Overview
An early career research group on historical text reuse – Methodologies & research fields

Marco Büchler, Emily Franzini, Greta Franzini, Maria Moritz
DH Estonia,
Tartu
20 October
2015
Text Reuse

Intratextuality
Internal relations within a text or an author

Intertextuality
External relations with other texts

Text Reuse
Spoken and written repetition of text/content
Motivation

Question

Why is it so relevant for Humanities and Computer Science?

Humanities

General transmissions of ideas under different conditions
E.g. lines of transmission & textual criticism

Computer Science

Text decontamination for authorship attribution
Text mining, corpus linguistics
Our Objectives

How
Were paraphrases, allusions or translations reused by authors over time?

Why
Was part of the text reused? What influenced the text?

Aim
Develop a methodology to „measure“ historical text reuse in spite of its diversity

Through
Big Data, texts in Ancient Greek, German, English, Italian, Latin

"IN THE BEGINNING GOD CREATED THE HEAVENS AND THE EARTH"
Genesis, Chap. 1
Team

Marco Büchler

Big Data Processing

Stability and Volatility of the Minutiae

Emily Franzini

Translation

Manuscripts

Greta Franzini

Maria Moritz

Reuse Diversity

Student

Student

Student

Student

Student

Student
Works on determining relevant features of text reuse and which of them can be calculated by machines?

What can’t be calculated?

How can paradigmatic relations support the text reuse analysis (e.g. query expansion on search engines)
Greta Franzini, Humanist

Can the study of text reuse…

› …tell us something about the **accuracy** and **confidence** of **variation**?

Which authors quote the source more literally and why?

› … tell us something about **transmission contamination**?
What are the **differences** and **similarities**, what do they **tell** us about the texts?

- **How** does one **classify** the similarities and the differences (**Wittgenstein**)?
  
  How and can those findings help to improve **NLP**?
The reuse of text across languages is a gold mine for transcultural studies.

When observing translated text,…

› …what is the scale of divergence from one translation to the other?

› …what is the divergence caused by? Translator competence or linguistic, cultural, ideological and political norms?

› Could machine translators be used to clean a translation of any contextual influences?
Challenge 1: Reuse Type Diversity

How to detect reuse automatically?

Stability (yellow): syntactic vs. semantic

Purpose/Intention (green)

Size of Text Reuse (blue)

Literary Classification (light blue)

Degree of distribution (purple)
Challenge 2: Author specific Reuse Styles

How to search reuse on big corpus where every author refers to another in a different way
Challenge 3: Historical Changes

How to deal with text variants?
- language evolution
- dialects
- “spelling errors”
- copy errors (by scribes in the Middle Ages)

https://commons.wikimedia.org/wiki/File:Greek-dialects-mod.jpg#file
Tracer – Current Approach

Split the problem into 6 levels of processing (Hackathon)
Thank you for your attention!

http://etrap.gcdh.de

Marco Büchler, Emily Franzini, Greta Franzini, Maria Moritz
Introduction
Team
Challenges