Detecting reported speech as a token classification task: an application to Classical Latin?

Agustin Dei
Sorbonne Université - Paris, France
EDITTA (Edition, Interprétation et traduction des textes anciens)
CERES (Centre d'expérimentation en méthodes numériques
pour les recherches en Sciences Humaines et Sociales)
agustin.dei@sorbonne-universite.fr

Abstract

This paper presents the first application of an automatic token-classification approach for detecting reported speech spans in Classical Latin using transformer-based neural architectures. Focusing on Seneca the Elder's Declamatory Anthology, the study addresses the text's highly polyphonic nature, resulting from the use of reported speech. Instead of relying exclusively on sentence-level syntactic information, the proposed approach treats reported speech detection as a token-level sequence labeling problem. This enables the identification of reported speech spans extending across multiple sentences. We fine-tune three Latin neural language models —LatinBERT, LaBERTa, and PhilBERTa— for binary token-level classification and conduct experiments both with and without punctuation. The results show that RoBERTa-based models effectively identify reported speech, with LaBERTa achieving the best performance (F1 scores above 0.90).

1 Introduction

Reported speech detection in ancient classical languages (Ancient Greek and Latin) remains a task not only unsolved but also unexplored.

This paper presents a work in progress on the first method for reported speech detection in Classical Latin, developed through an automatic approach based on the fine-tuning of encoder-only Latin pre-trained language models[1]. At this stage, we base our method on a use case drawn from the declamatory anthology of Seneca the Elder (henceforth Seneca), father of the renowned philosopher. The text was written around 30 AD. This anthology is the oldest surviving record of Latin declamatory practice and brings together thousands of quotations from the practice of the *Controuersiae* (mock judicial speeches) and the *Suasoriae* (mock speeches of private advice), which were delivered in pedagogical and public contexts. In this way, the text is distinguished by its highly polyphonic nature, resulting from the use of reported speech in different modalities. Up to 116 speakers cited by Seneca have been identified by scholars (Echavarren, 2007). Due to the text's manuscript tradition, the only complete books that have been transmitted to us are the following: *Controuersiae* 1, 2, 7, 9, 10, and the book of *Suasoriae*. The whole corpus contains 96,465 tokens, with punctuation excluded.

2 State of the art and Seneca's corpus

Reported speech is a powerful linguistic tool. By choosing how to report someone's words, it is possible to reshape the original meaning of a text and influence how others perceive the cited speaker. Seneca writes his anthology of declamations with a cultural and a historical purpose: he seeks to preserve the memory of the speakers he claims to have heard by transcribing their speeches. The resulting text therefore constitutes an important stylistic testimony. The speakers are cited and integrated in a network of imitation and competition: each declaimer is positioned in relation to others, who also treat the same subjects. Seneca's own words and the speakers utterances are then mixed in one text through the implementation of reported speech. In this way, quotations in both direct and indirect speech reveal the relationships among the different declaimers and this cultural and literary activity.

The anthology is structured in two macro-sections: the *sententiae*, where long sequences of the declaimers' speeches are cited and the *diuisiones et colores*, which gather Seneca's critical comments alongside other quotations (mostly from the same speakers). These quotations may be either long or short and introduced as direct or indirect reported speech. The work also includes a preface to each book, where the author mainly focuses on

[1]The code and dataset are available at https://github.com/agu-oli/latin_reported_speech_detection.git.

Proceedings of SIGHUM 2026, pages 251–256

the presentation of a speaker.

In Latin, reported speech has been the object of linguistic studies (see Bolkestein, 1996; Shalev, 2005; Rosén, 2013; Charles et al., 2020). These studies have focused on the use of non-finite verbal constructions and subordinate clauses, the use of markers of reported speech in Latin prose, speech verbs (*uerba dicendi*), as well as deictic elements such as the adverbs *hic* and *ecce*, personal pronouns, and the use of the vocative case, (see also Dalbera, 2024).

In Seneca, an important indicator of reported speech can be the presence of a syntagm or lemmas that not only announce the reported speech cataphorically but also introduce a metalinguistic comment on the cited text; in such cases, a speech verb may not be governing the sentence. For instance, Seneca frequently uses the lemma *color*, which in this context refers to the perspective from which the declamator develops the argumentation, for introducing reported speech with a critical motivation. The lemma *color* functions often as a complement of the verb *utor*, which is the root of the phrase, as in the example *Albucius hoc* ***colore*** *usus est*[2] (*Controuersiae* 2,1,31; 7,6,14; 7,7,13; 7,7,15; 9,5,13), but sometimes the verb is elided, see *Controuersiae* 1,1,17; 10,5,17.

Also, Latin's free word order affects the structure of reported speech at the phrase level: reported speech spans can be discontinuous and therefore interrupted by governing verbs or other phrase elements that do not belong to the reported span.

Finally, as it also happens in Seneca the Elder's anthology, reported speech in Ancient Greek and Latin can include sudden transitions from indirect to direct reported speech or vice versa. Syntactically this implies that, for instance, non-finite verb constructions can be juxtaposed with finite verbal clauses[3]. Seneca's texts present examples of this reported-speech modality, as in *Controuersiae* 2,2,7 where the speech of Marullus is first cited as indirect speech through a non-finite verb construction, followed by a sudden transition to direct reported speech: *Marullus praeceptor noster licenter uerbo usus est [. . .] cum diceret* ***uxorem intellexisse mariti mendacium: et ipsa aduersus temerarios mariti iocos relusit***[4].

[2] "Albucius made use of this colour".

[3] This phenomenon that has been explained as a mixed quotation style by Maier (2012).

[4] "Marullus, our teacher, made a bold use of a word [. . .] when he said that the woman had understood the husband's lie: 'she joked back to her husband's jests'".

In the field of Computational Linguistics applied to Latin, several resources have been developed including Treebanks and neural models such as LatinBERT, PhilBERTa, LaBERTa, and morphosyntactic parsers (Sprugnoli et al., 2024). There have also been advances in Named Entity Recognition task (Erdmann et al., 2016; Beersmans et al., 2023). Nevertheless, no attention has been paid to the automatic extraction of reported speech in Classical Latin. In modern literary studies, we can cite the development of a tagger framework including recognizers for detecting four different types of reported speech in German (Brunner et al., 2020).

Since v.2.10, the Universal Dependencies (UD) introduced changes for reported speech, which has become a subtype of clausal complements such as *csubj* and *ccomp*, instead of a subtype of the *parataxis* relation.

3 Challenges

Within this context, some challenges need to be addressed in order to detect reported speech in Latin as a token classification task:

- **UD reported subtype:** The reported subtype, already integrated in UD v. 2.10, is mostly absent in Latin treebanks. Even if already included in the gold standard of the EvaLatin dependency parsing 2024 campaign, among the five Latin treebanks only the Dante treebank includes some examples of reported speech as a subtype of clausal complements.

- **Reported speech as a within-sentence or span-level problem:** if we approached reported speech as an exclusively syntactic problem, one possible solution would be to enrich the Treebanks with the reported speech subtype in order to improve current parsers. However, such an approach would limit the analysis to the sentence level and the results would be inadequate: reported speech sequences can indeed cover multiple sentences as is the case of Seneca's quotations.

- **Punctuation-based approaches:** Punctuation marks could help identify reported speech since they can function as speech type delimiters. However, a punctuation-based approach would not be suitable for Latin or other ancient languages. The manuscript transmission

of these texts is not consistent with respect to punctuation, as punctuation conventions have evolved over time. Moreover, the punctuation found in Ancient Greek and Latin texts has largely been added by modern editors and may differ between editions.

4 Dataset

In this paper, we will focus on the reported speech in the *diuisiones et colores* section of the text, since it presents a wide variety of reported speech modalities, including different reporting markers (see *supra* 1 and 2). Moreover, because the *diuisiones et colores* section also includes the author's critical voice on the declaimers being cited and on the practice of declamation itself, the identification of reported speech spans in this context constitutes a fundamental first step towards analyzing quotation modalities within the text as a whole, as well as for later stylometric analysis of the cited speeches.

For this first experiment, we prepared a small gold standard of 6,203 tokens, including punctuation, as well as a version of the same dataset without punctuation, consisting of 5,068 tokens. The gold standard contains extracts primarily from the *diuisiones et colores* sections of Books 1, 2 and 9 of the *Controuersiae*, as well as an extract from the preface to Book 2 and from the *Suasoriae* (1,4), in order to obtain a more balanced sample of the work as a whole. The text corresponds to the A. Kiessling edition, which is in the public domain [5].

The dataset contains the following features: the word form as in the sentence, the lemma, the part-of-speech (PoS) tag for each word form, and a customized binary feature indicating whether the form corresponds to a non-finite verb (NFV). Under UD, participles and infinitives are classified as VERB, together with finite forms. However, in the case of reported speech, non-finite verb (infinitives) constructions are one of the possible ways of expressing reported speech in Latin. Finally, each token in the gold dataset is annotated with a binary label indicating reported or non-reported speech, which is then the target to be predicted by the developed model [6].

Although reported speech is not a homogeneous category—at a minimum, direct, indirect, and free indirect forms can be distinguished—the present study focuses on a binary distinction between reported and non-reported macro-categories, so that future work can build on this baseline and address the internal differentiation of reported speech types.

Form	Lemma	PoS	NFV	Label
Belle	belle	ADV	0	0
de	de	ADP	0	0
hoc	hic	DET	0	0
uitio	uitium	NOUN	0	0
illius	ille	DET	0	0
Scaurus	scaurus	PROPN	0	0
aiebat	aio	VERB	0	0
,	,	PUNCT	0	0
illum	ille	DET	0	**1**
acta	ago	VERB	1	**1**
in	in	ADP	0	**1**
aurem	auris	NOUN	0	**1**
legere	lego	VERB	1	**1**

Table 1: Excerpt of token-level annotation for reported speech from *Controuersiae* 2,1,39 including linguistic features and reported speech label from two different sentences. Labels indicate reported speech (1) and non-reported speech (0). The two sentences are not split.

5 Method

We approach reported-speech detection in Latin as a token-level sequence labeling problem, a perspective that enables the identification of spans that cover multiple sentences. This entails labeling tokens as reported or non-reported speech. In this manner, our approach can exploit features available at the sentence and token levels, such as lemmas, part-of-speech tags, and the presence of non-finite verbal forms, while analyzing them at the span-level in order to determine whether each token belongs to reported or non-reported speech.

This first experiment consisted of fine-tuning three encoder-only Latin neural models, Latin BERT (Bamman and Burns, 2020) and LaBERTa, which are monolingual, and the multilingual model PhilBERTa (Riemenschneider and Frank, 2023). All three were fine-tuned for binary token-level classification of textual sequences as reported or non-reported speech in Seneca the Elder's anthology, using token-level features (lemma, PoS, NFV) encoded as embeddings[7] and concatenated with the model's contextual token representations.

[5] Seneca, L. A. (1922). *Annaei Senecae oratorum et rhetorum sententiae, diuisiones, colores* (A. Kiessling, ed.). Leipzig: B. G. Teubner.

[6] Discontinuous spans (e.g. speech verb interrupting a reported sequence) are handled by labeling only the quoted material and treating interruptions as non-reported tokens.

[7] Lemma embeddings were initialized from the pretrained model's vectors.

Model	Punctuation	Acc	Prec	Rec	F1	PR AUC
LatinBERT	with punct.	0.744	0.769	0.892	0.826	0.803
LatinBERT	without punct.	0.681	0.737	0.855	0.791	0.822
LaBERTa	with punct.	**0.877**	0.921	**0.897**	**0.909**	**0.949**
LaBERTa	without punct.	0.873	**0.935**	0.883	0.908	0.933
PhilBERTa	with punct.	0.849	0.891	0.886	0.889	0.927
PhilBERTa	without punct.	0.847	0.910	0.870	0.890	0.937

Table 2: Performance of three fine-tuned models with and without punctuation on the test set.

We fine-tuned all three models with a learning rate of 1e-5, batch size 8, and weight decay 0.01. All experiments were run with a fixed random seed (42) (source code link in footnote 1). Checkpoints were selected by validation loss. Evaluation is performed at the token level after alignment of word-level labels to subword tokens.

Preprocessing choices: we conducted two experiments per model on two preprocessing choices on the same dataset, with and without punctuation.

Dataset splitting strategy: the span-level approach presented a new challenge for splitting the dataset in a training, validation, and test sets. Since the dataset consists of sequences of text containing reported speech spans which can extend across more than one sentence, it is not possible to adopt a sentence-level splitting strategy. Such a split would not be appropriate and would prevent the model from learning to detect sequences of tokens belonging to the same speech type across multiple sentences. To address this issue, the dataset was then organised into 18 sections, which do not necessarily coincide with the book and chapter divisions in Seneca's work. Each section contains both labels; section boundaries were defined during annotation by distinguishing coherent units.

In short, the dataset was split at the level of these sections. We randomly assign 70% of the sections to training, 20% to validation, and 10% to test sets using a fixed random seed (42). No section appears in more than one split.

6 Results

The results show (see Table 2) that transformer-based Latin models are able to identify reported speech at the token level even after a first experiment in a small data setting and without relying on sentence-level segmentation.

Among the three models, LaBERTa consistently achieves the best performance, with F1 scores above 0.90 and the highest PR AUC value, indicating robust discrimination between reported and non-reported tokens despite the limited size of the dataset. LatinBERT exhibits higher recall than precision in the test set, indicating sensitivity to reported speech spans, a behavior that can be desirable for literary and stylometric analysis because it can be refined via a post-processing step. However, LatinBERT shows clearly lower performance than RoBERTa-based models, which could reflect differences in tokenization strategies (see Bamman and Burns, 2020). In contrast, both LaBERTa and PhilBERTa (see Riemenschneider and Frank, 2023), which share the same tokenization method, demonstrate more stable performance, indicating better generalization.

LaBERTa's superior performance could be due to its pretraining data, which is limited to the *Corpus Corporum*. In future work, it should be investigated whether this training corpus provides more stable lexical and contextual representations for Latin. By contrast, PhilBERTa achieves very similar performance to LaBERTa, even though it is a multilingual Greek–Latin model and was pretrained on more heterogeneous data: the Open Greek & Latin project, the CLARIN corpus of Greek Medieval Texts, the *Patrologia Graeca*, the *Corpus Corporum* and also Project Gutenberg texts.

The comparison between punctuation-aware and punctuation-free datasets does not reveal considerable differences between LaBERTa and PhilBERTa performances. While the presence of punctuation slightly coincides with a recall improvement, LaBERTa and PhilBERTa models trained without punctuation remain highly competitive, which suggests that lexical and contextual information play a more important role than punctuation for reported speech detection in Latin. This is particularly relevant for the future of this study on a larger dataset and for the deployment of a reported speech detection tool on Latin corpora given the inconsistency

and instability of punctuation in the manuscript tradition and editorial practices.

7 Error analysis

At the multi-word level, subordinate clauses whose structure closely resembles that of reported speech remain difficult to distinguish from genuinely reported sequences. All models appear to overfit to reported clauses introduced by the Latin particle and subordinating conjunction *an*, which are frequent in the Seneca corpus because they are used to indicate which points are addressed by the speakers (the *diuisio* of the speech). Clauses introduced by *an* are also used by Seneca in his own authorial voice to analyze how this is addressed by the various speakers[8].

The proximity of tokens to a speech verb may also lead to false positives across all models, as in *Controuersia* 1,1,15[9], where the speech verb is followed by a direct object and then by a clausal modifier of that object which constitutes the proper reported speech span. The object is incorrectly tagged as part of the reported segment. This example illustrates a construction that is particularly frequent in Seneca, in which an object or oblique nominal is followed by a clausal modifier that introduces the reported-speech sequence. It functions as a transition between the author's critical comments on the speakers and the strictly reported text. The question of how best to handle these two constructions will be addressed in future work.

In both punctuation and non-punctuation training settings, the models tend to tag adverbs such as *deinde* ("then"), from Seneca's voice, which interrupt the quotations, as reported speech. By contrast, while LaBERTa and PhilBERTa models trained on a dataset containing punctuation correctly identify the speech verb *adicio* ("to add") when it interrupts a reported sequence, the models trained without punctuation and LatinBERT fail to recognise it in this function.

Furthermore, models trained without punctuation show a tendency to produce isolated false-negative tokens within an otherwise correctly identified reported span. Models trained with punctuation, on the other hand, exhibit errors on punctuation marks both within reported spans and at span boundaries. Both types of error can be addressed in a post-processing step.

8 Conclusions and future work

These initial results support the feasibility of token-level reported speech detection in Latin, as well as the advantages of a span-based segmentation strategy, despite the small size of the dataset used in this first experiment. A span-level approach based on fine-tuning transformer-based Latin models, to detect reported speech sequences at the token level, enables the effective extraction of reported speech sequences that extend beyond sentence boundaries.

Future work will focus on enlarging the dataset and further examining how linguistic features and contextual cues affect the performance of token-level reported speech detection, with the aim of improving model robustness. It will also involve testing the approach on the internal types of reported speech and conducting further error analysis in order to develop a post-processing strategy that could include rule-based approaches or syntax-based heuristics.

Just as reported speech is a powerful linguistic tool in Seneca's text, advances in its detection can lead to new contributions to information extraction from Latin texts—including non-literary ones—in which other sources are cited, and may also provide inspiration for comparable frameworks in other languages.

Acknowledgments

This research was funded by the Émergence project, at Sorbonne University and Rouen Normandie University (France), "SenecAI: Artificial Intelligence for the Stylistic and Authorial Analysis of Seneca the Elder's Declamatory Anthology"[10].

References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *Computing Research Repository*, arXiv:2009.10053.

Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training and evaluation of named entity recognition models for classical latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 1–12, Varna, Bulgaria. INCOMA Ltd.

[8] *An abdicari debeat per haec quaesiit: an...* "On the question 'should he be disinherited', he examined..." (*Controuersia*. 1,1,13)

[9] *et sua figura* ***dixit*** *omnia propter quae uelle deberet. 'Quare ergo abdicas?'*"and through his figure he said all the reasons why he should want it : 'Why do you disinherit me?'"

[10] https://editta.hypotheses.org/senecai

A. Machtelt Bolkestein. 1996. Reported speech in Latin. In Theo Janssen and Wim van der Wurff, editors, *Reported Speech: Forms and Functions of the Verb*, pages 121–140. John Benjamins Publishing Company, Amsterdam.

Ann Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020. To BERT or not to BERT: Comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In *Proceedings of SwissText/KONVENS*.

Lise Charles, Frédérique Fleck, and Lyliane Sznajder. 2020. Les formes de discours rapporté : repérage et interprétations. *Lalies*. Langues et littérature : actes des sessions de linguistique et de littérature.

Joseph Dalbera. 2024. Verbes introducteurs et stratégies d'introduction du discours direct dans la narration romanesque latine (le Satyricon de Pétrone et les Métamorphoses d'Apulée). In Concepción Cabrillana, editor, *Recent Trends and Findings in Latin Linguistics*, pages 503–520. De Gruyter, Berlin, Boston.

Arturo Echavarren. 2007. *Nombres y personas en Séneca el Viejo*. Ediciones Universidad de Navarra.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan. The COLING 2016 Organizing Committee.

Emar Maier. 2012. Switches between direct and indirect speech in ancient greek. *Journal of Greek Linguistics*, 12(1):118–139.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Hannah Rosén. 2013. About non-direct discourse: Another look at its parameters in latin. *Journal of Latin Linguistics*, 12(2):231–268.

Donna Shalev. 2005. Action nouns in reports of speech acts. *Journal of Latin Linguistics*, 9(2):719–730.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the evalatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Turin, Italy. ELRA and ICCL.

Annex: inter-annotator agreement

100% of the data were tagged independently by two experts[11], leading to a Cohen's kappa of 0.92. Disagreements were resolved through joint review to produce the final gold standard.

[11] The author of this work and Pierre Belenfant.