BROTHERS GRIMM, JANE AUSTEN AND PAULUS OROSIUS HAVE ONE THING IN COMMON:
THE eTRAP RESEARCH TEAM AND ITS DH PROJECTS

Emily Franzini and Greta Franzini
February 23, 2017
1. Who are we?

2. What is text reuse?

3. Our research
INTERRUPT US AT ANY TIME!
WHO ARE WE?
**Electronic Text Reuse Acquisition Project (eTRAP)**

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

**Budget:** €1.6M.

**Duration:** March 2015 - February 2019. Research since October 2015.

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

- **Interdisciplinary:** Classics, Computer Science, German Literature, Mathematics, Philosophy, Cognitive Psychology and Literature Studies.
- **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.
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.
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.
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 is language-independent.
Tested on: Ancient Greek, Arabic, Coptic, English, German, Hebrew, Latin, Tibetan.
TEXT REUSE INVESTIGATION INTO THE FIRST
CHRISTIAN HISTORY OF ROME
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.
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, diachronic corpus affect automatic text reuse detection and its performance?
CHALLENGES: DIACHRONIC CORPUS

Orosius’ Histories

~1.3M words
Classical Latin
Late/Ecclesiastical
Future additions
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]
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: OROSIUS’ REUSE OF TACITUS’ HISTORIAE

- Reuses documented in primary sources (precision): 15 (= 12+3?)
- Reuses TRACER can attempt to match: 5 (10 are of lost text)
- Reuses identified by TRACER (recall)*: 40
  - verbatim;
  - near-verbatim: “true” and “false” (spelling conventions);

*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.
JANE AUSTEN AND TEXT SIMPLIFICATION
PRIDE
AND
PREJUDICE:
A NOVEL.
IN THREE VOLUMES.

BY THE
AUTHOR OF "SENSE AND SENSIBILITY."

VOL. I.

London:
PRINTED FOR T. EGERTON,
MILITARY LIBRARY, WHITEHALL.
1813.
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
Structural changes:

- Elizabeth is **exceedingly handsome**.
- Elizabeth is **very beautiful**.

Cognitive changes:

- ... *Soon after this event, Elizabeth received a visit...*

Structural & cognitive changes:

- Elizabeth is **exceedingly beautiful**.
Readability tests aim to classify texts by their degree of complexity and understandability. Measured primitives are sentence length and difficulty of the words.

Two tests, the ARI score and the Dale-Chall-Index have been selected:

The ARI score is based on the word length and the sentence length:

$$R_{ARI} = 4.71 \left( \frac{\text{characters}}{\text{words}} \right) + 0.5 \left( \frac{\text{words}}{\text{sentences}} \right) - 21.43$$  \hspace{1cm} (1)

The Dale-Chall-Index is based on the word frequency (3000 most frequent words) and the sentence length:

$$R_{DCI} = 0.1579 \left( \frac{\text{difficult words}}{\text{words}} \times 100 \right) + 0.0496 \left( \frac{\text{words}}{\text{sentences}} \right)$$  \hspace{1cm} (2)
Readability test result matrix:

<table>
<thead>
<tr>
<th></th>
<th>ARI</th>
<th>Dale-Chall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Novel</td>
<td>14-15 year olds</td>
<td>14-16 year olds</td>
</tr>
<tr>
<td>Graded Reader</td>
<td>11-12 year olds</td>
<td>11-13 year olds</td>
</tr>
</tbody>
</table>
COMPARISON OF SENTENCE LENGTH

Sentence length distribution

- Original text
- Graded reader
COMPARISON OF WORD LENGTH

Word length distribution

Original text
Graded reader

Length of word
Probability
An example of a structural text simplification > many-to-one.
chapter 1 it be a truth universally understand that a single man in possession of a good fortune must need a wife

so when a wealthy gentleman arrive in a neighbourhood it be clear that he must soon become the property of someone daughter

my dear Mr. Bennet say he wise to he one day have you hear that someone have rent the house at Netherfield Park at last

Mr. Bennet reply that he have not

yes she continue Mrs. Long have just be here and she tell I all about it

Mr. Bennet do not answer

do you not want to know who have take it
cry be wise impatiently

you want to tell I and I have no objection to hear it

well Mrs. Long say that Netherfield have be take by a young man of large fortune from the north of England

what be he name
Bingley
be he marry or single

oh

single my dear

a single man of large fortune four or five thousand pound a year

what a fine thing for we girl

ON

however little known the feeling or view of such a man may be on he first enter a neighbourhood this truth be so well fix in the mind of the surround family that he be consider the rightful property of some one or other of they daughter

my dear Mr. Bennet say he lady to he one day have you hear that Netherfield Park be let at last

Mr. Bennet reply that he have not

but it be return she for Mrs. Long have just be here and she tell I all about it

Mr. Bennet make no answer

do you not want to know who have take it
cry be wise impatiently

want to tell I and I have no objection to hear it

this be invitation enough

why my dear you must know Mrs. Long say that Netherfield be take by a young man of large fortune from the north of England that he come down on Monday in a chaise and four to see the place and be so much delighted with it that he agree with Mr. Morris immediately that he be to take possession before Michaelmas and some of be servant be to be in the house by the end of next week

what be he name
Bingley
be he marry or single

oh
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:
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

Research Question:
When does Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) noise begin to interfere with automatic text reuse and authorship detection?

Case study: Correspondence of Brothers Grimm
Digitised letters written by Jacob & Wilhelm Grimm from childhood to death.

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

HTR Version
Transkribus

OCR Version
Ocropy & Tesseract

Threshold

Precision

100
80
60
40
20
0

10% 20% 30% 40% 50%
Noise

Text reuse
Stylometry
DIGITAL Breadcrumbs of
BROTHERS GRIMM
The collection and automatic detection of folktale motifs as text reuse units across languages and traditions.
RQ: How to computationally detect a motif despite its variants?

For example:

- DE [Grimm]\(^1\): *Schneewittchen und die sieben Zwerge*
- EN [Briggs]\(^2\): *Snow White and the three robbers*
- IT [Calvino]\(^3\): *Bella Venezia e i dodici ladroni*
- SQ [von Hahn]\(^4\): *Schneewittchen und die vierzig Drachen*
- RU [Pushkin]\(^5\): 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).
**DATA COLLECTION AND CURATION**

**Tasks:** Verify presence of motifs in different collections and record their “base form” as text reuse **training data**.

![Image](https://example.com/image)

**Figure 4:** Microsoft Excel matrix of motifs. Left column lists AT motifs in *Snow White* (AT 709); top row lists languages and collections covered.

![Table](https://example.com/table)

**Figure 5:** Grimm motifs reduced to keywords.
Thompson Motif Index (TMI) ontology (OWL/RDF), by Antónia Koštová, Thierry Declerck and Tyler Klement (Declerck et al., 2016).

Figure 6: Representation of a motif in the TMI ontology. Image reproduced with permission of Thierry Declerck.
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
REFERENCES


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.