ADVANCING MACHINE-ASSISTED INTERTEXTUAL RESEARCH ON HISTORICAL DATA

eTRAP: ELECTRONIC TEXT REUSE ACQUISITION PROJECT

Greta Franzini
Institute of Computer Science, University of Göttingen, Germany
WHO ARE WE?
Education

- Humanities & Further Maths Diploma (IT)
- Classics BA Honours (UK)
- Digital Humanities MA (UK)
- Part-time PhD student (UCLDH, UK): Digital Editions
  - Catalogue of Digital Editions (now a collaboration with ACDH)
  - Digital edition of an ancient Latin manuscript

Work

- Full-time post-doctoral researcher for eTRAP Early Career Research Group (DE): Automatic Text Reuse Detection and Analysis
Electronic Text Reuse Acquisition Project (eTRAP)

Interdisciplinary Early Career Research Group funded by the German Ministry of Education & Research (BMBF).

Budget: €1.6M.
Team: 4 core staff; 5-9 research & student assistants; Bachelor, Masters and PhD thesis students.

- Interdisciplinary: Classics, Computer Science, German Philology, Mathematics, Philosophy, Cognitive Literature.
- International: Currently from eight nationalities.
WHAT IS TEXT REUSE?
Text reuse = spoken and written repetition of text across time and space.

Figure 1: Text reuse styles.
“[...] a text is [...] a multidimensional space in which a variety of writings, none of them original, blend and clash. The text is a tissue of quotations drawn from the innumerable centres of culture... the writer can only imitate a gesture that is always anterior, never original. His only power is to mix writings [...].” (Barthes, 1977, pp. 146-47)

“[...] any text is constructed as a mosaic of quotations [...].” (Kristeva, 1980, p.66)
Question:
Why is text reuse detection relevant for Humanities and Computer Science?

- **Humanities:**
  - Lines of transmission and textual criticism.
  - Transmissions of ideas & thoughts under different circumstances and conditions.

- **Computer Science:**
  - Text decontamination for stylometry and authorship attribution, dating of texts.
  - Text Mining, Corpus Linguistics.
CURRENT CHALLENGES

Text reuse challenges:

- Detecting text reuse at scale (Big Data: information overload vs. information poverty);
- Detecting text reuse across languages;
- Detecting looser forms of text reuse, e.g. allusion;
- Diversity of historical texts: language evolution, copy errors, etc.
“The fundamental methodological fact that historical linguists have to face is that they have no control over their data... The great art of the historical linguist is to make the best of this bad data - ‘bad’ in the sense that it may be fragmentary, corrupted or many times removed from the actual productions of native speakers.” (Labov, 1972, p. 100)
OUR RESEARCH
OVERVIEW OF OUR PROJECTS: HISTORICAL DATA

TRACER:
Reuse detection algorithms

- Paulus Orosius: Reuse at scale & across periods
- Greek/Latin Patristics: Reuse diversity
- Jane Austen: Cognitive/structural reuse
- TrAiN: Reuse vs. OCR/HTR noise
- Brothers Grimm: Cross-lingual reuse
TRACER: suite of **700 algorithms** developed by Marco Büchler. Command line environment with no GUI.

**Figure 2:** Detection task in six steps. More than 1M permutations of implementations of different levels are possible.

TRACER tested on: Ancient Greek, Arabic, Coptic, English, German, Hebrew, Latin, Tibetan.
Webpage: http://www.etrap.eu/research/tracer
Repository: http://vcs.etrap.eu/tracer-framework/tracer.git
Upcoming tutorials:

• **AIUCD 2017** (Jan 2017): pre-conference workshop with DiXiT, Rome, Italy.
• **DATECH 2017** (May 2017): pre-conference workshop, Göttingen, Germany.
• Three more tutorials in 2017 pending confirmation.
The collection and automatic detection of folktale motifs as text reuse units at scale and across languages.
Motif: "1. A minimal thematic unit" (Prince, 2003, p. 55), a measurable primitive.

Measurable primitives from an interdisciplinary standpoint:

- **Literature**: tracing MOTIFS
- **Cultural Studies**: tracing MEMES
- **Linguistics**: tracing PATTERNS
- **Computer Science**: tracing FEATURES
- **Forensics**: tracing MINUTIAE
RQ: How to computationally detect a motif despite its variants?

For example:

- **DE** [Grimm]: *Schneewittchen und die sieben Zwerge*
- **EN** [Briggs]: *Snow White and the three robbers*
- **IT** [Calvino]: *Bella Venezia e i dodici ladroni*
- **SQ** [von Hahn]: *Schneewittchen und die vierzig Drachen*
- **RU** [Pushkin]: Сказка о мертвой царевне и о семи богатырях
- ...

**A**: We strike a balance between precision and recall. That is, finding the balance between a specific motif (Aarne-Thompson-Uther index) and its ontological root (Propp’s typological unity).

**How?** Adapt a Named Entity Recognition tool based on neural networks by replacing default categories (place name, person name, etc.) with the motifs and the top-level concepts.
Tasks: Verify presence of motif in different collections and record its “base form” as text reuse training data.

Figure 3: Microsoft Excel matrix of motifs. Left column lists AT motifs in Snow White (AT 709); top row lists languages and collections covered.

Figure 4: Grimm motifs reduced to keywords.
JANE AUSTEN AND TEXT SIMPLIFICATION
Definition:

Graded readers are “simplified books written at varying levels of difficulty for second language learners”, which “cover a huge range of genres ranging from adaptation of classic works of literature to original stories, to factual materials such as biographies, reports and so on” [Waring 2012].
To computationally analyse the process Y and classifying the changes:

- Do the changes follow strict rules?
- Do they form patterns?
- Can they be computationally reproduced?

Categories of changes:

- Cognitive
- Structural
- Cognitive and structural
An example of a structural text simplification > many-to-one.
The **Dotplot view** of original novel against the graded reader on a sentence-wise segmentation uncovers which passages were taken over in the graded reader and which not:
ANCIENT GREEK AND LATIN PATRISTIC TEXT REUSE
Inspired by Shannon's noisy-channel & Kolmogorov Complexity we study Greek and Latin text reuse to understand how text is transformed.

- We identify operations that characterise word changes.
- We show how linguistic resources can help detect non-literal reuse.
- We complement the automated approach with a manual analysis.
ANCIENT GREEK AND LATIN DATA-SETS

“Salvation for the Rich”
Clement of Alexandria (Greek)
Christian theologian, 2nd cent.

Extracts from 12 works & 2 collections
Bernard of Clairvaux (Latin)
French abbot, 12th cent.

• Reuse extracted by Biblindex team (Mellerin, 2014, 2016)
• 199 aligned reuses
• Pointing to 15 Bible books

• Reuse extracted by Biblindex team (Mellerin, 2014, 2016)
• 162 aligned reuses
• Pointing to 31 Bible books
### Table 1: Operation list for the automated approach

<table>
<thead>
<tr>
<th>operation</th>
<th>description</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOP(reuse_word, orig_word)</td>
<td>Original and reuse word are equal.</td>
<td>NOP(maledictus, maledictus)</td>
</tr>
<tr>
<td>upper(reuse_word, orig_word)</td>
<td>Word is lowercase in reuse and uppercase in original.</td>
<td>upper(kai.Kai) - in Greek</td>
</tr>
<tr>
<td>lower(reuse_word, orig_word)</td>
<td>Word is uppercase in reuse and lowercase in original.</td>
<td>lower(Gloriam, gloriam)</td>
</tr>
<tr>
<td>lem(reuse_word, orig_word)</td>
<td>Lemmatization leads to equality of reuse and original.</td>
<td>lem(penetrat, penetrabit)</td>
</tr>
<tr>
<td>repl_syn(reuse_word, orig_word)</td>
<td>Reuse word replaced with a synonym to match original word.</td>
<td>repl_syn(magnificavit, glorificavit)</td>
</tr>
<tr>
<td>repl_hyper(reuse_word, orig_word)</td>
<td>Word in bible verse is a hyperonym of the reused word.</td>
<td>hyper(cupit, habens)</td>
</tr>
<tr>
<td>repl_hypo(reuse_word, orig_word)</td>
<td>Word in bible verse is a hyponym of the reused word.</td>
<td>hypo(dederit, tolet)</td>
</tr>
<tr>
<td>repl_co-hypo(reuse_word, orig_word)</td>
<td>Reused word and original have the same hyperonym.</td>
<td>repl_co-hypo(magnificavit, fecit)</td>
</tr>
<tr>
<td>NOPmorph(reuse_tags, orig_tags)</td>
<td>Case or PoS did not change between reused and original word.</td>
<td>NOPmorph(na, na)</td>
</tr>
<tr>
<td>repl_pos(reuse_tag, orig_tag)</td>
<td>Reuse and original contain the same cognate, but PoS changed.</td>
<td>repl_pos(n, a)</td>
</tr>
<tr>
<td>repl_case(reuse_tag, orig_tag)</td>
<td>Reuse and original have the same cognate, but the case changed.</td>
<td>repl_case(g, d) - cases genitive, dative</td>
</tr>
<tr>
<td>lemma_missing(reuse_word, orig_word)</td>
<td>Lemma unknown for reuse or original word</td>
<td>lemma_missing(tentari, inlectus)</td>
</tr>
<tr>
<td>no_rel_found(reuse_wword, orig_word)</td>
<td>Relation for reuse or original word not found in AGWN</td>
<td>no_rel_found(gloria, arguitur)</td>
</tr>
</tbody>
</table>
Latin Text Reuse Detection at Scale
1. To test TRACER’s capabilities under **stressful conditions**:  
   • Large corpus (millions of words);  
   • Different types of Latin;  
   • **Different reuse styles** requiring different window sizes;  
   • Computational power and resources needed.

2. To work towards the establishment of a **Gold Standard** for Latin lemmatisation.
Challenges:
- Scale
- Reuse styles
- Messy reuse
- Latin(s)

- Classical Latin
- Late/Ecclesiastical
- Future additions

Paulus Orosius

- 1.3M words
RESULTS & NEXT STEPS

Method

• Multiple experiments with different window sizes to address the reuse diversity;
• Check computed results against identified reuse in commentaries.

Derivative research

• Optimise detection by parallelising TRACER computation;
• Improvement of TreeTagger in collaboration with its developers.
Tracing Authorship in Noise
“It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson, 2003). Data preparation is not just a first step, but must be repeated many times over the course of analysis as new problems come to light or new data is collected.” (Wickham, 2014, p. 1)
Duration: 6 months
Budget: 20,000€
Funder: University of Göttingen, Campuslabor Digitalisierung
Final expert workshop (March 2017): text reuse meets stylometry

RQ: When does OCR/HTR noise begin to interfere with automatic text reuse and style detection?

Case study: Correspondence of Brothers Grimm
CONCEPT: GRIMM CORRESPONDENCE

Digitised letters written by Jacob & Wilhelm Grimm from childhood to death.

HTR: Transkribus

OCR: Tesseract

OCR: Ocropy

H. Rölleke's 2001 print edition of the letters.

TRACER

Threshold:

Precision:

Noise:

Text reuse

Stylometry
SUMMARY
CONCLUSION

- **Complexity of text reuse detection**: big data, incomplete historical data, noisy digitised data;
- **Our holistic approach** to this complexity for a comprehensive understanding of text reuse;
- **Advanced our understanding** of the process of text reuse: what are our primitives and how we can measure them (towards the improvement of TRACER).
Team
Marco Büchler, Greta Franzini, Emily Franzini and Maria Moritz.

Visit us
🌐 http://www.etrap.eu
✉ contact@etrap.eu

*Stealing from one is plagiarism, stealing from many is research*
 *(Wilson Mitzner, 1876-1933)*
• 1. Grimm (1812-1857) *Kinder- und Hausmärchen*.


REFERENCES
DATECH 2017: A DH CONFERENCE ON DIGITAL TRANSFORMATION

- **Submission Deadline:** 7 January 2017
- **Topics:**
  - Improved OCR and special OCR techniques for historical documents.
  - Innovative views and tools for the exploitation of digital content by both experts and non-expert communities in the humanities.
  - Advanced tools for a higher productivity and quality in the creation of useful digital content.
  - Improved treatment of historical languages (diachronic language development) and multilingualism.
  - New mining techniques on historical text collections (addressing e.g., historical text reuse or person and event detection).
The theme this presentation is based on is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Changes to the theme are the work of eTRAP.