebanks. In *Language Technology for Cultural Heritage*, pages 79–98. Springer, 2011. URL https://doi.org/10.1007/978-3-642-20227-8_5.

Bamman, David, Marco Passarotti, Gregory Crane, and Savina Raynaud. *Guidelines for the Syntactic Annotation of Latin Treebanks*. Tufts University Digital Library, Boston, MA, 2007. URL https://itreebank.marginalia.it/doc/2007_Passa+Bamman+Crane+Raynaud_Guidelines%20Tb.pdf.

Batstone, William. Word at War: The Prequel. In *Citizens of Discord: Rome and Its Civil Wars*, pages 45–72. OUP USA, 2010.

Broughton, Thomas RS. Was Sallust Fair to Cicero? In *Transactions and Proceedings of the American Philological Association*, pages 34–46. JSTOR, 1936.

Conley, Duane F. The Interpretation of Sallust Catiline 10. 1-11. 3. *Classical Philology*, 76(2): 121–125, 1981.

Demske, Ulrike. Das MERCURIUS-Projekt: Eine Baumbank für das Frühneuhochdeutsche. *Sprachkorpora: Datenmengen und Erkenntnisfortschritt*, pages 91–104, 2007. URL https://doi.org/10.1515/9783110439083-007.

Earl, Donald C. *The political thought of Sallust*. PhD thesis, University of Cambridge, 1958.

Eckhoff, Hanne Martine and Aleksandrs Berdicevskis. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta*, 14:15, 2015.

Ferrer i Cancho, Ramon, Ricard V Solé, and Reinhard Köhler. Patterns in syntactic dependency networks. *Physical Review E*, 69(5):051915, 2004. URL https://doi.org/10.1103/physreve.69.051915.

González Saavedra, Berta and Marco Passarotti. 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, 2014. URL http://tlt13.sfs.uni-tuebingen.de/tlt13-proceedings.pdf#page=273.

Hajič, Jan, Alena Böhmová, Eva Hajičová, and Barbora Vidová Hladká. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Kluwer, 2000.

Haug, Dag and Marius Jøhndal. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008)*, pages 27–34. ELRA, 2008. URL http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf#page=31.

Inglese, Guglielmo. Towards a Hittite Treebank. Basic Challenges and Methodological Remarks. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)*, pages 59–68. Institute of Computer Science of the Polish Academy of Sciences, 2015.

Kroch, Anthony, Beatrice Santorini, and Lauren Delfs. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM. *Department of Linguistics, University of Pennsylvania, CD-ROM*, 2004.

Martineau, France. Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*, 7, 2008.

McConaghy, Mary Lee Sivess. *Sallust and the Literary Portrayal of Character*. UMI, Washington, 1974.

Meyer, Roland. New wine in old wineskins?—Tagging Old Russian via annotation projection from modern translations. *Russian linguistics*, 35(2):267–281, 2011. URL https://doi.org/10.1007/s11185-011-9075-x.

Mikulová, Marie et al. *Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank*. Institute of Formal and Applied Linguistics, Prague, Czech Republic, 2006. URL https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf.

Passarotti, Marco. Language Resources. The State of the Art of Latin and the Index Thomisticus Treebank Project. In *Corpus anciens et Bases de données*, pages 301–320. Presses universitaires de Nancy, 2011.

Passarotti, Marco and Gabriele Cantaluppi. A Statistical Investigation into the Corpus of Seneca. In *Latinitatis Rationes. Descriptive and Historical Accounts for the Latin Language*, pages 684–706. De Gruyter, 2016.

Passarotti, Marco and Berta González Saavedra. The Treebanked Conspiracy. Actors and Actions in Bellum Catilinae. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 18–26, 2017. URL http://www.aclweb.org/anthology/W17-7605.

Passarotti, Marco, Berta González Saavedra, and Christophe Onambele. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *LREC*, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/96_Paper.pdf.

R Development Core Team. *R: A language and environment for statistical computing*. Foundation for Statistical Computing, Vienna, Austria, 2012.

Ramsey, John T. *Sallust. The war with Catiline. The war with Jugurtha*. Harvard University Press, The Loeb Classical Library 116, Cambridge, MA, 2014.

Schröder, Wilt Aden. Zu Sallust, Catilina 3, 3 (und zum Gedankengang des Proömiums). In *Lemmata: Beiträge zum Gedenken an Christos Theodoridis*, pages 203–219. Walter de Gruyter GmbH & Co KG, 2015.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, NL, 1986.

Štěpánek, Jan and Petr Pajas. Querying Diverse Treebanks in a Uniform Way. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pages 1828–1835. ELRA, 2010. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/381_Paper.pdf.

Syme, Ronald. *Sallust*. University of California Press, Berkeley, 1964.

Tannenbaum, RF. What Caesar Said: Rhetoric and History in Sallust's Coniuratio Catilinae 51. In *Roman Crossings: Theory and Practice in the Roman Republic*, pages 209–223. Classical Press of Wales, 2005.

Taylor, Ann. The York—Toronto—Helsinki parsed corpus of old english prose. In *Creating and digitizing language corpora*, pages 196–227. Springer, 2007. URL https://doi.org/10.1057/9780230223202_9.

Taylor, Ann and Anthony S Kroch. The Penn-Helsinki Parsed Corpus of Middle English. *MS. University of Pennsylvania*, 1994. URL http://www.ling.upenn.edu/mideng/documentation/manual.ps.

Urešová, Zdeňka. Building the PDT-VALLEX valency lexicon. In *On-line proceedings of the fifth Corpus Linguistics Conference. University of Liverpool*, 2009. URL http://ufal.ms.mff.cuni.cz/pcedt2.0/publications/Uresova2011.pdf.

Waters, Kenneth H. Cicero, Sallust and Catiline. *Historia: Zeitschrift für Alte Geschichte*, H. 2: 195–215, 1970.

Wilkins, Ann Thomas. *Villain or hero: Sallust's portrayal of Catiline*, volume 15. Peter Lang Pub Inc, 1994.

Žabokrtský, Zdeněk. Treex – an open-source framework for natural language processing. In *Information Technologies – Applications and Theory*, pages 7–14. Univerzita Pavla Jozefa Šafárika v Košiciach, 2011. URL http://ceur-ws.org/Vol-788/paper2.pdf.

Address for correspondence:
Marco Passarotti
marco.passarotti@unicatt.it
Università Cattolica del Sacro Cuore. Largo Gemelli, 1 - 20123 Milan, Italy