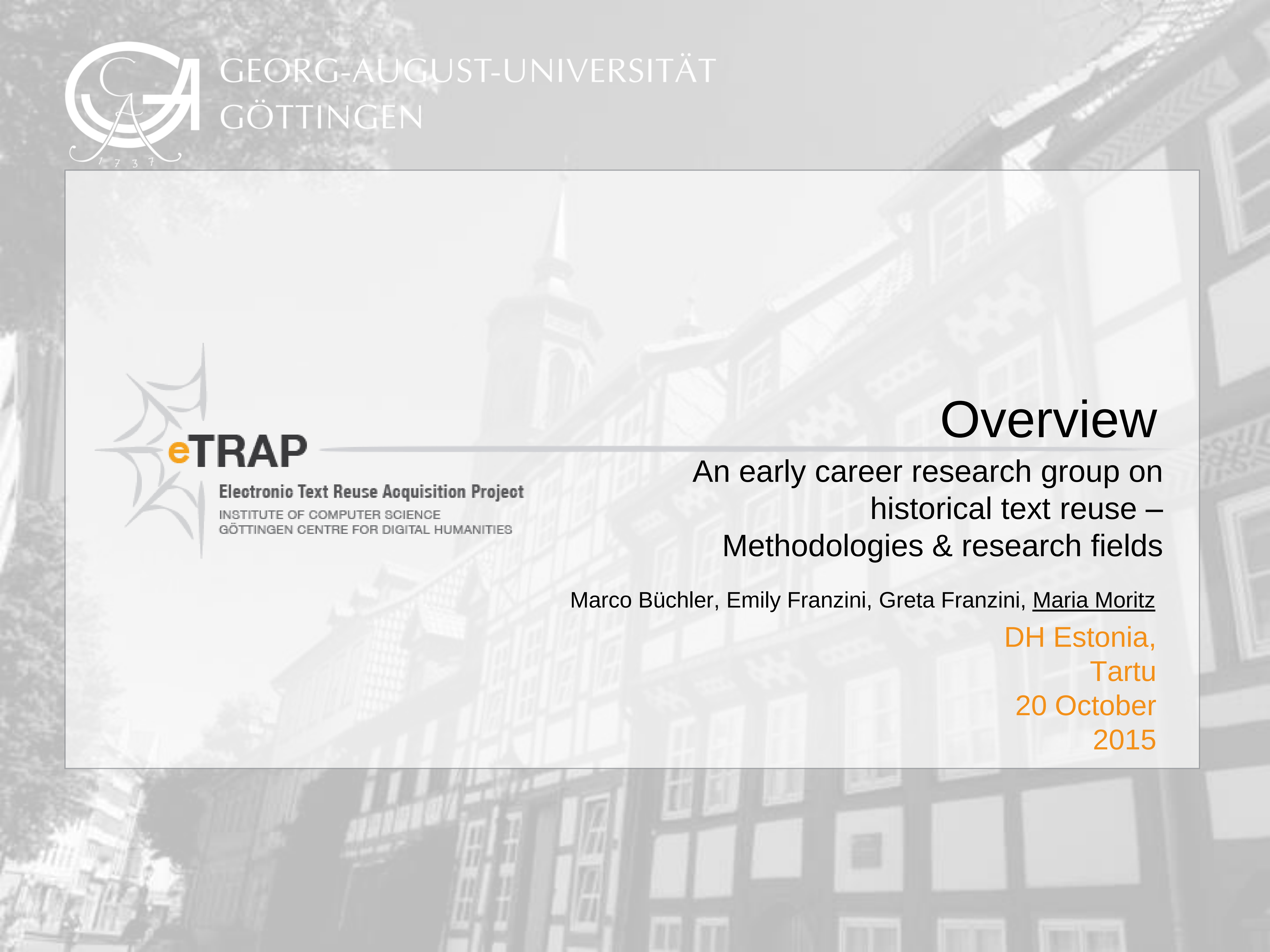


GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

eTRAP

Electronic Text Reuse Acquisition Project

INSTITUTE OF COMPUTER SCIENCE
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

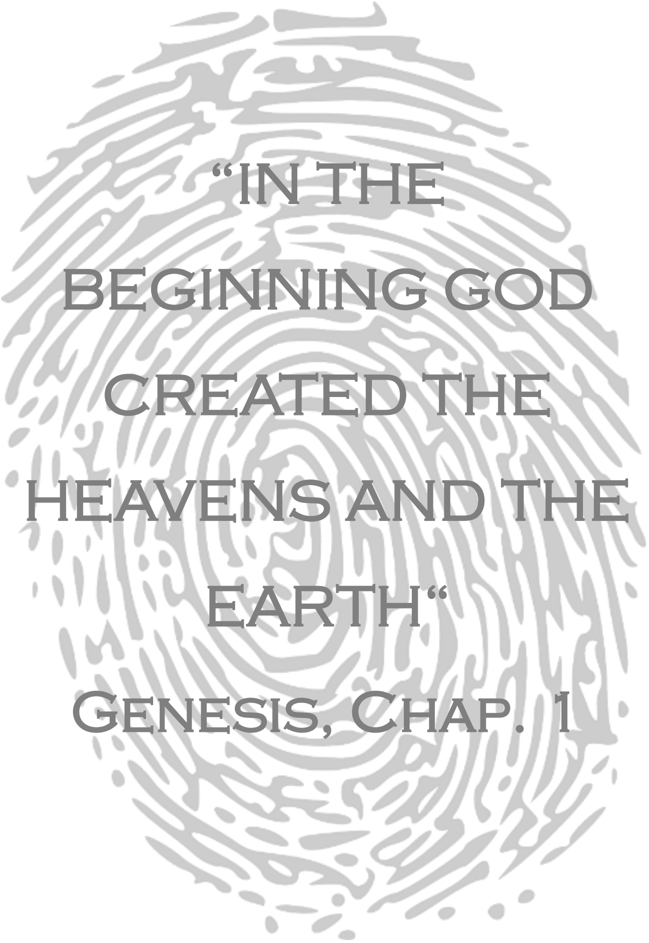
Overview

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



“IN THE BEGINNING GOD CREATED THE HEAVENS AND THE EARTH“

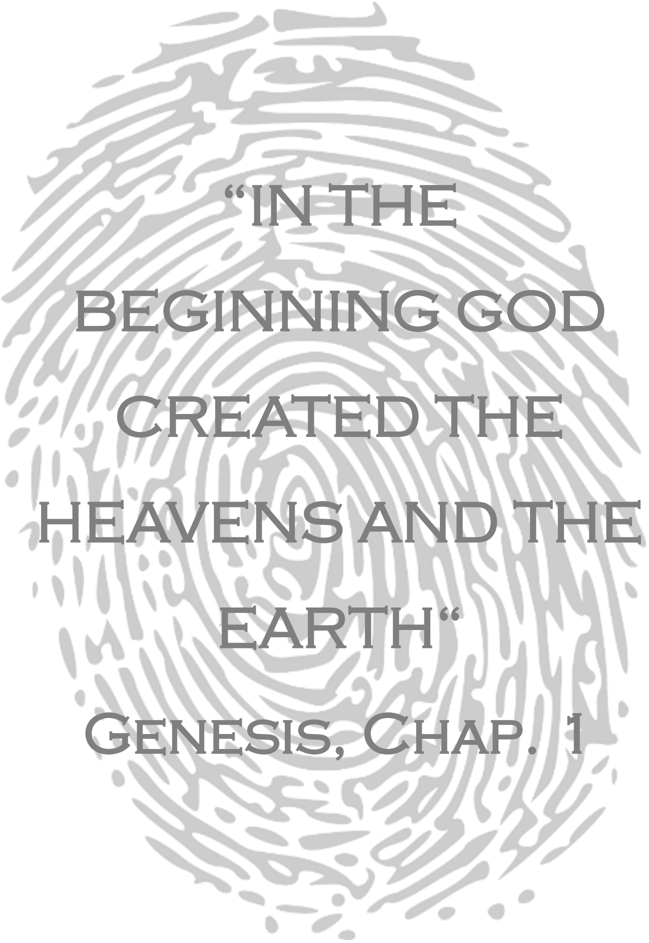
GENESIS, CHAP. 1

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



"In the beginning God created the heavens and the earth"

Genesis, Chap. 1

Question

Why is it so relevant for Humanities and Computer Science?

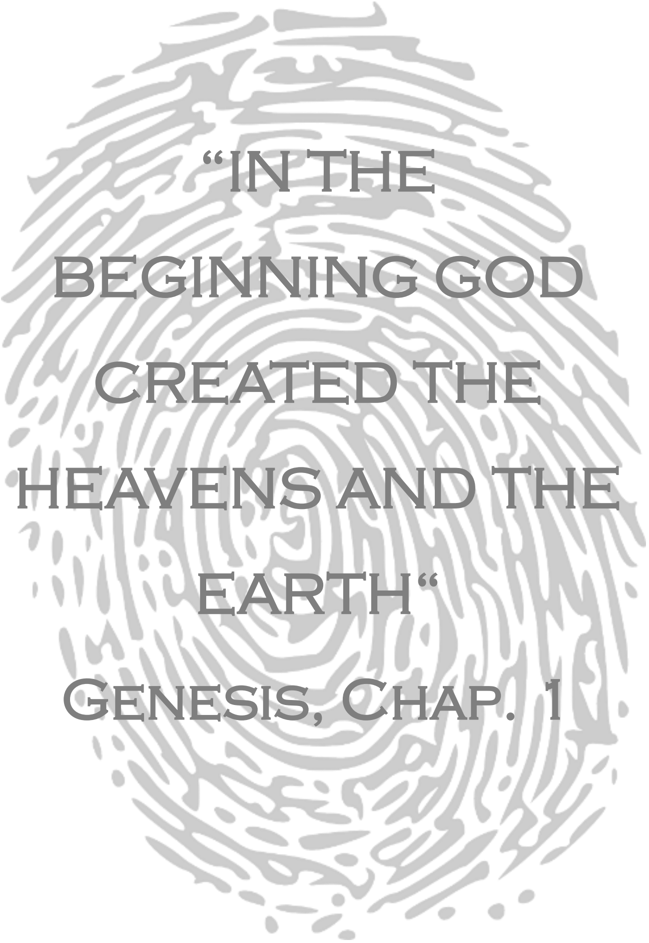
Humanities

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



“IN THE BEGINNING GOD CREATED THE HEAVENS AND THE EARTH“

GENESIS, CHAP. 1

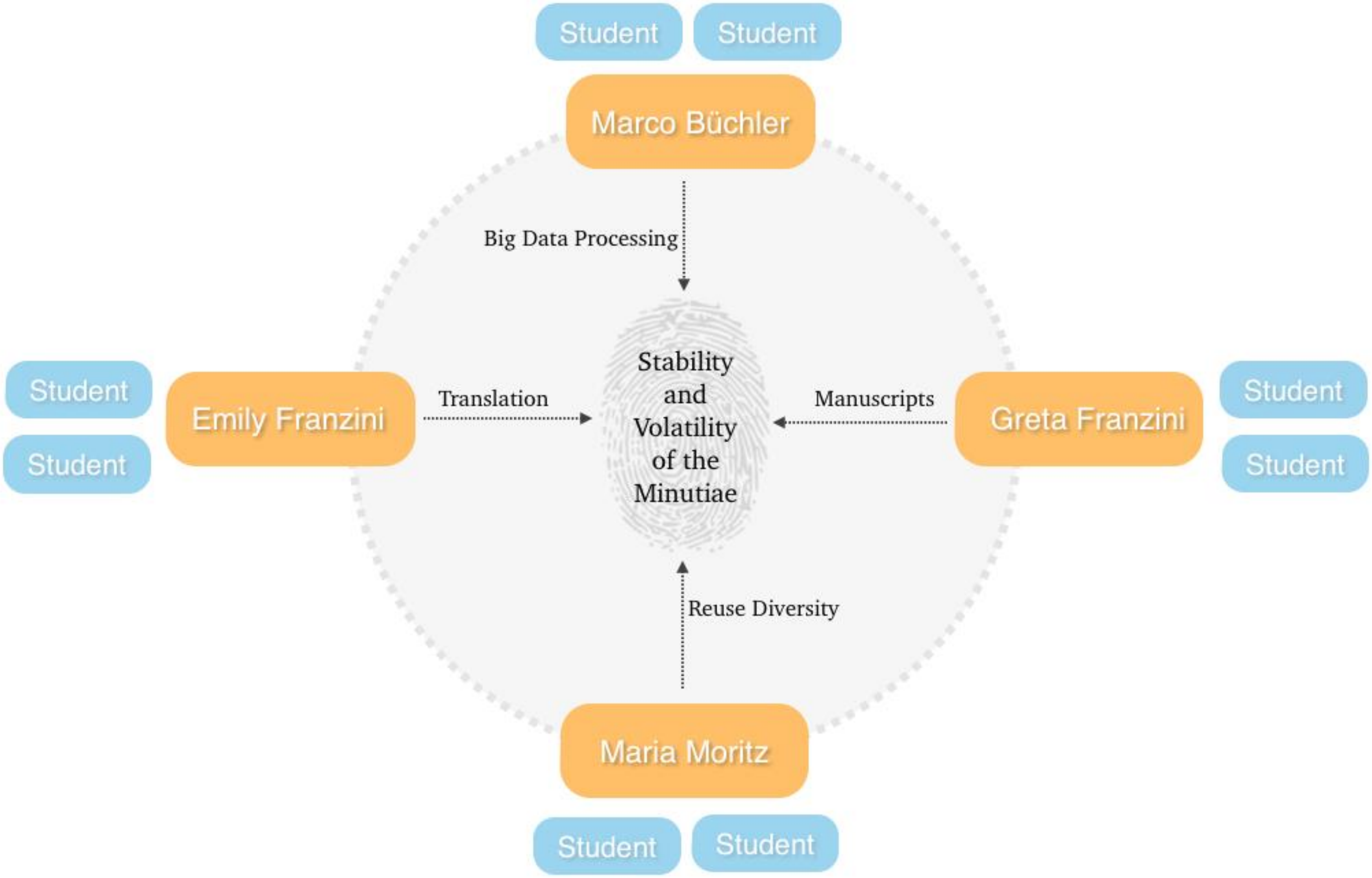
How
Were paraphrases, allusions or translations **reused** by **authors** over **time**?

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

Team



Student

Student

Marco Büchler

Big Data Processing

Student

Student

Emily Franzini

Translation

Stability and Volatility of the Minutiae

Manuscripts

Greta Franzini

Student

Student

Reuse Diversity

Maria Moritz

Student

Student

Marco Büchler, Comp. Sc.

???

- 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

???

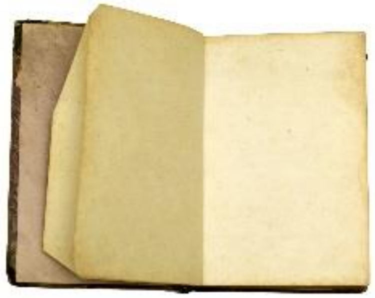
Can the study of text reuse…

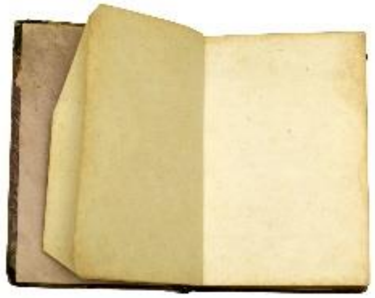
- …tell us something about the **accuracy** and **confidence** of **variation**?

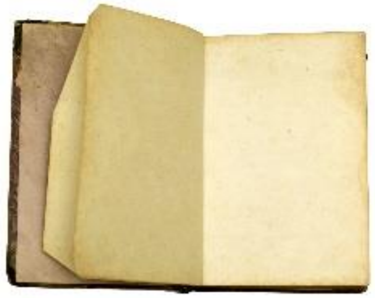
Which authors quote the source more literally and why?

- … tell us something about **transmission contamination**?

Maria Moritz, Comp. Sc.







???

- 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**?

Emily Franzini, Humanist

BIBLIA

τὰ βιβλία τὰ ἅγια

THE BIBLE

LA BIBBIA

DIE BIBEL

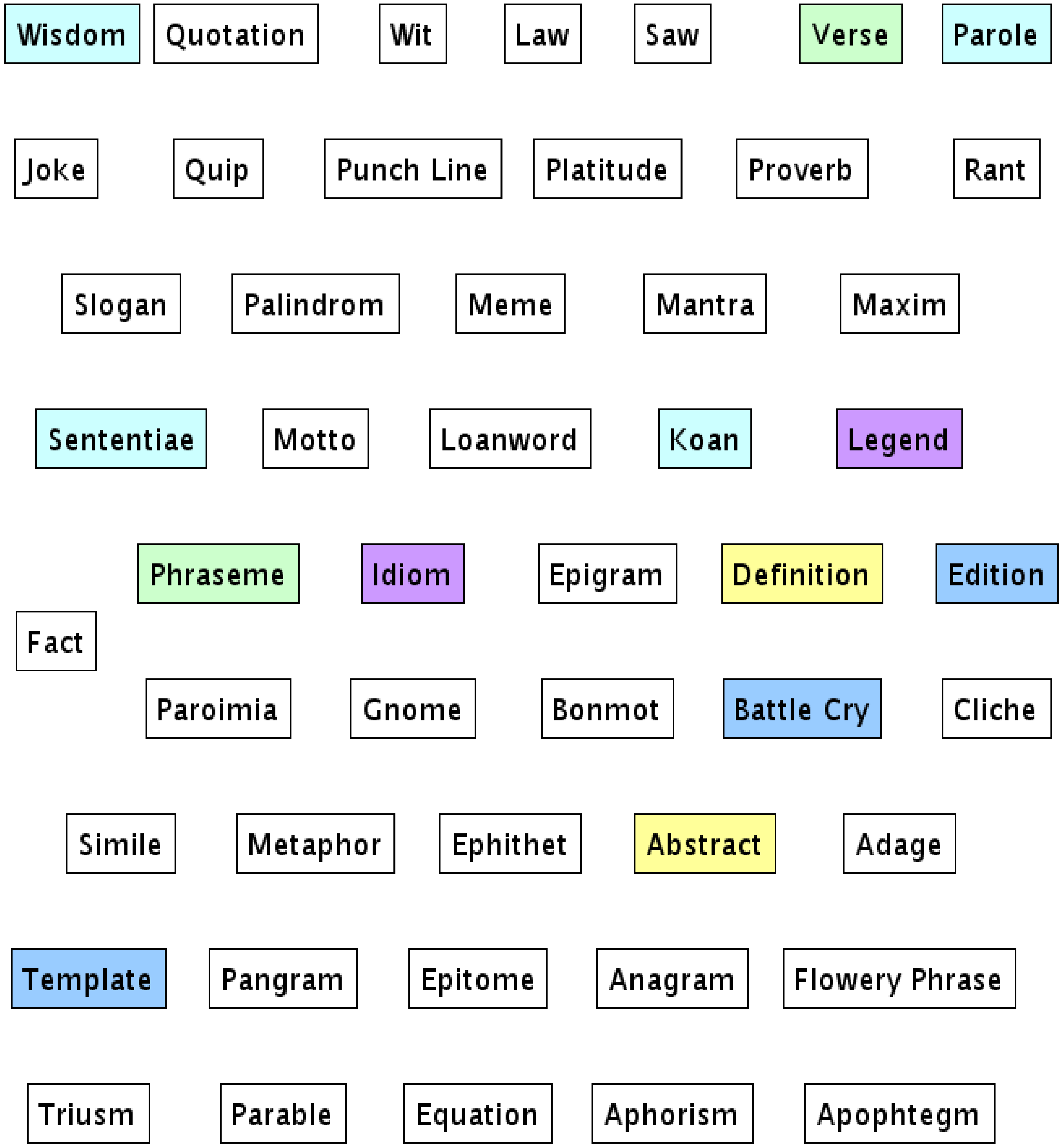
???

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



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

How to detect reuse automatically?

Stability (yellow):
syntactic vs. semantic

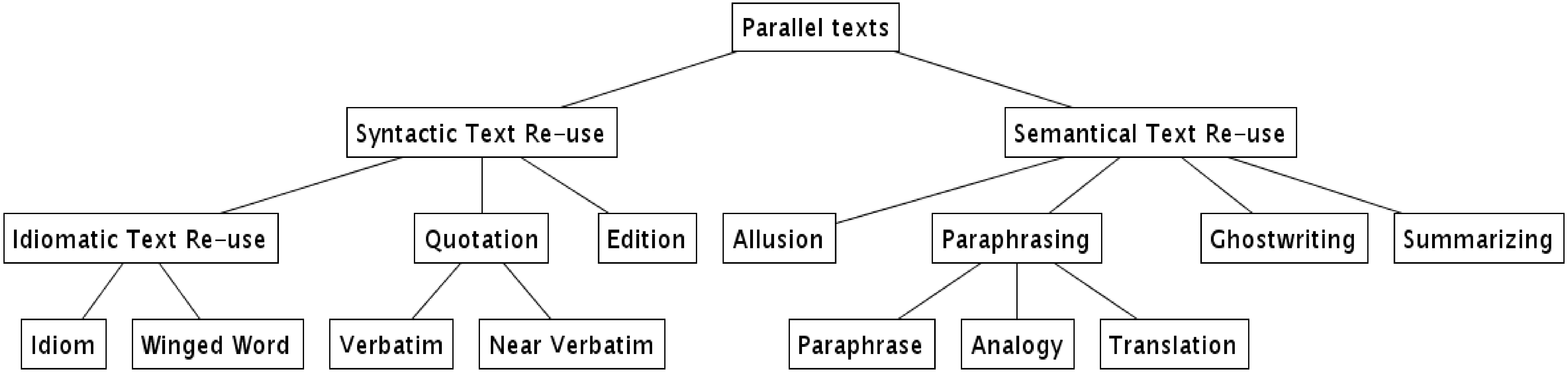
Purpose/Intention (green)

Size of Text Reuse (blue)

Literary Classification (light blue)

Degree of distribution (purple)

Challenge 2: Author specific Reuse Styles



Parallel texts

Syntactic Text Re-use

Semantical Text Re-use

Idiomatic Text Re-use

Quotation

Edition

Allusion

Paraphrasing

Ghostwriting

Summarizing

Idiom

Winged Word

Verbatim

Near Verbatim

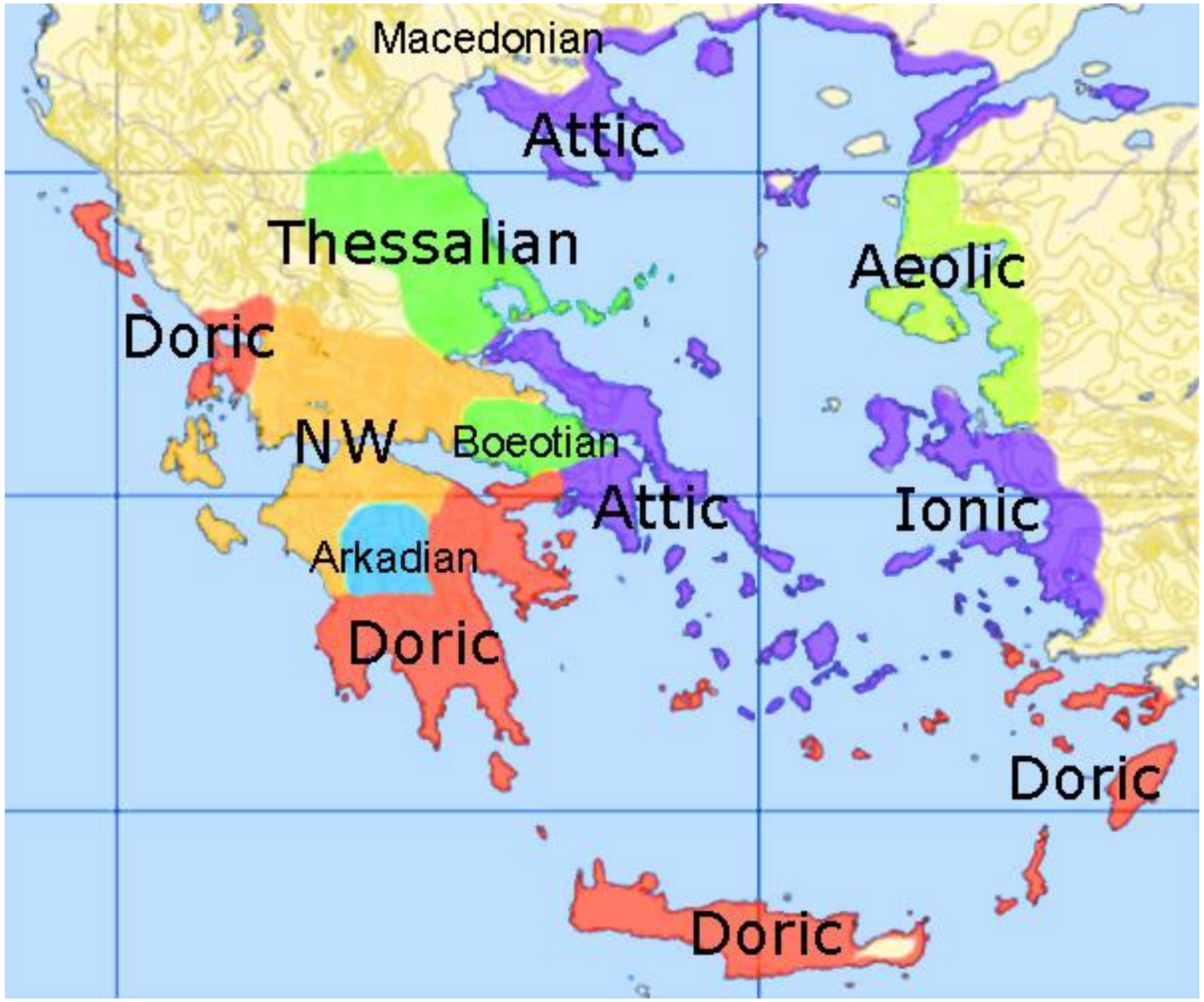
Paraphrase

Analogy

Translation

How to search reuse on big corpus where every author refers to another in a different way

Challenge 3: Historical Changes



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

How to deal with text variants?

language evolution

dialects

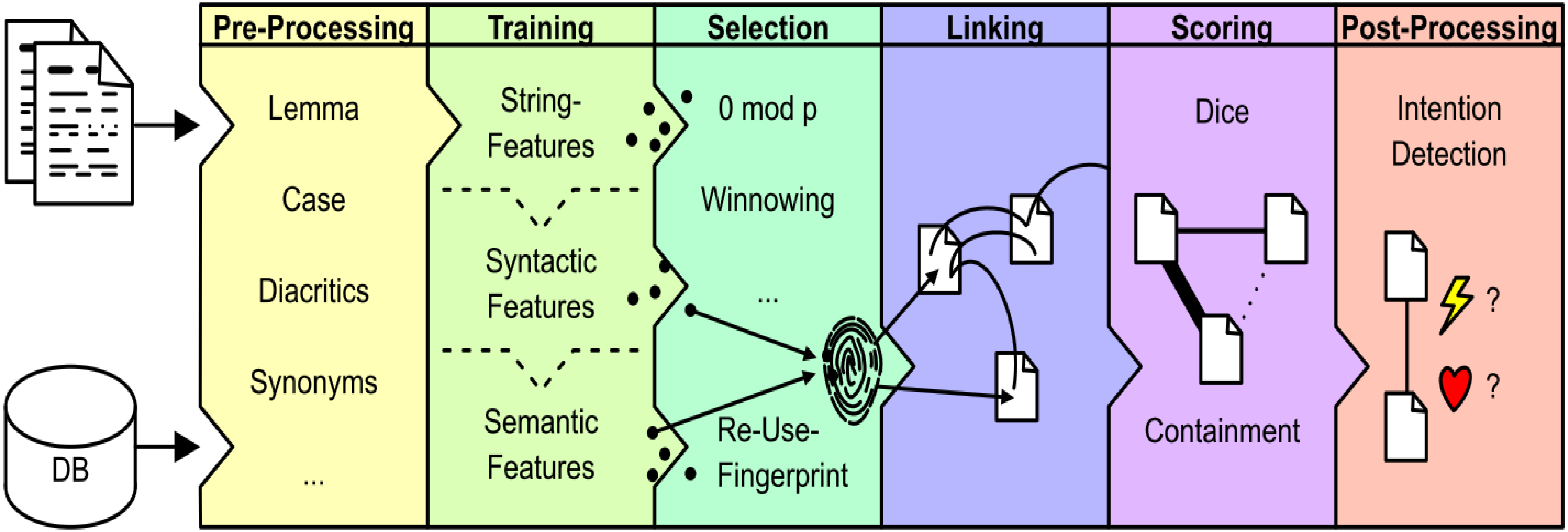
“spelling errors”

copy errors (by scribes in the Middle Ages)

eTRAP

Tracer – Current Approach

Split the problem into 6 levels of processing (Hackathon)



eTRAP

Electronic Text Reuse Acquisition Project

INSTITUTE OF COMPUTER SCIENCE
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

Thank you for your attention!

http://etrap.gcdh.de

Marco Büchler, Emily Franzini, Greta Franzini, Maria Moritz

Introduction

Team

Challenges