ELECTRONIC TEXT REUSE ACQUISITION PROJECT

Introduction & Motivation

Marco Büchler

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

WHO AM I?

WHO AM I?

- 2001-2002: Head of Quality Assurance department in a software company;
- 2006: Diploma in Computer Science on big scale co-occurrence analysis;
- 2007: Consultant for several SMEs in IT sector;
- 2008: Technical project management of the eAQUA project;
- 2011: PI and project manager of the eTRACES project;
- 2013: PhD in Digital Humanities on Text Reuse;
- 2014: Head of Early Career Research Group eTRAP at the University of Göttingen.

ABOUT ETRAP

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 DO YOU ASSOCIATE WITH TEXT REUSE?

TEXT REUSE

Text Reuse:

- spoken and written repetition of text across time and space.

For example:

- citations, allusions, translations.

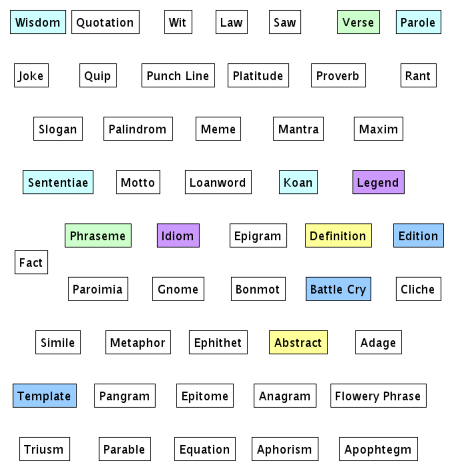
Detection methods are needed to support scholarly work.

- E.g. they help to ensure clean libraries or identify fragmentary authors.

Text is often modified during the reuse process.

EXPECTATIONS OF A HUMANIST: OVERSIMPLIFICATION

DIVERSITY (REUSE TYPES)



Wisdom Quotation Wit Law Saw Verse Parole

Joke Quip Punch Line Platitude Proverb Rant

Slogan Palindrom Meme Mantra Maxim

Sententiae Motto Loanword Koan Legend

Phraseme Idiom Epigram Definition Edition

Fact

Paroimia Gnome Bonmot Battle Cry Cliche

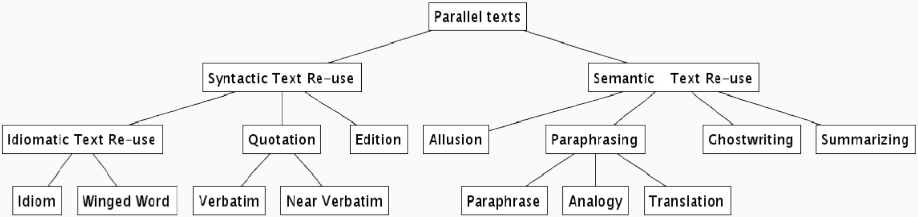
Simile Metaphor Ephithet Abstract Adage

Template Pangram Epitome Anagram Flowery Phrase

Triusm Parable Equation Aphorism Apophtegm

- **Stability** (yellow)
- **Purpose** (green)
- **Size of text reuse** (blue)
- **Classification** (light blue)
- **Degree of distribution** (purple)
- Written and oral transmission

DIVERSITY (REUSE STYLES)



KEY PROBLEM

Question:

The distribution of **Reuse Types** and **Reuse Styles** is often unknown - which model(s) should be chosen?

MOTIVATION

“REUSE FROM SAME SOURCE”: COMMONALITIES & DIFFERENCES

WITTGENSTEIN'S "FAMILY RESEMBLANCE"

Family resemblance is an equivalence relation that clusters common objects of similar and not identical characteristics together.

Family resemblance is hierarchical such as in the examples before "Greta", "Franzinis", "Human", ''creature".

ETRAP'S OBJECTIVE

Title: eTRAP - electronic Text Reuse Acquisition Project

Premise: Language is a changing system. Compared to biometry the volatility is much higher.

- Research on the characteristics
    - What are good characteristics?
    - Which characteristics are stable and which are volatile and therefore not helpful in the detection process?
- Research on the reuse process
    - Begins with: Why do we quote what we quote?
    - Passes by: If changes in the reuse process happen, why do they happen and what is the model behind (if one exists)?
    - Ends with: Understanding paraphrases and allusions

COMPARISON OF LUKE & MARK

TRACER: OVERVIEW

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

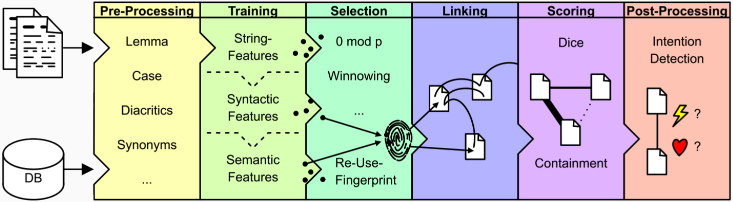
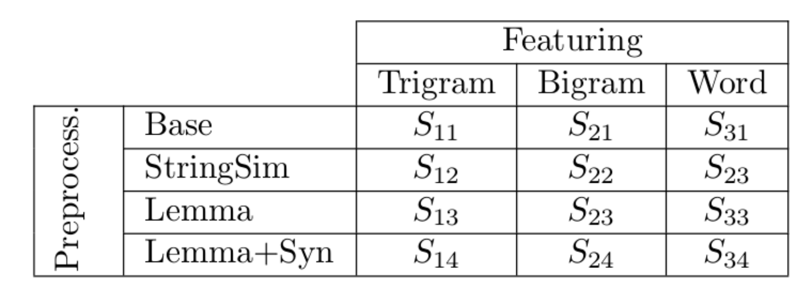


Figure 1: Detection task in six steps. More than 1M permutations of implementations of different levels are possible.

TRACER is language-independent.
Tested on: Ancient Greek, Arabic, Coptic, English, German, Hebrew, Latin, Tibetan.

TEXT REUSE IN ENGLISH BIBLE VERSIONS: SETUP

Segmentation: disjoint and verse-wise segmentation.



		Featuring		
		Trigram	Bigram	Word
Preprocess.	Base	S_{11}	S_{21}	S_{31}
	StringSim	S_{12}	S_{22}	S_{23}
	Lemma	S_{13}	S_{23}	S_{33}
	Lemma+Syn	S_{14}	S_{24}	S_{34}

Selection: max pruning with a Feature Density of 0.8;

Linking: Inter- Digital Library Linking (different Bible editions);

Scoring: *Broder's Resemblance* with a threshold of 0.6;

Post-processing: not used.

DATA SCIENCE & PRECISION AND RECALL

EXPECTATIONS OF A HUMANIST: OVERSIMPLIFICATION

TRACER: DISSEMINATION

Webpage: `http://www.etrap.eu/research/tracer`
Repository: `http://vcs.etrap.eu/tracer-framework/tracer.git`
Upcoming tutorials:

- **DATeCH 2017** (May 2017): pre-conference workshop, Göttingen, Germany.
- Three more tutorials in 2017 pending confirmation.

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

LICENCE