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

Greta Franzini and Marco Büchler
Digital Humanities Conference 2017, Montréal, Canada
1. Introduction

2. Research questions

3. Challenges

4. Methodology

5. Results

6. Conclusions and Next steps
INTRODUCTION
Greta Franzini, Classicist
Marco Büchler, Computer scientist (NLP)

Members of the eTRAP Early Career Research Group, an interdisciplinary team funded by the German Ministry of Education and Research (BMBF) and focussing on Automatic Text Reuse Detection.

Text Reuse: written repetition or borrowing of text.

http://www.etrap.eu
Our research goal
Advance technology for automatic text reuse detection in historical texts through the study of intertextuality. Philological discoveries are an added bonus!

**Q:** How do we advance the technology?

**A:** We test it on as many different text reuse scenarios or *styles* as possible.
**Figure 1**: Different reuse styles.

Orosius is an **ideal case study** for Latin as he draws from **500 years worth** of historical accounts providing **many different reuse styles** in one place.
Paulus Orosius [ca. AD 375-418]

- Roman historian and Christian from Spain;
- Wrote in Latin.

He wrote the *Historiae adversus Paganos* (*Histories against the Pagans*)

- First *Christian* history of Rome;
- Defense against pagan accusations that Rome’s ruin had been caused by the advent of Christianity;
- Heavily reuses both pagan and Christian authors to support his argument (*1688 known* instances of reuse: *18 authors* and *27 works*).
RESEARCH QUESTIONS
1. Can we automatically identify the 1688 instances of Orosian reuse reported by scholars ("gold standard")?

2. How does Orosius’ reuse diversity affect question 1?

3. What does this research tell us about the state-of-the-art of automatic text reuse detection in historical, and particularly Latin, texts?
CHALLENGES
Big Data: data that cannot be manually processed due to its large size.

- The corpus is over 2M words in size;
- The corpus contains both poetry and prose, as well as Classical (1st BC-3rd AD), Late (3rd-6th AD) and Ecclesiastical/Medieval Latin (6th/8th AD-onward) - differences in style.
Orosius:

- reuses text from two words to entire sentences or even paragraphs;
- quotes word-for-word (verbatim), near-verbatim or (very) loosely;
- doesn’t always cite the original author;
- occasionally misattributes text as probably citing from memory;
- deliberately distorts text to fit his argument;
- reuses text that no longer survives.

In many cases, editors can’t tell whether there is reuse or not.

No single algorithm can extract all of this!
CHALLENGE: RESOURCES USED ARE INCOMPLETE

- **LemLat 3** (Ruffolo, Passarotti): morphological analyzer & lemmatizer
  - lexical basis resulting from the collation of three Latin dictionaries (40,014 lexical entries; 43,432 lemmas):

- **Latin WordNet** (Minozzi): 9,124 lemmas; 8,973 synsets; 25,908 word senses

- **BabelNet API** (Navigli): 2,471,700 lemmas; 2,468,797 synsets (we use nouns & verbs only); 2,641,410 word senses

Quality in input = quality in output.
METHODOLOGY
Figure 2: TRACER contains ca. 700 algorithms.

http://www.etrap.eu/research/tracer
Methodology: Results Evaluation

Detection #1

TRACER results

1688 known reuses
Detection #1 → 1688 known reuses → True Positives
METHODOLOGY: RESULTS EVALUATION

Detection #1 TRACER results

1688 known reuses

False Negatives

True Positives
METHODOLOGY: RESULTS EVALUATION

Detection #1 → 1688 known reuses → False Negatives
                                      → True Positives
                                      → False Positives

TRACER results
METHODOLOGY: RESULTS EVALUATION

Detection #1

TRACER results

1688 known reuses

→ False Negatives

→ True Positives

→ False Positives

Errors

→ Lemmatization

→ Thesaurus

→ TRACER

→ Dirty text

NLP research
METHODOLOGY: RESULTS EVALUATION

Detection #1

TRACER results

1688 known reuses

False Negatives

False Positives

True Positives

Errors

Lemmatization
Thesaurus
TRACER
Dirty text

New parallels

Unnoticed reuse
Unintended reuse

NLP research

Classics research
RESULTS
• Given previous scholarship spanning 1600 years, we don’t expect to find many new reuses.

• We usually expect an average of 5-10 new reuses. Colleague in Coptic Studies recently found 14 new reuses with TRACER.
Gold Standard - Experts tell us Orosius reuses Tacitus 15 times*:

- 10 we can’t trace because they refer to lost books;
- 5 remaining* - 1 uncertain reuse:

Figure 3: CSEL reports 4 reuses of Tacitus, including 1 uncertain which we ignore.

We optimise for RECALL
In one detection task**, TRACER identified 40 reuse candidates:

- it identified the 4 reported reuses (=true positives)!
- it returned 36 other parallels (=false positives), of which 2 “new” discoveries.

*Corpus Scriptorum Ecclesiasticorum Latinorum edition (1882) lists 4 reuses; the Patrologia Latina (1846) lists 5.
**Detection parameters: moving window of 15 words; 0.8 feat. density; syn. replacement; 0.5 (50%) sim. threshold.
Example of a Gold Standard match

**Tacitus.** *Sic conquisitum collectumque vulgus, postquam vastis locis relictum sit, ceteris per lacrimas torpentibus, Moysen unum exulum monuisse ne quam deorum hominum ve opem expectarent utrisque deserti, sed sibimet duce caelesti crederent, primo cuius auxilio praesentis miserias pepulissent.* (5.3)

**Orosius.** *sic conquisitum collectumque uulgus postquam vastis locis relictum sit, ceteris per lacrimas torpentibus Moysen, unum exulum, monuisse, ne quam deorum hominum ue opem exspectarent sed sibimet duci caelesti crederent, primo cuius auxilio praesentes miserias pepulissent.* (1.10)
New find: Syntactic text reuse; possibly intentional.

**Tacitus.** *Terra finesque qua ad Orientem vergunt Arabia terminantur, a meridie Aegyptus obiacet, ab occasu Phoenices et mare, septentrionem et latere Syriae longe prospectant.* (5.6)

**Orosius.** *Insula Cypros ab oriente mari Syrio, quem Issicum sinum uocant, ab occidente mari Pamphylico, a septentrione Aulone Cilicio, a meridie Syriae et Phoenices pelago cingitur.* (1.2)
New find: Semantic text reuse; possibly unintentional.

**Tacitus.** *Haud facile quis* [uni adsignaverit] *culpam quae omnium fuit.* (5.1)

**Orosius.** *quae tempora non* [uni tantum urbi adtributa sed orbi uniuerso constat esse communia.* (3.78)
Example of a false positive, to be added to the Gold Standard

TRACER output: 1200457 1101980 2 0.5
Example of a false positive, to be added to the Gold Standard

TRACER output: 1200457 1101980 2 0.5

**Tacitus** (1200457). *imus ad bellum*.
Reduced by TRACER to: *eo ad bellum*.
CORRECT
RESULTS: OROSIUS’ REUSE OF TACITUS’ HISTORIES

Example of a false positive, to be added to the Gold Standard

TRACER output: 1200457  1101980  2  0.5

Tacitus (1200457). *imus ad bellum.*
Reduced by TRACER to: *eo ad bellum*
CORRECT

Orosius (1101980). *itaque bellum nocte commissum est.*
Reduced by TRACER to: *eo bellum nox committo sum*
INCORRECT
CONCLUSIONS AND NEXT STEPS
Despite all previous scholarship, our method makes it possible to detect new parallels for further study.

- Our methodology works but there is room for more automation.
- Our research limited by the linguistic resources at our disposal.
**NEXT STEPS**

**Project**

- Consolidate our methodology as we go through the entire corpus.

**Research**

Work towards an **automatic taxonomy of reuse styles**:

- Add *verba dicendi* (verbs introducing reported statements), idioms, proverbs, law maxims, figures of speech, etc. as embedded TRACER data resources to help filter results for evaluation;

- Incorporate *WordFormation Latin* (Litta, Passarotti) data (derivational dictionary) as additional text reuse detection dimension (word class).
参观我们
🌐 http://www.etrap.eu
✉️ contact@etrap.eu

Stealing from one is plagiarism, stealing from many is research
(Wilson Mitzner, 1876-1933)
The authors wish to thank Marco Passarotti and Paolo Ruffolo (CIRCSE) for their continuous support with LemLat3.
The theme this presentation is based on is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Changes to the Metropolis theme are the work of eTRAP.