eTRAP
electronic Text Re-use Acquisition Project

Marco Büchler

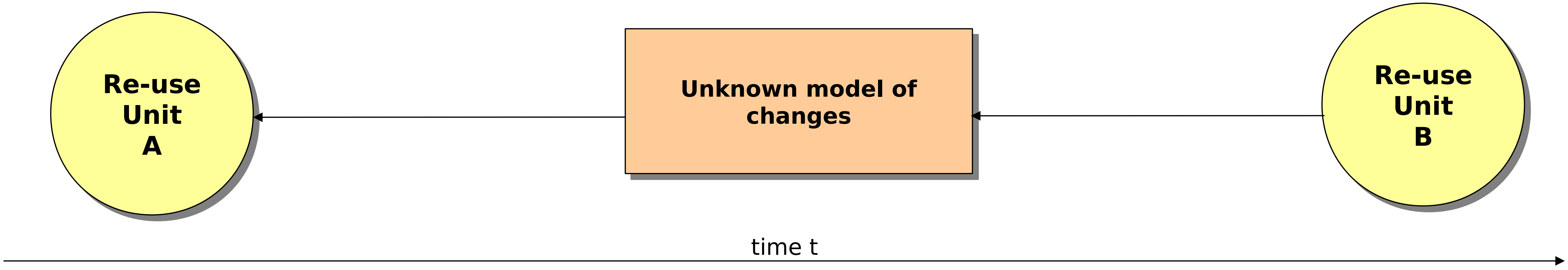
Dhd conference, Graz, Austria
February 24th 2015

What is (Historical) *Text Re-use*?

General: Text Re-use describes the spoken and written repetition of content.

Example: quotations, paraphrases but also translations

Historical changes: language evolution, different dialects, “spelling errors” but also copy errors (by monks in the Middle Ages)



Historical Text Re-use as an Opportunity for Humanities and Computer Science

Question: Why is Text Re-use so fundamental for Humanities and Computer Science?

Premise: the amount of digitally available data grows exponentially (Big Data)

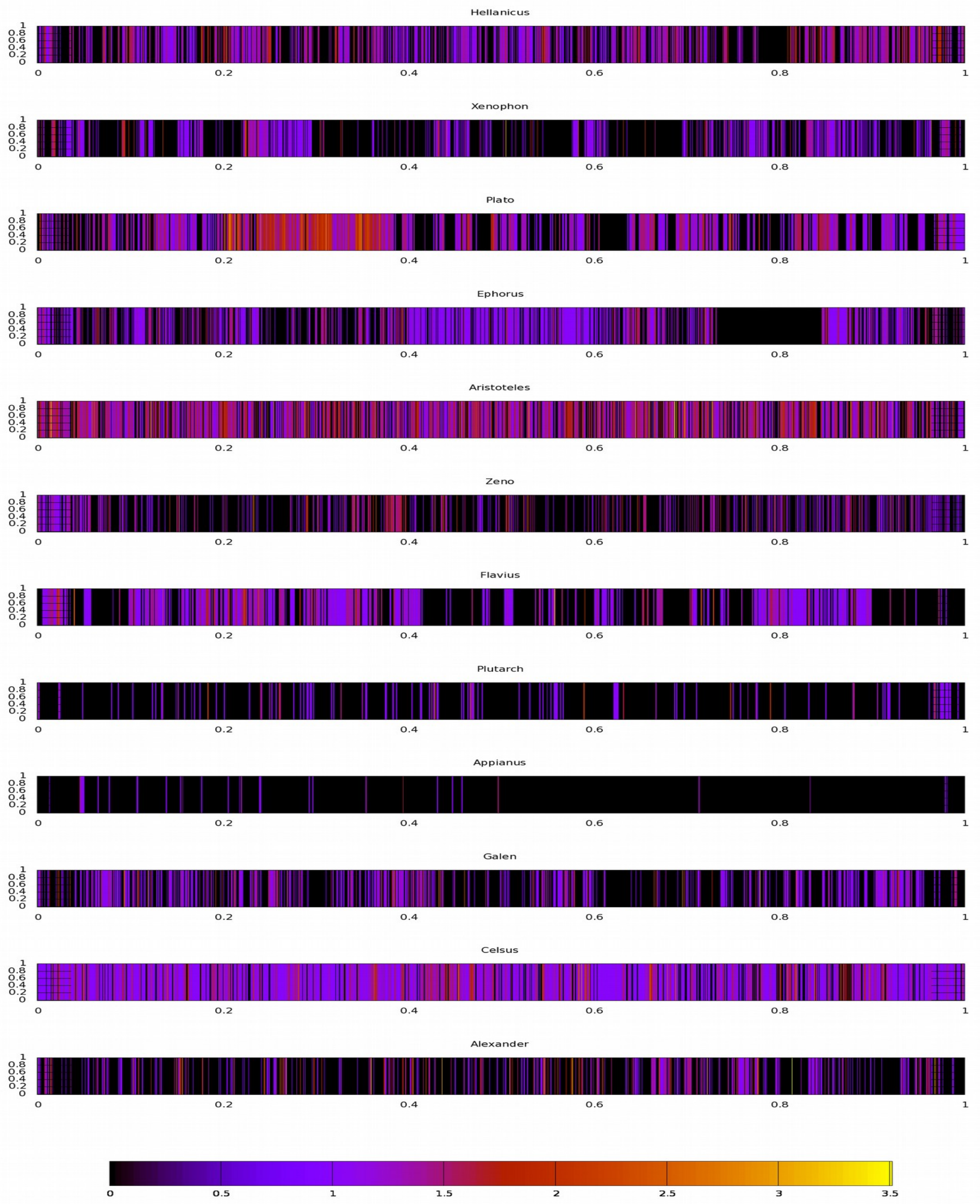
Humanities:

- Lines of transmissions and **textual criticism**
- **Transmissions of ideas/thoughts** under different circumstances and conditions

Computer Science:

- **Text Decontamination** for stylometry and authorship attribution, dating of texts
- gen. Text Mining, Corpus Linguistics

Temperature Map



„Pecunia Non Olet“

„Money does not smell“

Source: http://cf.geekdo-images.com/images/pic606797_md.jpg

What is *Big Data*?

3 aspects of Big Data (by Ulrike Rieß, *Big Data bestimmt die IT-Welt*):

1) **Huge amount of data** that can‘t be processed and analyzed manually

2) **Less structured data**; e. g. in comparison to databases and data warehouse systems

3) **Linked data** between heterogeneous and distributed resources

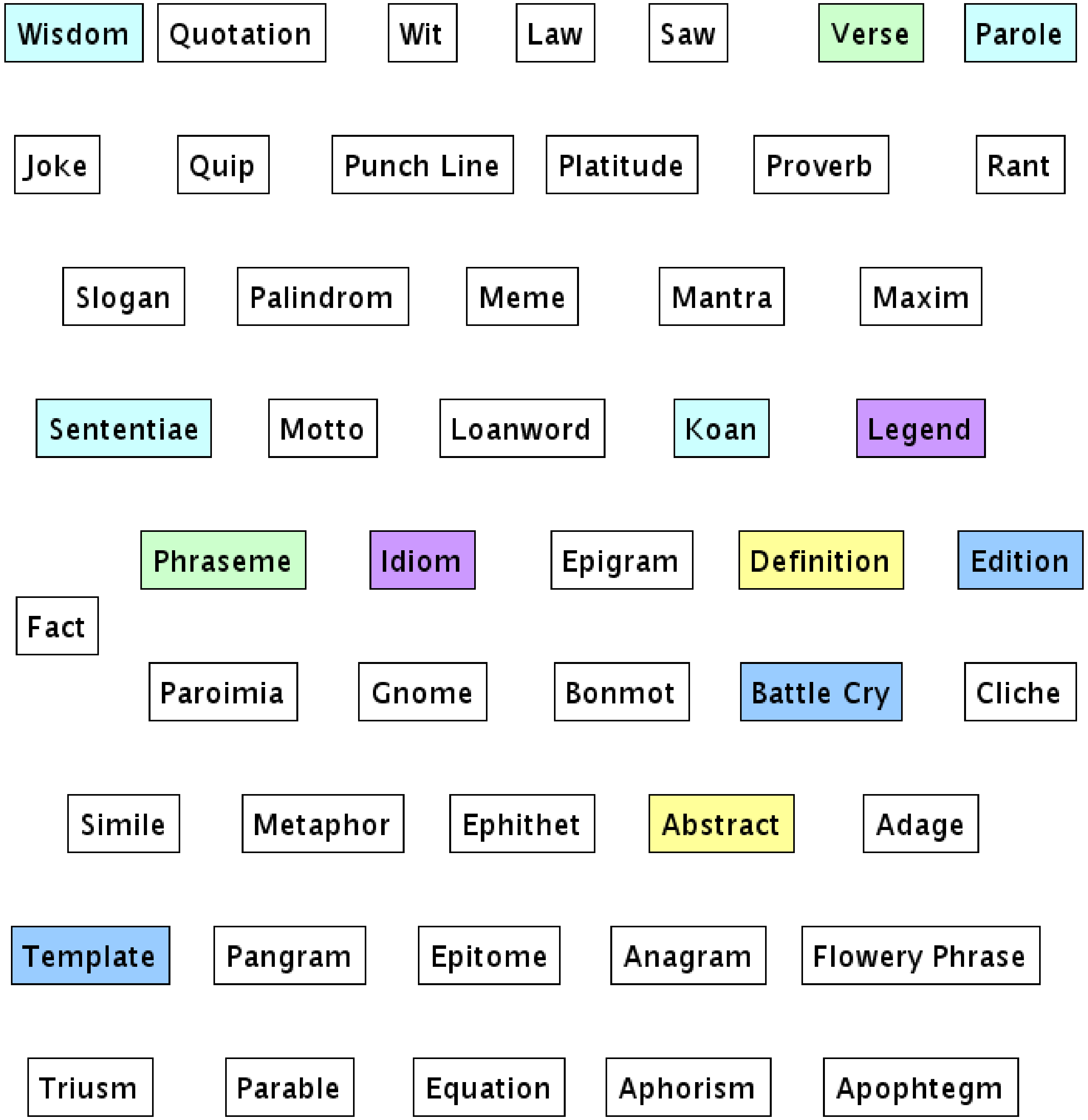
The fastest growing sources of Big Data are **text** and **images**.

Researchers easily get lost in the **information overload** (Big Data) and in the **information poverty** (Humanities Data).

A basic question

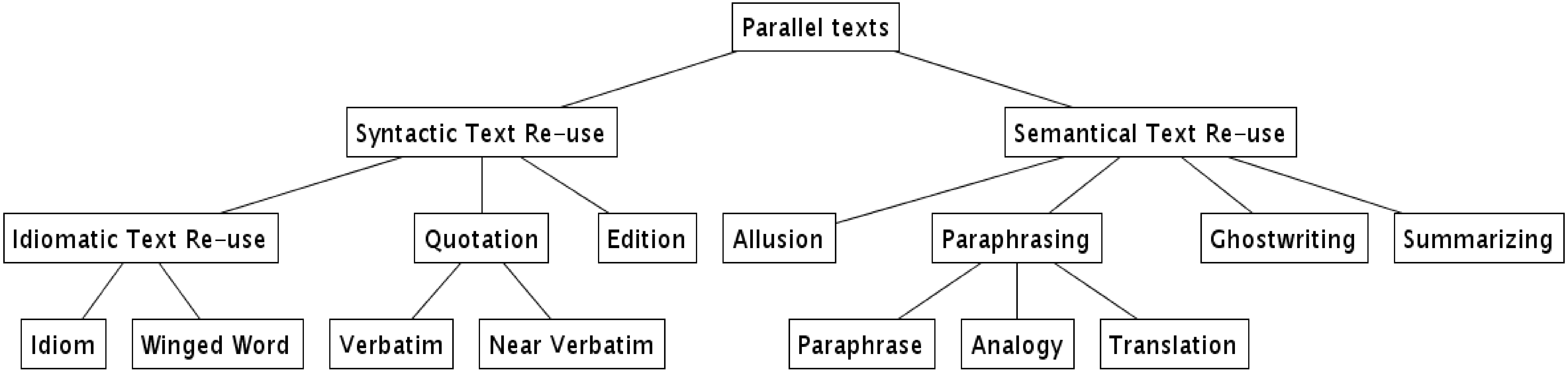
What are the algorithm’s requirements for text re-use?

Complete view: **Re-use Types**



- **Stability (yellow)**: syntactic vs. semantic
- **Purpose (green)**
- **Size of Text Re-use (blue)**
- **Literary classification (light blue)**
- **Degree of distribution (purple)**
- **Written and oral transmission**

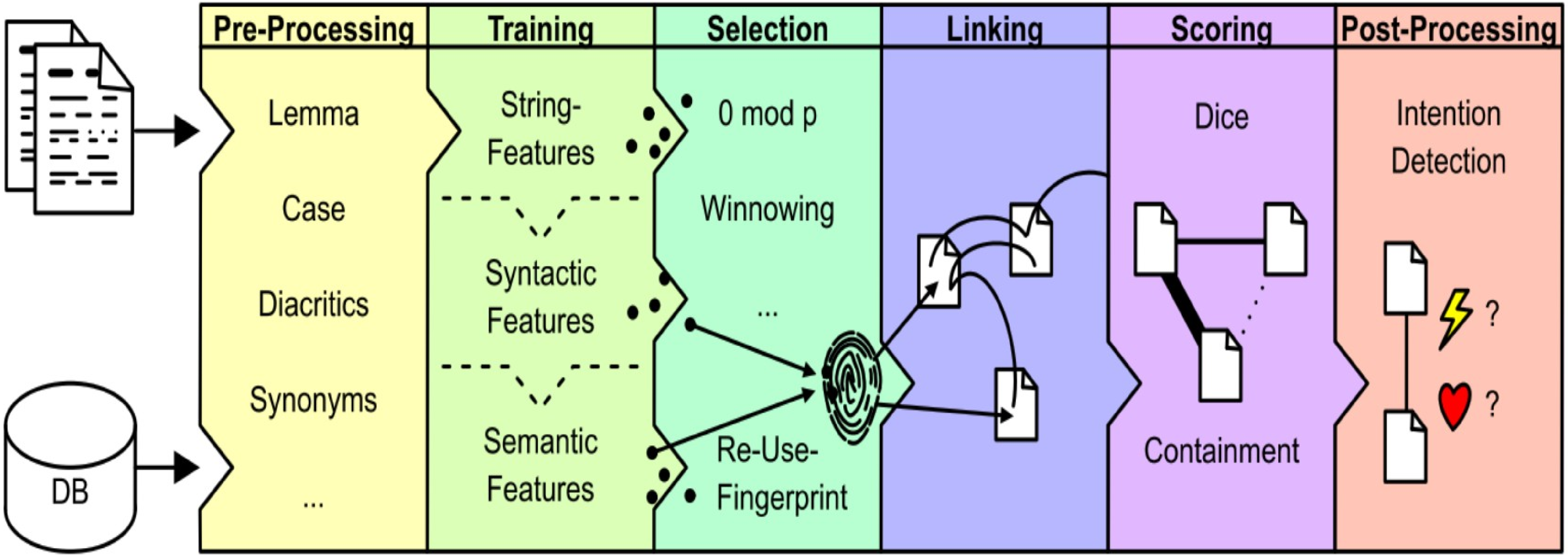
Complete view: **Re-use Styles**



Basic Question

Basic question: Distribution of *Re-use Types* und *Re-use Styles* are often unknown. Question: Which model(s) should be chosen and how to evaluate the results?

Recent Approach

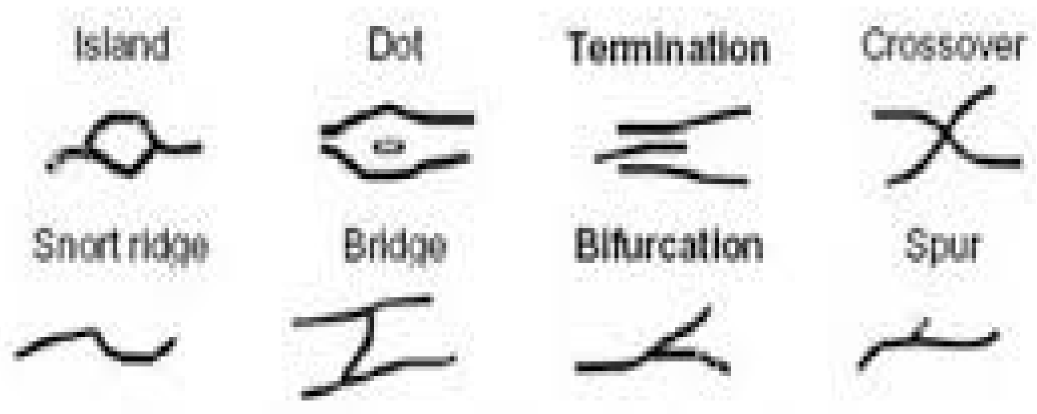


Implemented in TRACER software: more than a million permutations of implementations of different levels are now possible

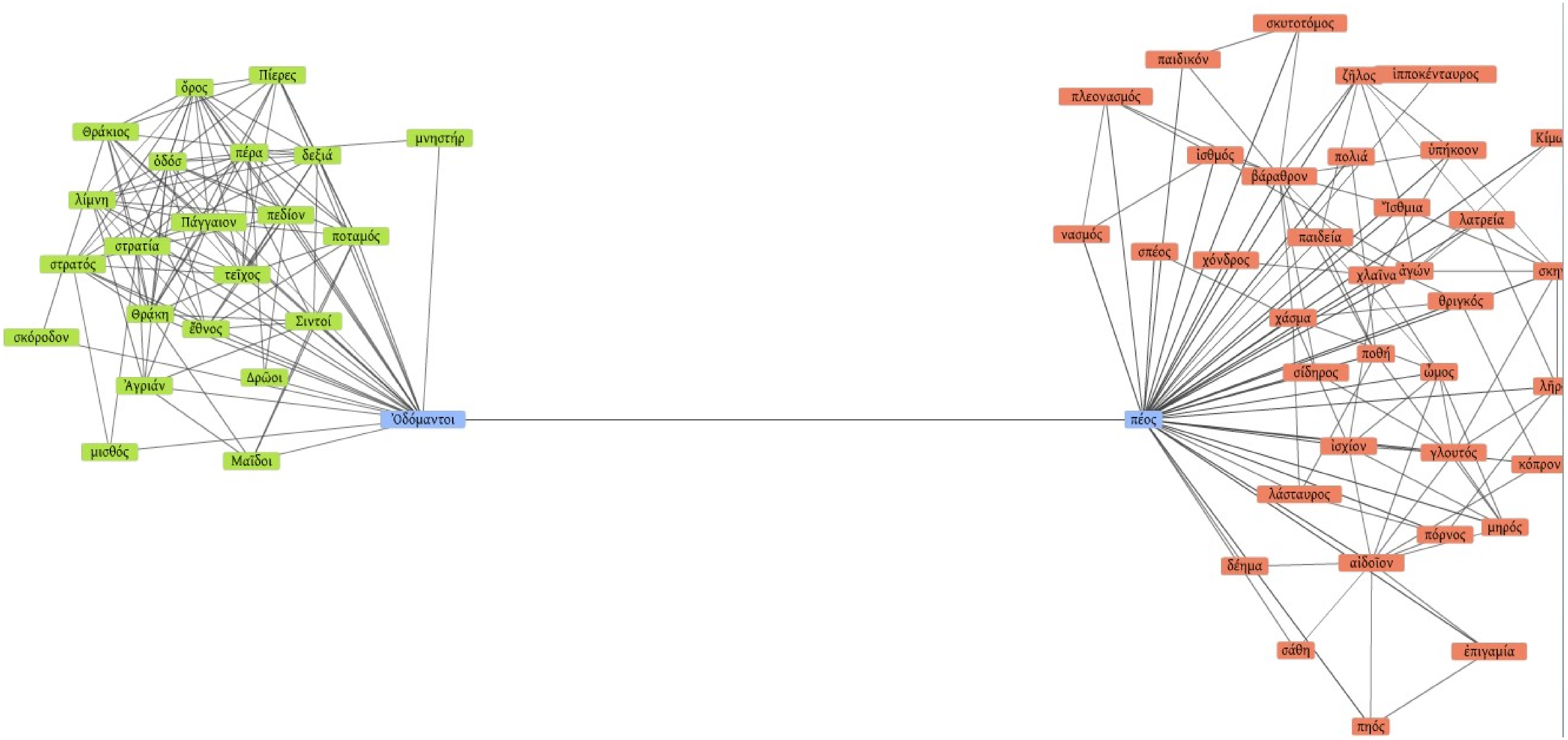
eTRAP: Resulting Questions

Question: What are the common primitives in the re-use diversity?

From biometry (Minutiae):

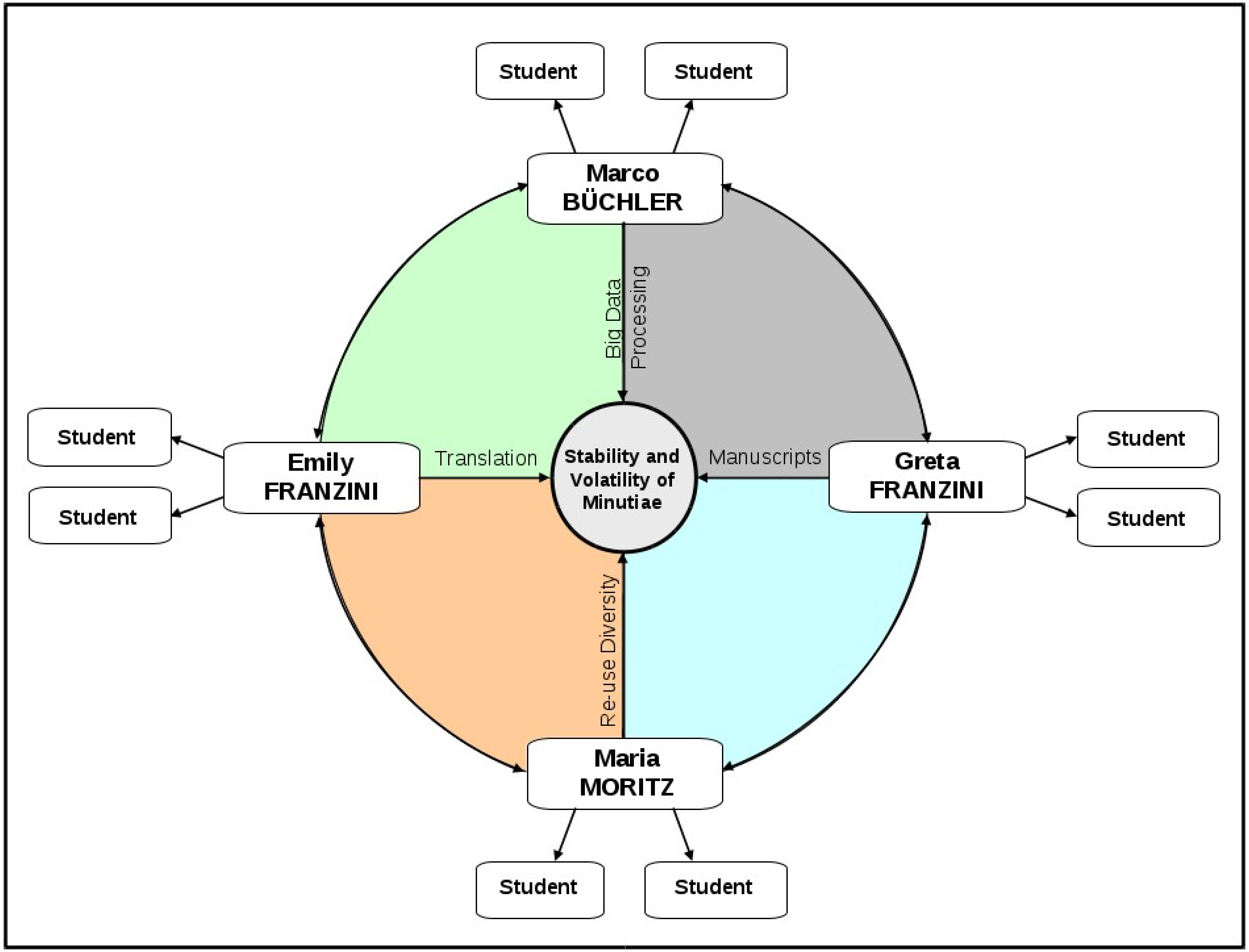


Identifying Passages of Interest in Text: Visualising Contrastive Semantics



Source: F. Baumgardt: Visualisierung von Kookkurrenzgraphen. Bachelorarbeit Abteilung Automatische Sprachverarbeitung, Universität Leipzig, 2010.

eTRAP: staff setup



eTRAP: The Team (Lyon 2014)

eTRAP: Marco Büchler (Big Data Processing)

Marco Büchler holds a Diploma in **Computer Science**. Since 2006 he has worked as a Research Associate in the **Natural Language Processing Group** at Leipzig University. From April 2008 to March 2011 Marco served as the technical Project Manager for the **eAQUA** project and continued to work in the capacity for the following eTRACES project. In March 2013 he received his PhD in the field of eHumanities. Since May 2014 he leads a Digital Humanities Research Group at the Göttingen Centre for Digital Humanities. His research includes Natural Language Processing on Big Humanities Data. Specifically, he works on **Historical Text Re-use Detection** and its **application in the business world**. In addition to his primary responsibilities, Marco manages the Medusa project (Big Scale co-occurrence and ngram framework) as well as the **TRACER** framework for detecting historical text re-use.

eTRAP: Greta Franzini (Manuscripts)

A Liceo Classico graduate, Greta completed her **Classics BA** and **Digital Humanities M**A degrees at **King’s College** London. Greta is currently doing a PhD at the **UCL Centre for Digital Humanities** where her research will ultimately produce a digital edition of an ancient Latin manuscript. Greta’s interests lie within the fields of Classics, Philology, Manuscript Studies and **Electronic Editing**. Previously, she worked as a Research Associate at the **Humboldt Chair of Digital Humanities** at the University of Leipzig.

Greta is fluent in Italian and English, her native languages, speaks Spanish as well as conversational Modern Greek, German and French.

eTRAP: Maria Moritz (Text Re-use Diversity)

Maria followed up her **Bachelor of Computer Science** with a Master’s thesis on *information extraction from Ancient Greek* texts. She worked as a Research Associate for the **Natural Language Processing** (NLP) Chair at the University of Leipzig before she joined the **Humboldt Chair of Digital Humanities** at the same university. Maria’s interests revolve around the adoption of natural language approaches to research questions in the humanities, particularly by means of **annotation application**s and **pattern recognition**.

Maria is fluent in German and English, with elementary proficiency in French.

eTRAP: Emily Franzini (Translations)

Emily completed a **Classics BA** and a **Management Science & Innovation MSc** degree at **University College London**. This unique combination of areas of study has led her to work in **strategy consulting** and for a not-for-profit organisation supporting the **preservation of cultural heritage**. Emily's interests lie within the fields of **Classics, Translation Studies, Bilingualism and Machine Translation**. Before Göttingen, Emily worked as a Research Associate at the **Humboldt Chair of Digital Humanities** at the University of Leipzig.

Emily is fluent in her native Italian and English and can converse in German, Spanish and French.

eTRAP: Conclusion

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

eTRAP

GCDH GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

Visit us via http://etrap.gcdh.de