

ADVANCING MACHINE-ASSISTED INTERTEXTUAL RESEARCH ON HISTORICAL DATA

eTRAP: ELECTRONIC TEXT REUSE ACQUISITION PROJECT

Greta Franzini

Institute of Computer Science, University of Göttingen, Germany



TABLE OF CONTENTS

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.

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 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.

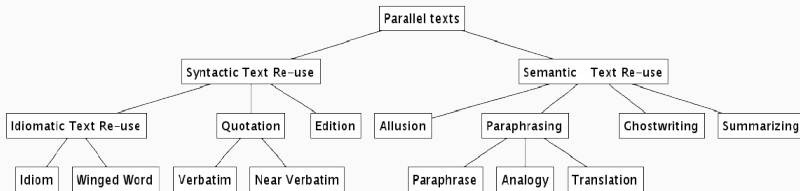


Figure 1: Text reuse styles.

“[...] 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.

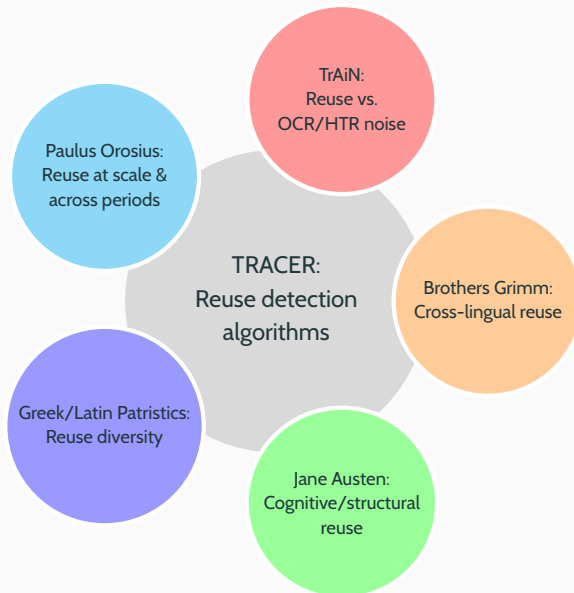
Text reuse challenges:

- 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

TEXT REUSE DETECTION MACHINE

TRACER: OVERVIEW

TRACER: suite of **700 algorithms** developed by Marco Büchler.
Command line environment with no GUI.

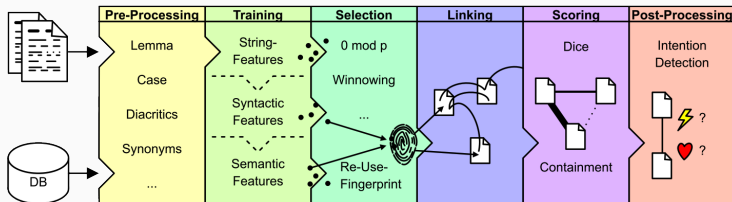


Figure 2: Detection task in six steps. More than 1M permutations of implementations of different levels are possible.

TRACER tested on: Ancient Greek, Arabic, Coptic, English, German, Hebrew, Latin, Tibetan.

Webpage: <http://www.etrp.eu/research/tracer>

Repository: <http://vcs.etrp.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.



DIGITAL BREADCRUMBS *of*
BROTHERS GRIMM

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**



EXAMPLE CASE STUDY: SNOW WHITE

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).

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.

DATA COLLECTION AND CURATION

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

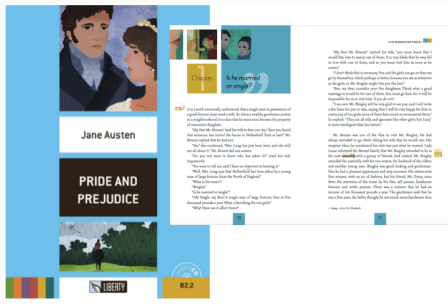
ISO Language Codes https://www.loc.gov/standards/iso639-2/php/code_list.php		GER						RUS	ITA	GLA	ARM	ENG		ARA					
Aarne-Thompson: 709		Grimm_1819 VIAF: 187449723	Grimm_1837 VIAF: 187449723	Grimm_1840 VIAF: 187449723	Grimm_1843 VIAF: 187449723	Grimm_1850 VIAF: 187449723	Grimm_1857 VIAF: 187449723	Pushkin_1833 VIAF: 312344013	Tsvetaeva_1911 VIAF: 185098476	Calvino_1956 VIAF: 181208131	Jacobs_1892 VIAF: 315397813	Bruford_1994 VIAF: 12471835	Hoogasian- Villa_1966 VIAF: 186329063	Campbell_1958 VIAF: 25969242	Taylor_1823 VIAF: 59071527	Briggs_1970 VIAF: 46803237	El-Shamy_1989 VIAF: 276573319	El Koudia_2003 VIAF: 5206198	Jason_1977 VIAF: 9970253
D1300-D1379. Magic objects effect changes in persons																			
D1364. Object causes magic sleep		x	x	x	x	x	x	x	null	x	x	x	x	x	x	x	x	x	x
D1364.4. Fruit causes magic sleep		x	x	x	x	x	x	x	null	null	null	null	null	x	x	x	null	null	null
D1364.4.1. Apple causes magic sleep		x	x	x	x	x	x	x	null	null	null	null	null	x	x	x	null	null	null
D1364.9. Comb causes magic sleep		x	x	x	x	x	x	null	null	null	null	null	null	x	x	null	null	null	null
D1364.13. Cloth causes magic sleep		x	x	x	x	x	x	null	null	null	null	null	null	null	x	null	null	null	null
D1364.13.1. Lace causes magic sleep		x	x	x	x	x	x	null	null	null	null	null	null	null	x	null	null	null	null

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

Q400-Q599. Kinds of punishment		
Q411. Death as punishment		zu todt tanzen
Q414. Punishment: burning alive		glühende Pantoffeln, zu todt tanzen
Q414.4. Punishment: dancing to death in red-hot shoes		eiserne Pantoffeln, Feuer, glühend, anziehen, tanzen, Füße jämmerlich verbrannt, nicht aufhören, zu todt tanzen

Figure 4: Grimm motifs reduced to keywords.

JANE AUSTEN AND TEXT SIMPLIFICATION

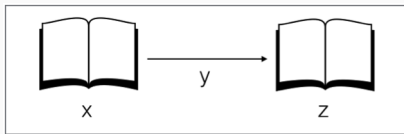


Definition:

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

SIMPLIFICATION & SENTENCE LENGTH

An example of a structural text simplification > many-to-one.

Text Re-use Alignment Visualization

X

GR

chapter 1 it be a truth universally understand that a single man in possession of a good fortune must need a wife

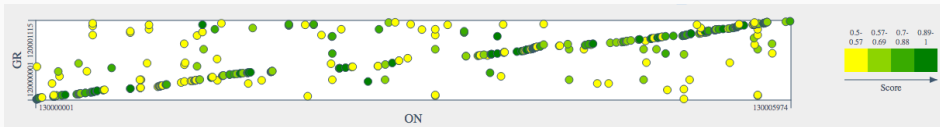
ON

chapter 1 it be a truth universally acknowledge that a single man in possession of a good fortune must be in want of a wife



MACRO SCALE: VISUALISATION OF THE SELECTION PROCESS

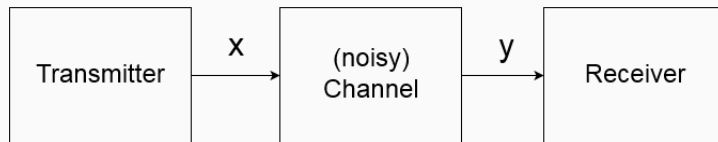
The **Dotplot view** of original novel against the graded reader on a sentence-wise segmentation uncovers which passages were taken over in the graded reader and which not:



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.



“Salvation for the Rich”

Clement of Alexandria (Greek)

Christian theologian, 2nd cent.

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

Extracts from 12 works & 2 collections

Bernard of Clairvaux (Latin)

French abbot, 12th cent.

- Reuse extracted by Biblindex team (Mellerin, 2014, 2016)
- **162 aligned reuses**
- Pointing to 31 Bible books

Table 1: Operation list for the automated approach

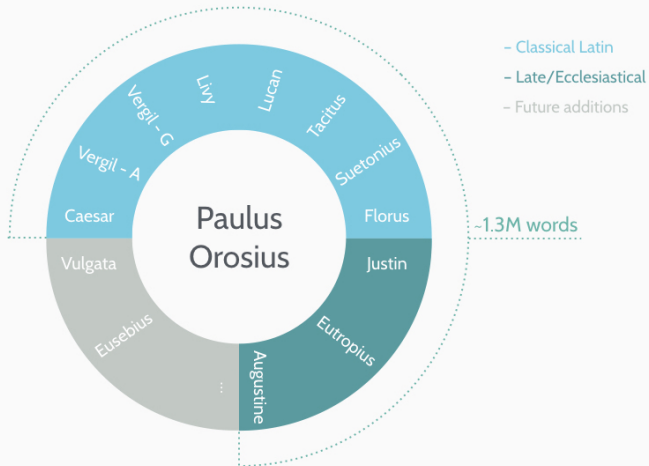
operation	description	example
<i>NOP(reuse_word, orig_word)</i>	Original and reuse word are equal.	<i>NOP(maledictus,maledictus)</i>
<i>upper(reuse_word, orig_word)</i>	Word is lowercase in reuse and uppercase in original.	<i>upper(kai,Kai)</i> - in Greek
<i>lower(reuse_word, orig_word)</i>	Word is uppercase in reuse and lowercase in original.	<i>lower(Gloriam,gloriam)</i>
<i>lem(reuse_word, orig_word)</i>	Lemmatization leads to equality of reuse and original.	<i>lem(penetrat,penetrabit)</i>
<i>repl_syn(reuse_word, orig_word)</i>	Reuse word replaced with a synonym to match original word.	<i>repl_syn(magnificavit,glorificavit)</i>
<i>repl_hyper(reuse_word, orig_word)</i>	Word in bible verse is a hyperonym of the reused word.	<i>hyper(cupit,habens)</i>
<i>repl_hypo(reuse_word, orig_word)</i>	Word in bible verse is a hyponym of the reused word.	<i>hypo(dederit,tollet)</i>
<i>repl_co-hypo(reuse_word, orig_word)</i>	Reused word and original have the same hyperonym.	<i>repl_co-hypo(magnificavit,fecit)</i>
<i>NOPmorph(reuse_tags, orig_tags)</i>	Case or PoS did not change between reused and original word.	<i>NOPmorph(na,na)</i>
<i>repl_pos(reuse_tag, orig_tag)</i>	Reuse and original contain the same cognate, but PoS changed.	<i>repl_pos(n,a)</i>
<i>repl_case(reuse_tag, orig_tag)</i>	Reuse and original have the same cognate, but the case changed	<i>repl_case(g,d)</i> - cases genitive, dative
<i>lemma_missing(reuse_word, orig_word)</i>	Lemma unknown for reuse or original word	<i>lemma_missing(tentari, inlectus)</i>
<i>no_rel_found(reuse_word, orig_word)</i>	Relation for reuse or original word not found in AGWN	<i>no_rel_found(gloria,arguitur)</i>

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.

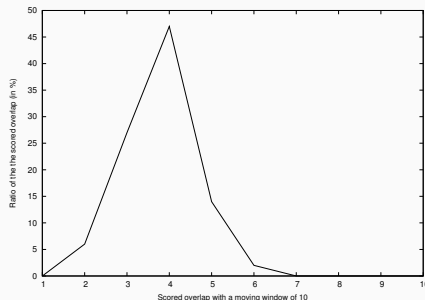
Challenges:

- Scale
- Reuse styles
- Messy reuse
- Latin(s)



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)

OVERVIEW & RESEARCH OBJECTIVE

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.



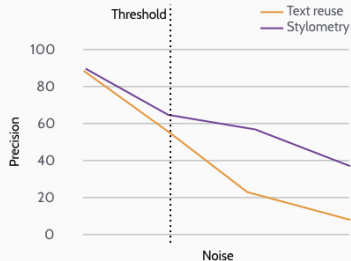
H. Rölleke's 2001 print edition of the letters.

HTR: Transkribus

TRACER

OCR: Tesseract

OCR: Ocropy



SUMMARY

- **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.etrp.eu>



contact@etrp.eu

Stealing from one is plagiarism, stealing from many is research
(Wilson Mitzner, 1876-1933)



SPONSORED BY THE

Federal Ministry
of Education
and Research

- 1. Grimm (1812-1857) *Kinder- und Hausmärchen*.
- 2. Briggs, K. M. (1970) *A Dictionary of British Folk-Tales in the English Language: Part A: Folk Narratives*. London: Routledge & Kegan Paul.
- 3. Calvino, I. (1956) *Fiabe Italiane*. Mondadori.
- 4. Hahn, J. G. von (1864) *Griechische und Albanesische Märchen*, Zweiter Theil. Leipzig: Engelmann, pp. 137.
- 5. Пушкин, Александр Сергеевич (1799-1837). Сказка о мертвой царевне и о семи богатырях. Available at:
<http://rvb.ru/pushkin/01text/03fables/01fables/0800.htm> (Accessed: 27 June 2016).

REFERENCES

- Barthes, R. (1977) *Image-Music-Text*, Stephen Heath (trans.). London: Fontana.
- Dasu T., Johnson T. (2003) *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons.
- Kristeva, J. (1980) *Desire in Language: A Semiotic Approach to Literature and Art*. Thomas Gora, Alice Jardine, Leon S. Roudiez (trans.), Leon S. Roudiez (ed.). New York: Columbia University Press.
- Labov, W. (1972) 'Some principles of linguistic methodology', *Language in Society*, 1(1), pp. 97-120 [Online]. At: <http://www.jstor.org/stable/4166672>
- Waring, R. (2012) *Writing graded readers*. At: <http://www.er-central.com/authors/writing-a-graded-reader/writing-graded-readers-rob-waring/>
- Wickham, H. (2014) 'Tidy Data', *Journal of Statistical Software*, 59(10). At: <https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf>

- **Link:** <http://ddays.digitisation.eu/datech-2017/>
- **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.

