
Orosius' *Histories*: A Digital Intertextual Investigation into the First Christian History of Rome

Greta Franzini
gfranzini@etrap.eu
Georg-August-Universität Göttingen, Germany

Marco Büchler
mbuechler@etrap.eu
Georg-August-Universität Göttingen, Germany

Introduction

The research described in these pages is made possible by openly available Classical texts and linguistic resources. It aims at performing semi-automatic analyses of Paulus Orosius' (385-420 AD) most celebrated work, the *Historia adversum Paganos Libri VII*, against its sources. The *Histories*, as this work is known in English, constitute the first history (752 BC to 417 AD) to have been written from a Christian perspective. To do so, Orosius drew from earlier and contemporary christian and pagan authors, providing a rich narrative fraught with intertextual references to poetry and prose alike.

Orosius' vast network of references challenges automatic text reuse detection tasks both qualitatively and quantitatively. In fact, information retrieval algorithms face differences in reuse style –from verbatim quotations to paraphrase and allusions (Navarro, 1991)– and millions of words to sift through. To understand how Orosius reused texts, it is necessary to detect, extract, classify and evaluate all references and compare them to their sources, mindful of the balance between the *precision* of the results and their *recall* or number.

Related Work

Existing research on Orosius' sources for the *Histories* is scattered and often focusses on his relation to one author or work only, albeit acknowledging the full spectrum of sources (e.g. Coffin, 1936; Sihler, 1887). The size of the source- texts (see Table 1) makes it extremely difficult to produce a

comprehensive and detailed manual exploration of *all* of Orosius' references.

“It would be burdensome to list all of the Vergilian echoes [...]” (Coffin, 1936: 237)

What Coffin describes as “burdensome” can be accomplished with machine assistance. To the best of our knowledge, no existing study, traditional or computational, has quantified and analysed the reuse habits of Orosius.

The *Tesserae* project, which specialises in allusion detection, is the most similar to the research presented here (Coffee, 2013), with the difference that it does not yet contain the text of Orosius nor many of its sources, and that the results are automatically computed without user input.

In contrast, our approach, TRACER (Büchler et al., 2017), offers complete control over the algorithmic process, giving the user the choice between being guided by the software and to intervene by adjusting search parameters. In this way, results are produced through a critical evaluation of the detection.

Research Questions and Goal

Our research began with the following questions: how does Orosius adapt Classical authors? Can we categorise his text reuse styles and what is the optimal precision-recall retrieval ratio on this large historical corpus? How does detection at scale affect computational speed?

This project tests the stability of historical text reuse detection on a corpus of Latin authors where Orosius is our target text. We evaluate our computed results against known reuses published in commentaries to Orosius, thus corroborating existing findings but also potentially uncovering previously unnoticed reuse. In so doing, we refine our workflow and resources in order to advance historical text reuse detection for Latin.

Data

All of the public-domain works for this study were downloaded from *The Latin Library*. We chose this collection over other analogous resources as it provides clean and plain texts (.txt), the format required by our text reuse detection machine TRACER. Table 1 outlines the authors and works under investigation in chronological order. To give an idea of the size of the texts, the ‘Tokens’ column provides a total word-count for each work; the ‘Types’ column provides the total number of unique words; and the ‘Token-Type Ratio’ shows how often a type occurs in the

text (e.g. a TTR of 3 indicates that for every type in a text there are three tokens on average. Generally, the higher the ratio the less linguistic variance in a text). This table reveals the language and challenges we should expect when detecting reuse. For instance, Caesar, Lucan and Tacitus share similar text lengths but Caesar has a higher TTR; this tells us that Caesar’s text has less linguistic variety than Lucan and Tacitus. Conversely, if we look at Suetonius in comparison to Lucan and Tacitus, we notice a larger text but a similar TTR. This indicates a high linguistic variance in Suetonius’ text, and one that can prove challenging for text reuse detection.

Author [date]	Latin Style	Work (type)	Tokens	Types	Token-Type Ratio (TTR)
Julius Caesar [100-44BC]	Classical	De Bello Gallico (prose)	51,723	11,100	4.65
Vergil [70-19 BC]	Classical	Aeneid (epic poem)	63,715	16,799	3.79
Vergil [70-19 BC]	Classical	Georgics (epic poem)	14,175	6,974	2.03
Livy [59 BC-17 AD]	Classical	Ab urbe condita (prose)	507,120	50,774	9.98
Lucan [39-65 AD]	Classical	De Bello Civili sive Pharsalia (epic poem)	51,033	14,780	3.45
Tacitus [56-117 AD]	Classical	Historiae (prose)	51,417	15,347	3.35
Suetonius [69-ca.130 AD]	Classical	De Vita Caesarum (biography)	71,040	21,565	3.29
Florus [74-ca. 130AD]	Classical	Epitome de T. Livio Bellorum Omnium Annorum DCC Libri Duo (prose)	26,750	9,181	2.91
Entropius [n.d.-ca. 399AD]	Late	Breviarium ab Urbe Condita (prose)	18,873	5,575	3.38
St. Augustine [354-430AD]	Late (Ecclesiastical)	De civitate Dei contra Paganos (prose)	274,720	35,430	7.75
Orosius [385-420 AD]	Late (Ecclesiastical)	Historia adversum Paganos (prose)	74,929	19,748	3.79
Total tokens (words to be processed):			1,205,495		

Table 1. Overview of analysed texts.

Reuse Styles

Orosius employs a variety of reuse styles, ranging from verbatim quotations to allusions and paraphrase (Navarro, 1991). The reuses are as short as two words (*ibid.*) or as long as sixty-five words, and sometimes invert the word order of the original text (Elerick, 1994).

Methodology

Our workflow makes use of three resources: first, a *TreeTagger* Latin language model for Part-of-Speech (PoS) tagging and lemmatisation (Schmid, 2013). We chose to work with *TreeTagger* as, unlike other taggers, it comes with a pre-trained model for Latin (trained by Marco Passarotti). Since submitting this abstract, we also began experimenting with the *LemLat* morphological analyser. Secondly, we used the *Latin WordNet* lemma list and synonym set to support the detection of paraphrase and paradigmatic relations; and *TRACER*, our text reuse detection machine (see also, the [list of TRACER’s 700 algorithms](#)).

First, the data is acquired and prepared: the texts are downloaded, semi-automatically cleaned (by "cleaning" we mean the removal of footnotes, section

numbering, special characters and typos in the texts) and, where possible, spelling variants are normalised. Next, the texts are lemmatised and tagged for PoS. We then run TRACER with different parameters in order to define the diversity of the reuses in the corpus.

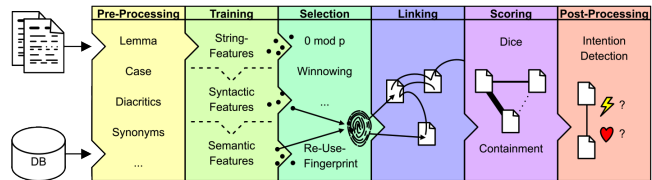


Figure 1: The six-step pipeline of TRACER (from left to right).

TRACER can split a detection task into six sub-tasks, each containing parameters that users can customise or (de)activate depending on the type of detection required (see Figure 1). The reader will notice that the *Pre-processing* step also contains lemmatisation. This does not mean that TRACER can lemmatise any text but it currently supports lemmatised input data from the *Stanford CoreNLP* English lemmatiser and the *TreeTagger* Latin lemmatiser.

Results

Tagged and lemmatised text accounts for 93.1% of the tokens in the corpus. A 7% of words could not be lemmatised due to typos in the text (e.g. missing white-spaces), which we are manually or semi-automatically correcting; similarly, some words could not be successfully tagged, as the lemmas are not included in the *TreeTagger’s* parameter file (e.g. named entities).

To perform detection with TRACER, all texts were initially segmented by sentence. The average sentence length measured across the whole corpus is thirty-one words per sentence. A first detection task between Orosius and all other authors was conducted at the sentence level. However, this failed due to the presence of very short text reuses. For this reason, the segmentation was changed to a moving window of ten words, thus giving TRACER smaller units to process. In the *Selection* step (see Figure 1), we first experimented with max-pruning (i.e. the removal of high frequency words) but eventually settled on a PoS-based selection, which considered nouns, verbs and adjectives as more relevant features than function words, thus significantly increasing the recall and the overall quality of the results.

For the *Scoring* (see Figure 1), we used the resemblance score, which measures the ratio of overlapping

features with the overall unique set of features of two alignment candidates. Figure 2 illustrates the results of this detection process: over 45% of reuses identified in Orosius overlap with the source texts by four words, and that roughly 93% of all candidates have overlaps of 3, 4 or 5 words, indicating a fragmentary reuse style rather than block-copying. This detection task took approximately 30 hours to compute.

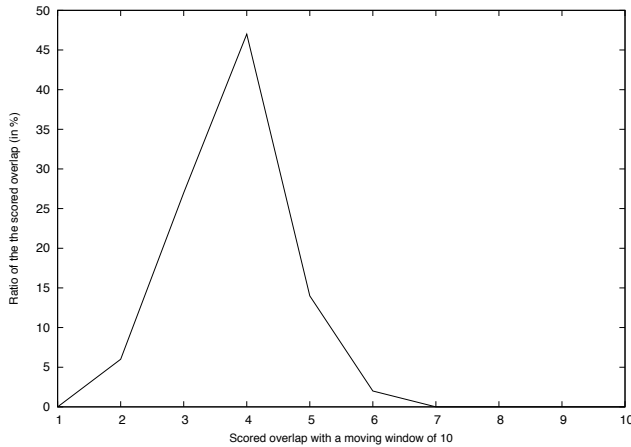


Figure 2. In this plot the x-axis represents a window of 10 words, while the y-axis the occurrence of the overlap in percentage. Over 45% of text reuse in Orosius overlaps with the source texts by four words.

A second TRACER experiment was run between Orosius and Tacitus to explore and evaluate the results of the detection on a smaller scale. Commentaries to Orosius reveal the presence of fifteen reuses of Tacitus, ten of which refer or allude to text that no longer survives. This means that TRACER can only try to match five known reuses, which differ in style. In this experiment, we used a moving window of fifteen words and synonym replacement in order to identify paraphrase as well as verbatim quotations. TRACER identified forty reuses, of which thirty-six false positives and two new finds. Figure 3 illustrates these results.

False positives	
Similarities identified by TRACER that don't constitute reuse	
Orosius	Tacitus
Locus: 4.19	2.43
Text: <i>... ad Africum ...</i>	<i>... necesse est ...</i>
Translation: ...	
Reasoning: ...	

New finds	
Similarities identified by TRACER that deserve further study and assembly semantic and syntactic detection.	
Orosius	Tacitus
Locus: 5.5	2.76
Text: <i>... quae ...</i>	<i>... in ...</i>
Translation: ...	
Reasoning: ...	

Figure 3. Top: an example of a false positive produced by TRACER in detecting reuse between Orosius and Tacitus. Bottom: two new finds yielded by TRACER, an analogy and

a potential reuse. Colours match up the similarities between the aligned candidates.

Limitations and Future Work

The retrieval accuracy of TRACER partly depends on the accuracy of the trained models of *TreeTagger* and the *Latin WordNet* data. An error analysis is needed in order to verify the accuracy of our cleaned and automatically-tagged data, and to determine the effect of this incorrect tagging on text reuse detection. Depending on the outcome of this analysis, we will consider re-tagging our corpus with a more advanced tagger, such as *LemLat* (Passarotti, 2007) and/or *LatMor*, or even training a tagger on the different types of Latin constituting our corpus.

We are currently running TRACER comparisons between Orosius and each of the other source authors in our corpus to verify the presence of previously unknown reuse, corroborate known reuse and improve our detection techniques.

Additionally, we plan on comparing Orosius against non-reused authors, such as Plautus or Apuleius, to examine TRACER's performance on "negative" texts.

Acknowledgements

This research is funded by the German Federal Ministry of Education and Research (grant 01UG1509).

Bibliography

- Büchler, M., Franzini, G., Franzini, E. Moritz, M. (2017 forthcoming). "TRACER - a multilevel framework for historical Text Reuse detection." *Journal of Data Mining and Digital Humanities - Special Issue on Computer Aided Processing of Intertextuality in Ancient Languages*.
- Coffee, N., Koenig, J. P., Poornima, S., Forstall, C. W., Ossewaarde, R., Jacobson, S. L. (2013). "The Tesseræ Project: intertextual analysis of Latin poetry." *Literary and Linguistic Computing*, 28(2): 221–28. DOI: 10.1093/ljc/fqs033
- Coffin, H. C. (1936). "Vergil and Orosius." *The Classical Journal*, 31(4): 235–41. Available at: <http://www.jstor.org/stable/3290976> (Accessed: 13 October 2016)
- Elerick, C. (1994). "How Latin Word Order Works." *Journal of Latin Linguistics*, 4(1): 99–118. DOI: 10.1515/joll.1994.4.1.99
- Fear, T. A. (2010). *Orosius: Seven Books of History against the Pagans*. Liverpool University Press.

Navarro, M.A.R. (1991). "Historiadores y poetas citados en las Historias de Orosio: Livio y Tácito, Virgilio y Lucano." *Fortunatae: Revista canaria de filología, cultura y humanidades clásicas*, (2): 277-86. Available at: <https://dialnet.unirioja.es/descarga/articulo/163829.pdf> (Accessed: 13 October 2016).

Passarotti, M. (2007). "LEMLAT. Uno Strumento per la Lemmatizzazione Morfologica Automatica del Latino." *Journal of Latin Linguistics*, 9(3): 107-128. DOI: 10.1515/joll.2007.9.3.107

Schmid, H. (2013). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In D. B. Jones and H. Somers (eds), *New Methods in Language Processing*. London and New York: Routledge, pp. 154-64. Available at: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf> (Accessed: 28 October 2016).

Sihler, E. (1887). "The Tradition of Caesar's Gallic Wars from Cicero to Orosius." *Transactions of the American Philological Association (1869-1896)*, 18: 19-29. Available at: <http://www.jstor.org/stable/2935772> (Accessed: 14 October 2016).