Treebank for Dialogue: a case study from Roman Tragedy

Federica Iurescia and **Giovanni Moretti**
Università Cattolica del Sacro Cuore Milano
federica.iurescia@unicatt.it, giovanni.moretti@unicatt.it

Abstract

This paper presents a case study on dialogues in dramatic texts, leveraging a treebank enhanced with annotation of speakers. Information on characters speaking contributes to investigate dialogues from various perspectives, including the study of interaction and linguistic characterisation.

1 Introduction and related works

This paper aims at investigating dialogues in dramatic texts by leveraging information provided by treebanks. More specifically, it takes Roman Tragedy as a case study, and explores the language of the characters in Seneca's *Agamemnon*, a Latin dramatic text dating back to 2nd century CE.
As such, the present paper draws inspiration from various lines of research. On one side, from qualitative analysis of dialogues from the perspective of Conversation Analysis and Historical Pragmatics applied to ancient dramatic texts.[1] On another side, from works on language of dramatic characters from a quantitative perspective, particularly those works that exploit treebanks, that is, syntactically annotated texts.[2] The perspective of computational linguistic research has recently benefited from contributions in the field of Computational Drama Analysis.[3] The present paper relies on such studies in that it explores how we can advance our knowledge and comprehension of dialogues in drama with computational methods, focusing on language and interactions between characters.

2 Corpus

The corpus used for this case study is the text of the tragedy enhanced with syntactic annotation following the UD framework.[4] The text originates from the *Opera Latina* corpus built by the LASLA laboratoires in Liége,[5] and is provided with sentence-splitting, tokenization, lemmatization, PoS-tagging and the annotation of morphological features according to a format developed by the LASLA team. The texts of the *Opera Latina* corpus were converted from the LASLA into the CoNLL-U format, and into the UD formalism. The syntactic annotation was performed manually. *Agamemnon*'s text consists of 5580 tokens distributed across 409 sentences. It is one of the three texts currently present in the UD_Latin-CIRCSE Treebank, and it is enhanced with the annotation of the speakers to whom each sentence is attributed. This annotation, manually performed, is formatted as a comment in the CoNLL-U file and follows the comment line that reports the text of each sentence. In cases where the same sentence includes words uttered by more than one speaker, the indication of speakers details the distribution of tokens among them (see Figure 1).

In cases of reported speech, the character who utters the reported speech is listed as first; the character who reports the speech is enclosed in round brackets, as exemplified in Figure 2, where the character named Eurybates reports words uttered by the people of Danaans.
Based on this annotation, we developed a Python script to extract all tokens attributed to each speaker.[6]

[1] See, e. g., (Martin et al., 2020), and the forthcoming proceedings of the conference on Conversation Analysis and Classical Languages (https://caclassics.wordpress.com/conferences/).

[2] For Ancient Greek Tragedy, see, e. g., (Mambrini, 2005).

[3] See, most recently, (Andresen and Reiter, 2024). Many contributions in that volume rely on texts collected under the Drama Corpora Project available at https://dracor.org/ (See (Fischer et al., 2019)).

[4] See (de Marneffe et al., 2021) and https://universaldependencies.org/. The Latin treebank is available at https://github.com/UniversalDependencies/UD_Latin-CIRCSE.

[5] https://www.lasla.uliege.be/cms/c_8508894/fr/lasla.

[6] The script is available at https://github.com/CIRCSE/UD_Latin-CIRCSE in the "scripts" folder. It takes as input a CoNLL-U file enhanced with annotation of speakers as

```
# sent_id = Latin_SenecaYounger_Ag_poetry-1
# text = opaca linquens Ditis inferni loca adsum profundo Tartari emissus specu incertus utras oderim sedes magis fugio
Thyestes inferos superos fugo
# speaker = Thyestis umbra

# sent_id = Latin_SenecaYounger_Ag_poetry-199
# text = sistito infestum mare uehit ista Danaos classis et Troas uehit nec plura possunt occupat uocem mare
# speaker = Danai (token 1-10), Eurybates (token 11-16)
```

Figure 1: Annotation of speakers as comment in the CoNLL-U file

```
# sent_id = Latin_SenecaYounger_Ag_poetry-194
# text = nil nobile ausos pontus atque undae ferunt
# speaker = Danai (Eurybates)
```

Figure 2: Annotation of reported speech as a comment in the CoNLL-U file

3 Speakers in Dialogue: Agamemnon

For each speaker in the *Agamemnon*, we extract a number of properties, including the number of tokens and the number of speeches,[7] the type/token ratio and the number of sentences, the sentence depth, and a graph showing the tree related to each sentence. These properties enable to compare the distribution and variation of the language of the speakers on several levels. Among the possible levels of analysis, this paper focuses on the character who lends the tragedy its title, Agamemnon.

In spite of lending the tragedy its title, Agamemnon is one of the characters who speaks the least.[8] He enters the stage and expresses his relief for being back home after the Trojan war. He sees on stage the seer Cassandra who tries to warn him: she knows that he will soon be murdered, but Agamemnon does not really engage in conversation with her.[9] After a brief invocation to the gods, he leaves the stage. In terms of the types of interactions he is involved in, he appears isolated: he engages in conversation with only one character and does not even comprehend what the other character is attempting to convey to him. This state of affairs is expressed on the syntactic level by short sentences, with a maximum of sentence depth equal to 2.[10] In contrast, the average length and depth of his sentences increase when he addresses characters who do not reply to him, as it is the case with gods, or servants who merely execute his commands.[11] The distribution and complexity of dependency relations varies accordingly; for this case study, we take only sentence depth into account, as a proxy of complexity of syntactic trees.

4 Conclusions and future work

This paper offers a syntax-based study of the linguistic characterisations of dialogues in Seneca's *Agamemnon*. Far from being a systematic analysis, it intends to give an idea of the potential for enhancing the study of dialogues in dramatic texts. A possible expansion of the present research envisages the comparison of the language of the same character across different works, both within the same literary genre,[12] and across different genres.[13] Focusing on the study of communicative situations, mapping the variation of the syntactic tree according to the different type of interaction may lead to interesting observations, such as studying differences in the syntactic trees between, e. g., persuasion[14] and quarrel.[15] We plan to include the analysis of dependency relations in order to investigate syntactic patterns on a more fine-grained level, such as the structure of reported speeches.

described in Section 2. The results can be downloaded as a markdown file with the linguistic profiling of each speaker as described in Section 3.

[7] As speech it is to be intended a sequence of tokens uttered by the same speaker.

[8] In Seneca's *Agamemnon* there are twelve speakers. Two of them figure only in reported speeches (see Section 2 for the people of Danaans as an example) speaking approximately fifty tokens each. Agamemnon himself is attributed 135 tokens across 10 speeches. Only one character speaks less than him, uttering slightly more than a hundred tokens.

[9] Their dialogue is a clear example of failure in communication that may be explained with lack in Common Ground: see (Iurescia, 2021).

[10] His interlocutor Cassandra shows a similar linguistic behaviour when interacting with him: short sentences with a maximum of sentence depth equal to 2. When engaging in conversation with other characters, Cassandra utters longer sentences with a greater variation in sentence depth.

[11] Sentence depth for these two cases is 6 and 3 respectively.

[12] For instance, is there any difference in the linguistic characterisation of Oedipus in the *Oedipus* and in the *Phoenissae*?

[13] E. g., Amphitryon in Seneca's tragedy *Hercules Furens* and Plautus' comedy *Amphitruo*.

[14] In the *Agamemnon*, the dialogue between the nurse and Clytemestra, Agamemnon's wife.

[15] In the *Agamemnon*, the dialogue between Clytemestra and her daughter Electra.

References

Melanie Andresen and Nils Reiter, editors. 2024. *Computational Drama Analysis*. De Gruyter, Berlin, Boston.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable corpora: Introducing dracor, an infrastructure for the research on european drama. In *Proceedings of DH2019: Complexities*. Zenodo.

Federica Iurescia. 2021. Common ground management in roman tragic dialogues. In *Linguisticae Dissertationes. Current Perspectives on Latin Grammar, Lexicon and Pragmatics. Selected Papers from the 20th International Colloquium on Latin Linguistics (Las Palmas de Gran Canaria, Spain, June 17-21, 2019)*, pages 689–702. Ediciones Clàsicas.

Francesco Mambrini. 2005. The syntax of the heroes? a treebank-based approach to the language of the sophoclean characters. *Classics@*, 20.

Gunther Martin, Federica Iurescia, Severin Hof, and Giada Sorrentino. 2020. *Pragmatic Approaches to Drama: Studies in Communication on the Ancient Stage*. Brill, Leiden, The Netherlands.