LATIN TEXT REUSE DETECTION AT SCALE

AN AUTOMATIC TEXT REUSE INVESTIGATION INTO THE FIRST CHRISTIAN HISTORY OF ROME

Greta Franzini and Marco Büchler

25 January 2017

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

1. Introduction

2. Research questions

3. Challenges

4. Methodology

5. Results

6. Research value and output

INTRODUCTION

PAULUS OROSIUS AND HIS HISTORIES

Paulus Orosius [ca. AD 375-418]

- Roman historian and a Christian from Spain;
- Student of St Augustine [AD 354-430].

Historiae adversus Paganos = *Histories against the Pagans**

- First Christian history of Rome;
- Complementary to St Augustine’s *De civitate Dei contra Paganos*;
- Defense against pagan accusations that Rome’s was decline caused by the advent of Christianity;
- Heavily reuses both pagan and Christian authors to reject pagan claims.

*Paganism = pantheism, polytheism, non-Christian.
*Christianity = monotheism. Declared *permitted religion* by Constantine the Great in 313 (Edict of Milan); declared official religion of the Empire by son Constantius II in 350.

PRIMARY SOURCES

1. *[ed.] [tr.]* Arnaud-Lindet, M. P., *Orose: Histoires contre les païens*, 3 vols, Collection des Universités de France, Paris: Les Belles Lettres, 1990–1991.
2. *[ed.]* Zangemeister, K., *Pauli Orosii historiarum adversum paganos libri VII*; accedit eiusdem, Liber apologeticus, Corpus Scriptorum Ecclesiasticorum Latinorum 5, Vienna, 1882.
   Internet Archive: `https://goo.gl/SnJJHy`
3. *[ed.]* Migne, J. P., *Pauli Orosii Hispanorum Chronologorum Opera Omnia*, Patrologia Latina Cursus Completus 31, Paris, 1846.
   Internet Archive: `https://goo.gl/AWRP8i`
4. *[ed.]* Zangemeister, K., *Pauli Orosii historiarum adversum paganos libri VII*, Bibliotheca scriptorum Graecorum et Romanorum Teubneriana, Leipzig: Teubner, 1889.
   Internet Archive: `https://archive.org/details/pavliorosiihist01orosgoog`
   Attalus.org: `http://www.attalus.org/latin/orosius.html`

RESEARCH QUESTIONS

RESEARCH QUESTIONS: BRIDGING CLOSE & DISTANT READING

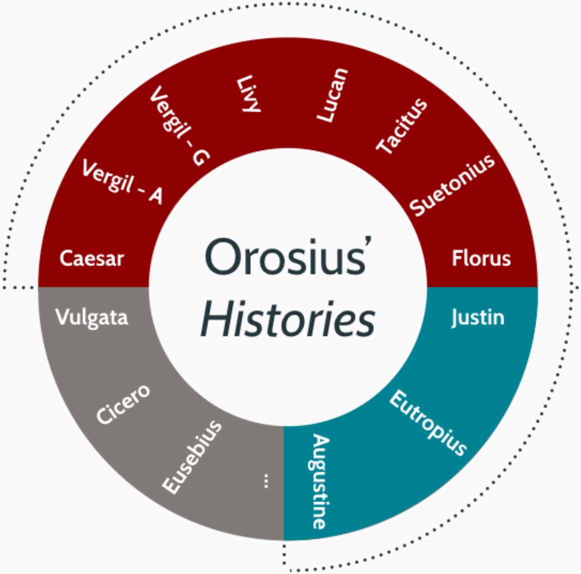
- **Close Reading**
  How does Orosius reuse text in order to build his defense?

- **Close + Distant Reading**
  Can we quantify and categorise Orosius’ reuse diversity (taxonomy)?

- **Distant Reading**
  How does a large corpus affect automatic text reuse detection and its performance?

CHALLENGES

CHALLENGES: DIACHRONIC CORPUS



~1.3M words

Classical Latin

Late/Ecclesiastical

Future additions

CHALLENGES: REUSE DIVERSITY

Orosius:

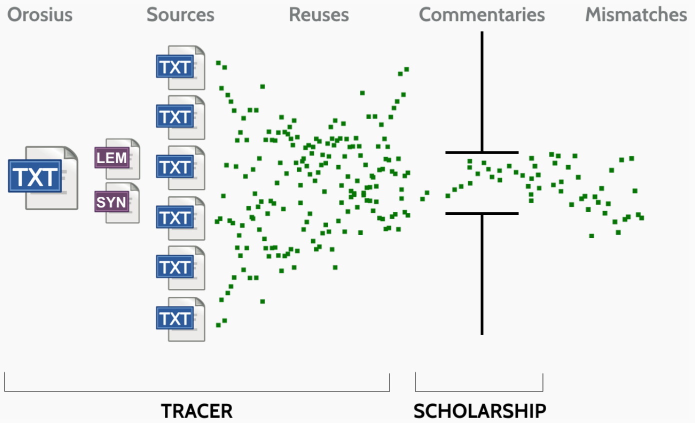
- reuses two words to entire sentences or even paragraphs;
- quotes word-for-word (i.e. *verbatim*), near-verbatim or (very) loosely;
- doesn't always cite the original author;
- occasionally misattributes words because citing from memory;
- reuses text that doesn't survive.

Nec tibi ***cura canum*** *fuerit* ***postrema*** (Georg. 3.404 - poetry)
[= Nor be your dogs last cared for]

non est tamen ***canum cura postrema*** (Oros. 1.1. - prose)
[= Dogs are not to be cared for last]

METHODOLOGY

METHODOLOGY: TRACER VS. COMMENTARIES



- How do the computed results compare to existing scholarship?
- Has TRACER identified reuses that existing scholarship hasn't?
- Has existing scholarship identified reuses that TRACER hasn't?

RESULTS

RESULTS: TRACER ON ENTIRE CORPUS

Window size: 10 words; Feature density: 0.8; Highest reuse overlap: 4 words; Computation: ca. 48 hours.

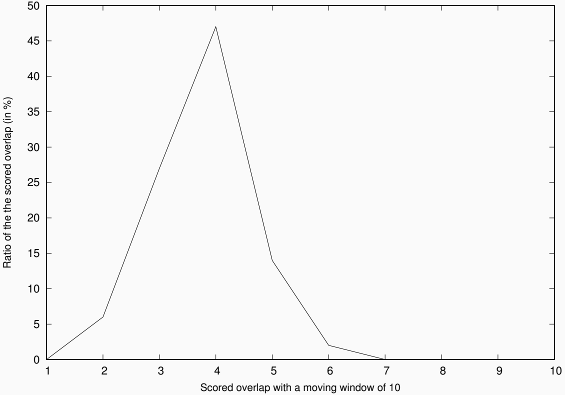
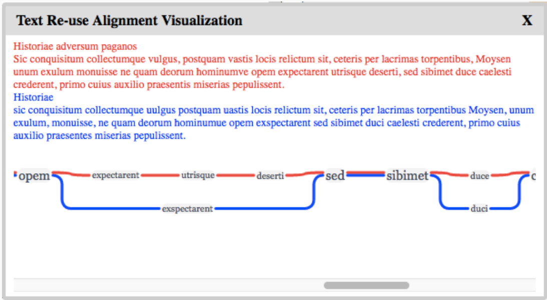


Figure 1: Orosius' general reuse pattern, across the entire corpus.

RESULTS: OROSIUS' REUSE OF TACITUS' HISTORIAE

- Reuses documented in primary sources (*precision*): 15 (= 12+3?)
- Reuses identified by TRACER (*recall*)*: 55
    - verbatim;
    - near-verbatim: "true" and "false" (spelling conventions);
    - no similarity: why? Synonym replacement? PoS? Feature density?



Text Re-use Alignment Visualization X

Historiae adversum paganos
Sic conquisitum collectumque vulgus, postquam vastis locis relictum sit, ceteris per lacrimas torpentibus, Moysen unum exulum monuisse ne quam deorum hominumve opem expectarent utrisque deserti, sed sibimet duce caelesti crederent, primo cuius auxilio praesentis miserias pepulissent.

Historiae
sic conquisitum collectumque uulgus postquam uastis locis relictum sit, ceteris per lacrimas torpentibus Moysen, unum exulum, monuisse, ne quam deorum hominumue opem exspectarent sed sibimet duci caelesti crederent, primo cuius auxilio praesentes miserias pepulissent.

Figure 2: Orosius 1.10, Tacitus 5.3.

*__Detection parameters__: moving window of 15 words; 0.8 feat. density; synonym replacement. Comparing 51,417 (T) against 74,929 words (OR). Computation: ca. 1 hour.

RESEARCH VALUE AND OUTPUT

RESEARCH VALUE AND OUTPUT

Research contribution

- Better understanding of Orosius' reuse behaviour.
- Detection strategy refinement; max extraction with min algorithms.
- Better understanding of degree of influence of (noisy) text on computed results.
- Refinement of existing linguistic resources towards Gold Standard for Latin lemmatisation + PoS-tagging.

Research data-sets

- Reuse pairs manually *and* computationally identified in Orosius.
- The cleanest `.txt` corpus.
- PoS-tagged+lemmatised corpus.

CONTACT

Visit us

🌐 http://www.etrap.eu

✉ contact@etrap.eu

Stealing from one is plagiarism, stealing from many is research

(Wilson Mitzner, 1876-1933)

eTRAP

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

SPONSORED BY THE

Federal Ministry
of Education
and Research

ACKNOWLEDGMENTS

The authors wish to thank Marco Passarotti and Paolo Ruffolo (CIRCSE) for their continuous support in parsing the corpus with *LemLat*.

LITERATURE

LITERATURE

- Büchler, M. *TRACER: Text Reuse Detection Machine*. At: `http://www.etrap.eu/research/tracer/`
- Jänicke, S., Franzini, G., Faisal, C., Scheuermann, G. (2015) ‘On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. A State-of-the-Art (STAR) Report’, In: (Proceedings) *EuroVis 2015: The EG/VGTC Conference on Visualization*. Cagliari, May 2015, 25-29. DOI: `10.2312/eurovisstar.20151113`
- *LemLat 3.0: Morphological Analyser and Lemmatiser for Latin*. CNR-ILS, UCSC-CIRCSE, Italy. At: `http://www.lemlat3.eu/`
- Minozzi, S. (2009) *The Latin WordNet Project*, Innsbrucker Beiträge zur Sprachwissenschaft, vol. 137, pp. 707–716. Institut für Sprachen und Literaturen der Universität Innsbruck, Innsbruck.
- Passarotti, M. (2004) ‘Development and perspectives of the Latin morphological analyser LEMLAT’, in A. Bozzi, L. Cignoni and J. L. Lebrave (eds.) *Digital Technology and Philological Disciplines, ≪Linguistica Computazionale≫*, XX-XXI, pp. 397-414.
- Budassi, M., Passarotti, M. (2016) ‘Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon’, in *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, Berlin, Germany, The Association for Computational Linguistics, pp. 90-94.

MORE INFORMATION

CORPUS STATISTICS

Author [date]	Work (type)	Tokens	Types	TTR
Caesar [100-44BC]	De Bello Gallico (prose)	51,723	11,100	4.65
Vergil [70-19 BC]	Aeneid (epic poem)	63,715	16,799	3.79
Vergil [70-19 BC]	Georgics (epic poem)	14,175	6,974	2.03
Livy [59 BC-17 AD]	Ab urbe condita (prose)	507,120	50,774	9.98
Lucan [39-65 AD]	De Bello Civili sive Pharsalia (epic poem)	51,033	14,780	3.45
Tacitus [56-117 AD]	Historiae (prose)	51,417	15,347	3.35
Suetonius [69-ca.130 AD]	De Vitis Caesarum (biography)	71,040	21,565	3.29
Florus [74-ca. 130AD]	Epitome de T. Livio Bellorum Omnium Annorum DCC Libri Duo (prose)	26,750	9,181	2.91
*Justin [3rd century]	Historiarum Philippicarum T. Pompeii Trogi Libri XLIV in Epitomen Redacti (prose)	61,256	15,134	4.04
Eutropius [n.d.-ca. 399AD]	Breviarium ab Urbe Condita (prose)	18,873	5,575	3.38
Augustine [354-430AD]	De civitate Dei contra Paganos (prose)	274,720	35,430	7.75
Orosius [385-420 AD]	Historia adversum Paganos (prose)	74,929	19,748	3.79
				Total tokens (words to be processed): 1,266,751

Table 1: Token-type ratio across the corpus (January 2017).

LICENCE