

TEACHING DATA SCIENCE

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

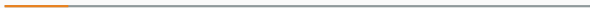
Greta Franzini, Emily Franzini, Elisa Cugliana, Nicoletta Guido, Marco Büchler



TABLE OF CONTENTS

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.

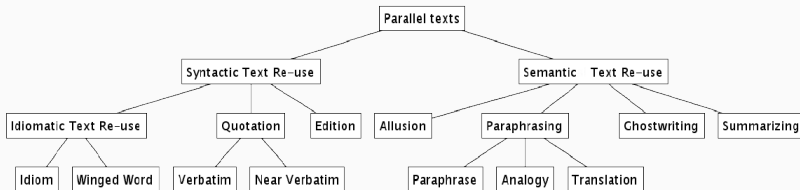


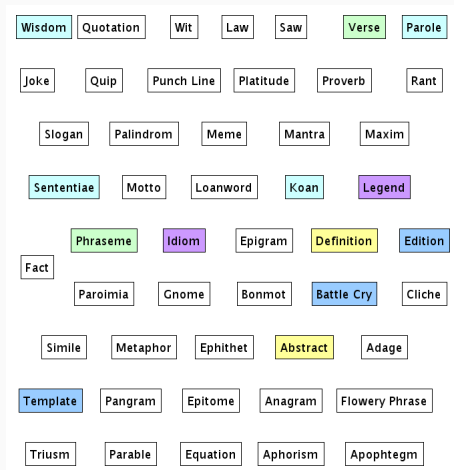
Figure 1: Text reuse styles.

EXPECTATIONS OF A HUMANIST: OVERSIMPLIFICATION



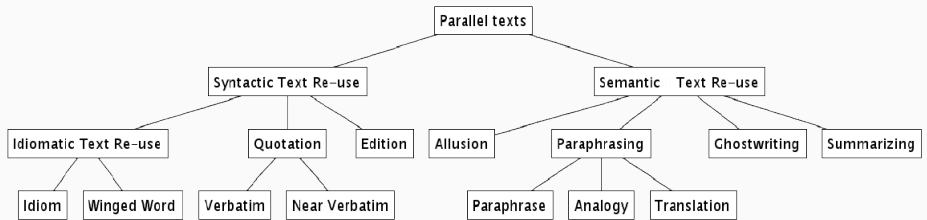
WHAT IS DATA SCIENCE?

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)



Question:

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

COMPARISON OF LUKE & MARK

TRACER: OVERVIEW

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

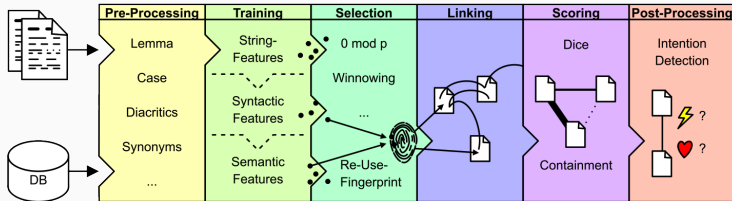


Figure 2: Detection task in six steps. More than 1M permutations of implementations of different levels are possible.

TRACER is language-independent.

Tested on: Ancient Greek, Arabic, Coptic, English, German, Hebrew, Latin, Tibetan.

Segmentation: disjoint and verse-wise segmentation.

		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;

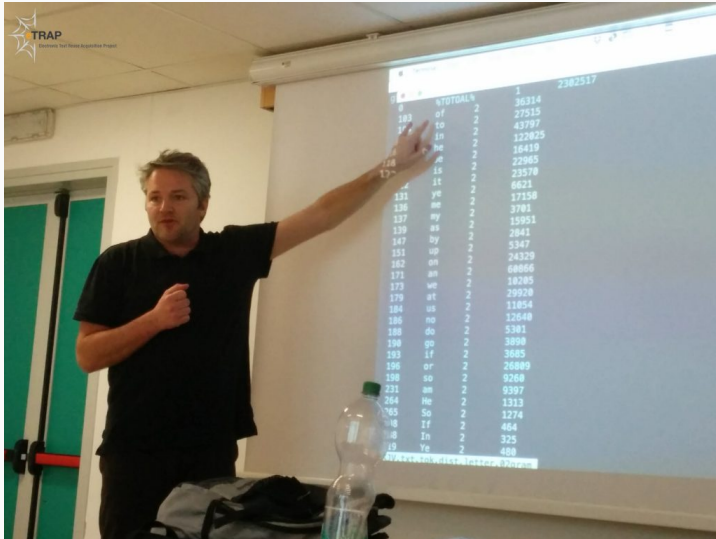
Linking: Inter- Digital Library Linking (different Bible editions);

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

Post-processing: not used.

ACQUIRING KNOWLEDGE, PRACTIS- ING KNOWLEDGE, SHARING KNOWL- EDGE

ACQUIRING KNOWLEDGE



The image shows a man in a black shirt pointing at a projected list of words and their frequencies on a screen. The list is titled "TOTAL%" and shows the frequency of various words. The man is standing next to a green door and a black bag. A water bottle is visible in the foreground.

Rank	Word	Frequency
1	of	36314
2	to	27515
3	in	43797
4	he	122825
5	is	16419
6	it	22965
7	at	23570
8	on	6621
9	by	17158
10	me	3701
11	my	15951
12	as	2841
13	for	5347
14	up	24329
15	an	68866
16	with	18285
17	the	29920
18	and	11854
19	us	12640
20	no	5381
21	do	3890
22	go	3685
23	if	26809
24	or	9260
25	so	9397
26	am	1313
27	He	1274
28	So	464
29	If	325
30	In	488
31	Ye	

A group of people, including students and staff, are working on laptops in a classroom. A tablet in the foreground displays a terminal window with the word 'HACKING' and some code. The background shows a classroom setting with large windows and a red wall. The people are focused on their work, and the atmosphere appears to be one of collaborative learning and technical exploration.

PRACTISING KNOWLEDGE IN TEAMS



PRACTISING KNOWLEDGE WITH TEAMS



SHARING KNOWLEDGE



SHARING KNOWLEDGE



DATA SCIENCE & PRECISION AND RE- CALL

EXPECTATIONS OF A HUMANIST: OVERSIMPLIFICATION



Webpage: <http://www.etrp.eu/research/tracer>

Repository: <http://vcs.etrp.eu/tracer-framework/tracer.git>

Upcoming tutorials:

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

ETRAPSTERS WARMLY WELCOME YOU IN GÖTTINGEN!



Visit us



<http://www.etrp.eu>



contact@etrp.eu

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



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN



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.

