TEACHING DATA SCIENCE

AN EXPERIENCE REPORT FROM SIX INTERNATIONAL TRACER TUTORIALS AND WORKSHOPS WITH MIXED CLASSES

Greta Franzini, Emily Franzini, Elisa Cugliana, Nicoletta Guido, Marco Büchler
1. Who am I?

2. What do you associate with text reuse?

3. What is data science?

4. Comparison of Luke & Mark

5. Acquiring knowledge, practising knowledge, sharing knowledge

6. Data science & precision and recall
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.
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 = spoken and written repetition of text across time and space.

Figure 1: Text reuse styles.
WHAT IS DATA SCIENCE?
DIVERSITY (REUSE TYPES)

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

Parallel texts

- Syntactic Text Re-use
  - Idiomatic Text Re-use
    - Idiom
    - Winged Word
  - Quotation
  - Edition
  - Allusion

- Semantic Text Re-use
  - Paraphrasing
    - Paraphrase
    - Analogy
    - Translation
  - Ghostwriting
  - Summarizing
Question:

The distribution of **Reuse Types** and **Reuse Styles** is often unknown - which **model(s)** should be chosen?
COMPARISON OF LUKE & MARK
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 is language-independent.
Tested on: Ancient Greek, Arabic, Coptic, English, German, Hebrew, Latin, Tibetan.
**Segmentation:** disjoint and verse-wise segmentation.

<table>
<thead>
<tr>
<th>Preprocess</th>
<th>Featuring</th>
<th>Trigram</th>
<th>Bigram</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>$S_{11}$</td>
<td>$S_{21}$</td>
<td>$S_{31}$</td>
<td></td>
</tr>
<tr>
<td>StringSim</td>
<td>$S_{12}$</td>
<td>$S_{22}$</td>
<td>$S_{23}$</td>
<td></td>
</tr>
<tr>
<td>Lemma</td>
<td>$S_{13}$</td>
<td>$S_{23}$</td>
<td>$S_{33}$</td>
<td></td>
</tr>
<tr>
<td>Lemma + Syn</td>
<td>$S_{14}$</td>
<td>$S_{24}$</td>
<td>$S_{34}$</td>
<td></td>
</tr>
</tbody>
</table>

**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.
ACQUIRING KNOWLEDGE, PRACTISING KNOWLEDGE, SHARING KNOWLEDGE
ACQUIRING KNOWLEDGE
PRACTISING KNOWLEDGE IN TEAMS
PRACTISING KNOWLEDGE WITH TEAMS
DATA SCIENCE & PRECISION AND RE-CALL
EXPECTATIONS OF A HUMANIST: OVERSIMPLIFICATION
Webpage: http://www.etrap.eu/research/tracer
Repository: http://vcs.etrap.eu/tracer-framework/tracer.git
Upcoming tutorials:

- Three more tutorials in 2017 pending confirmation.
ETRAPSTERS WARMLY WELCOME YOU IN GÖTTINGEN!
Visit us

🌐 http://www.etrap.eu
✉️ contact@etrap.eu

_Stealing from one is plagiarism, stealing from many is research_  
_(Wilson Mitzner, 1876-1933)_

SPONSORED BY THE

Federal Ministry of Education and Research
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.