TRACER TUTORIAL: TEXT REUSE DETECTION

Introduction to Historical Text Reuse Detection

Marco Büchler, Emily Franzini and Greta Franzini

TRACER
TEXT REUSE DETECTION MACHINE

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

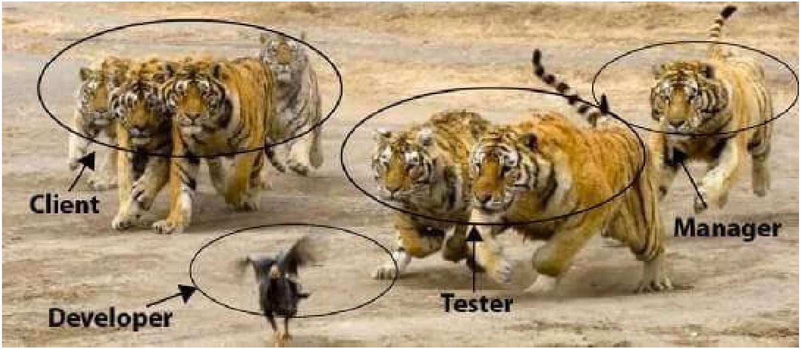
1. Who am I?
2. What is text reuse?
3. Aspects of text reuse
4. ACID for the Digital Humanities
5. Big (Humanities) Data
6. Language Model

WHO AM I?

WHO AM I?

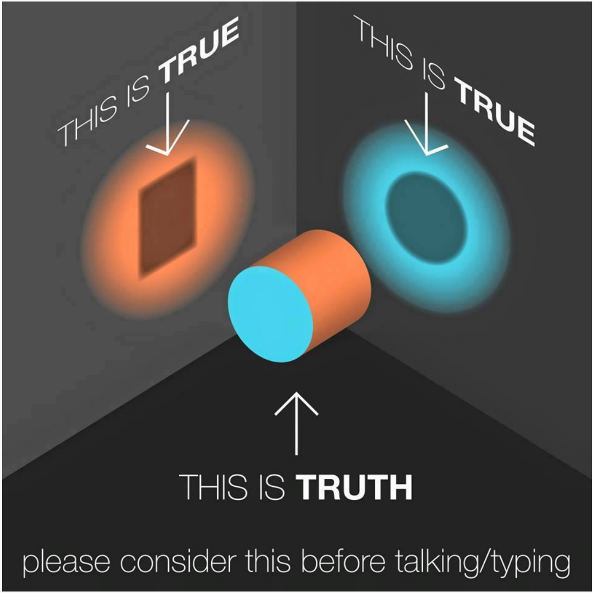
- 2001-2002: Head of Quality Assurance department in a software company;
- 2006: Diploma in Computer Science on big scale co-occurrence analysis;
- 2007: Consultant for several SMEs in IT sector;
- 2008: Technical project management of the eAQUA project;
- 2011: PI and project manager of the eTRACES project;
- 2013: PhD in Digital Humanities on Text Reuse;
- 2014: Head of Early Career Research Group eTRAP at the University of Göttingen.

MY INTERESTS :)



WHAT IS TEXT REUSE?

WHAT DO YOU ASSOCIATE WITH TEXT REUSE AND INTERTEXTUALITY?



ASPECTS OF TEXT REUSE

EXPECTATIONS OF A COMPUTER SCIENTIST: OVERSIMPLIFICATION

It was the best of times, it
was the worst of times,
the age of wisdom, it w
age of foolishness, it
epoch of belief, it was
epoch of incredulity, it was
the season of Light, it was
the season of Darkness, it
It was the best of times,
it was the worst of times,
it wait was the age of
om, it was|

EXPECTATIONS OF A HUMANIST: OVERSIMPLIFICATION

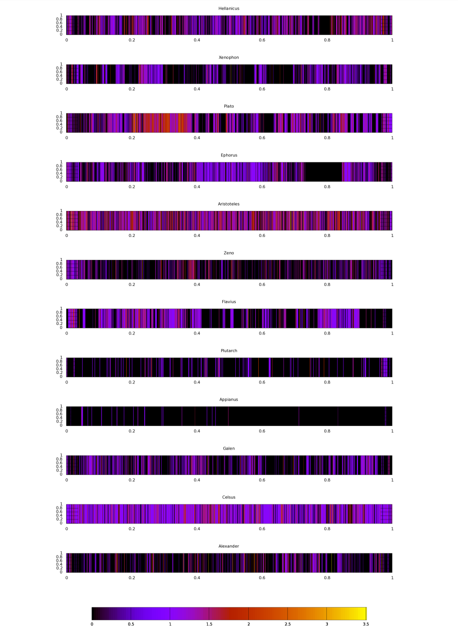
TEXT REUSE FOR HUMANITIES AND COMPUTER SCIENCE

Question:
Why is text reuse so relevant for Humanities and Computer Science?

Premise:
The amount of digitally available data is growing exponentially (Big Data).

- Humanities:
  - Lines of transmission 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



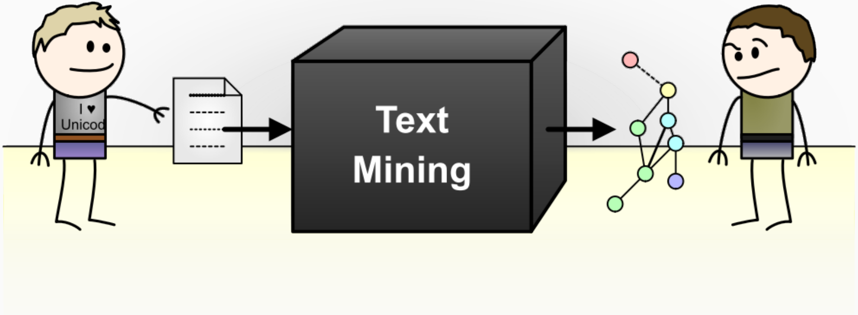
ACID FOR THE DIGITAL HUMANITIES

ACID PARADIGM

ACID for the Digital Humanities:

- **A**cceptance
- **C**omplexity
- **I**nteroperability
- **D**iversity

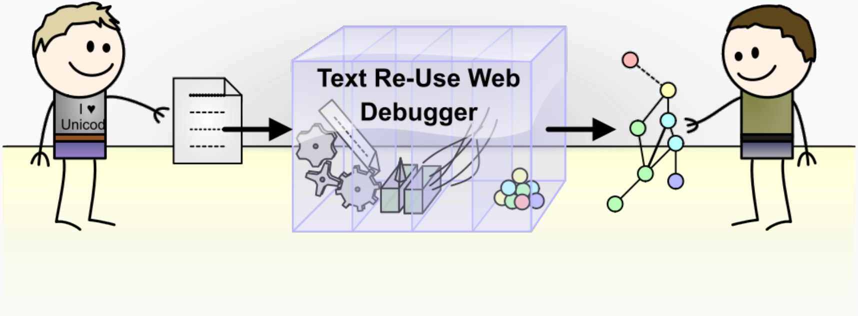
ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE I



ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE II

How to be accepted by humanists if text mining is a black box we can't look into?

ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE III



Transparency: How to provide user-friendly insights into complex mining techniques and machine learning?

BIG (HUMANITIES) DATA

WHAT IS BIG DATA?

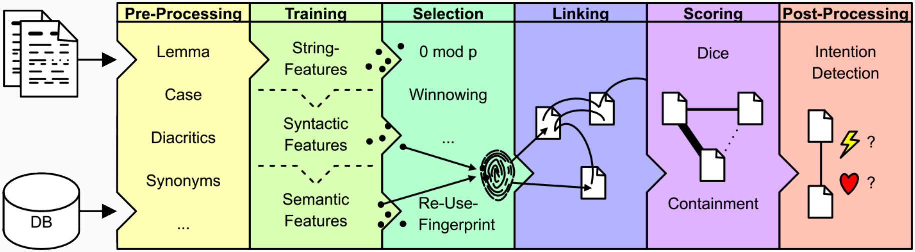
Ulrike Rieß (*Big Data bestimmt die IT-Welt*):

- Large amounts of data that can't be processed and analysed manually;
- Less structured data, e.g. in comparison to databases and data warehouse systems;
- Linked data between heterogeneous and distributed resources.

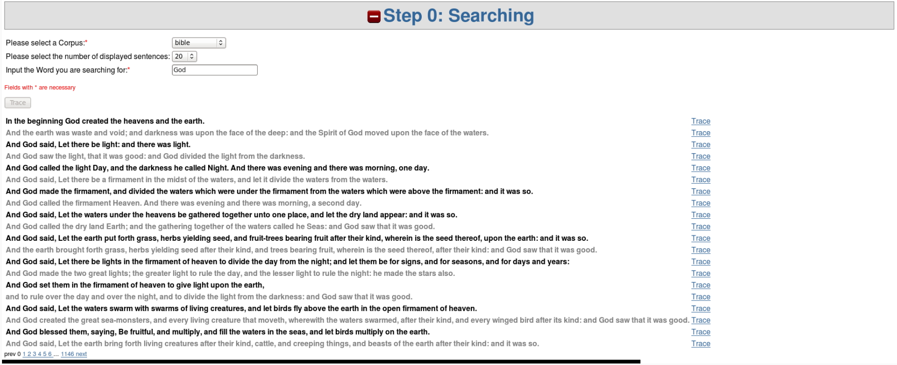
Information overload = large amounts of data (Big Data).
Information poverty = noisy, missing, fragmentary, oral data (Humanities Data).

COMPLEXITY

CURRENT APPROACH: TRACER



ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE IV



Step 0: Searching

Please select a Corpus:* bible

Please select the number of displayed sentences: 20

Input the Word you are searching for:* God

Fields with * are necessary

In the beginning God created the heavens and the earth. Trace

And the earth was waste and void; and darkness was upon the face of the deep: and the Spirit of God moved upon the face of the waters. Trace

And God said, Let there be light: and there was light. Trace

And God saw the light, that it was good: and God divided the light from the darkness. Trace

And God called the light Day, and the darkness he called Night. And there was evening and there was morning, one day. Trace

And God said, Let there be a firmament in the midst of the waters, and let it divide the waters from the waters. Trace

And God made the firmament, and divided the waters which were under the firmament from the waters which were above the firmament: and it was so. Trace

And God called the firmament Heaven. And there was evening and there was morning, a second day. Trace

And God said, Let the waters under the heavens be gathered together unto one place, and let the dry land appear: and it was so. Trace

And God called the dry land Earth; and the gathering together of the waters called he Seas: and God saw that it was good. Trace

And God said, Let the earth put forth grass, herbs yielding seed, and fruit-trees bearing fruit after their kind, wherein is the seed thereof, upon the earth: and it was so. Trace

And the earth brought forth grass, herbs yielding seed after their kind, and trees bearing fruit, wherein is the seed thereof, after their kind: and God saw that it was good. Trace

And God said, Let there be lights in the firmament of heaven to divide the day from the night; and let them be for signs, and for seasons, and for days and years: Trace

And God made the two great lights; the greater light to rule the day, and the lesser light to rule the night: he made the stars also. Trace

And God set them in the firmament of heaven to give light upon the earth, Trace

and to rule over the day and over the night, and to divide the light from the darkness: and God saw that it was good. Trace

And God said, Let the waters swarm with swarms of living creatures, and let birds fly above the earth in the open firmament of heaven. Trace

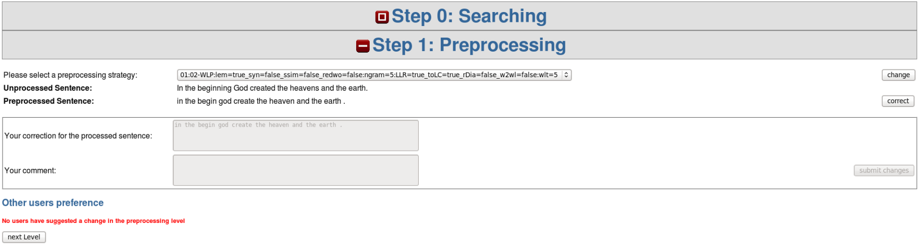
And God created the great sea-monsters, and every living creature that moveth, wherewith the waters swarmed, after their kind, and every winged bird after its kind: and God saw that it was good. Trace

And God blessed them, saying, Be fruitful, and multiply, and fill the waters in the seas, and let birds multiply on the earth. Trace

And God said, Let the earth bring forth living creatures after their kind, cattle, and creeping things, and beasts of the earth after their kind: and it was so. Trace

prev 0 1 2 3 4 5 6 ... 1146 next

ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE V



Step 0: Searching

Step 1: Preprocessing

Please select a preprocessing strategy: 01:02-WLP:lem=true_syn=false_ssim=false_redwo=false:ngram=5:LLR=true_toLC=true_rDia=false_w2wl=false:wlt=5

Unprocessed Sentence: In the beginning God created the heavens and the earth.

Preprocessed Sentence: in the begin god create the heaven and the earth .

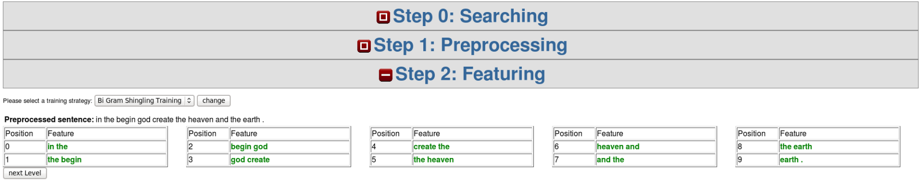
Your correction for the processed sentence: in the begin god create the heaven and the earth .

Your comment:

Other users preference

No users have suggested a change in the preprocessing level

ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE VI



Step 0: Searching

Step 1: Preprocessing

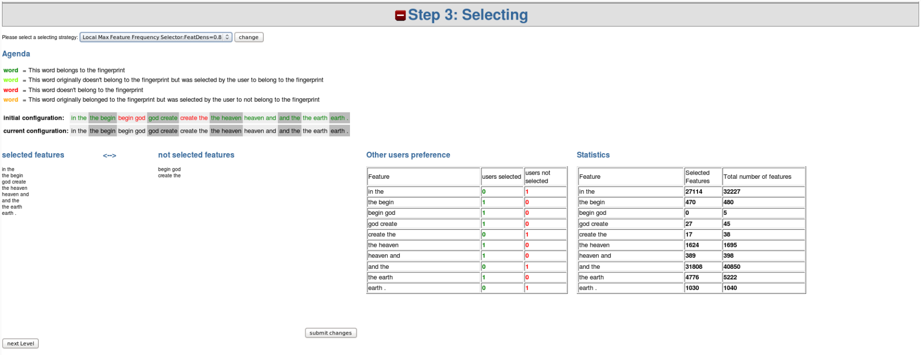
Step 2: Featuring

Please select a training strategy: Bi Gram Shingling Training

Preprocessed sentence: in the begin god create the heaven and the earth .

Position	Feature
0	in the
1	the begin
2	begin god
3	god create
4	create the
5	the heaven
6	heaven and
7	and the
8	the earth
9	earth .

ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE VII



Step 3: Selecting

Please select a selecting strategy: Local Max Feature Frequency Selector:FeatDens=0.8

Agenda

word = This word belongs to the fingerprint
word = This word originally doesn't belong to the fingerprint but was selected by the user to belong to the fingerprint
word = This word doesn't belong to the fingerprint
word = This word originally belonged to the fingerprint but was selected by the user to not belong to the fingerprint

initial configuration:	in the	the begin	begin god	god create	create the	the heaven	heaven and	and the	the earth	earth .
current configuration:	in the	the begin	begin god	god create	create the	the heaven	heaven and	and the	the earth	earth .

selected features <--> not selected features

selected features	not selected features
in the	begin god
the begin	create the
god create	
the heaven	
heaven and	
and the	
the earth	
earth .	

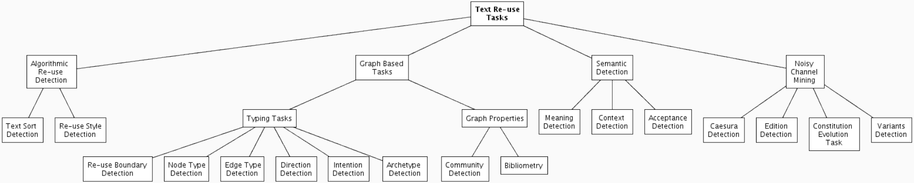
Other users preference

Feature	users selected	users not selected
in the	0	1
the begin	1	0
begin god	1	0
god create	1	0
create the	0	1
the heaven	1	0
heaven and	1	0
and the	0	1
the earth	1	0
earth .	0	1

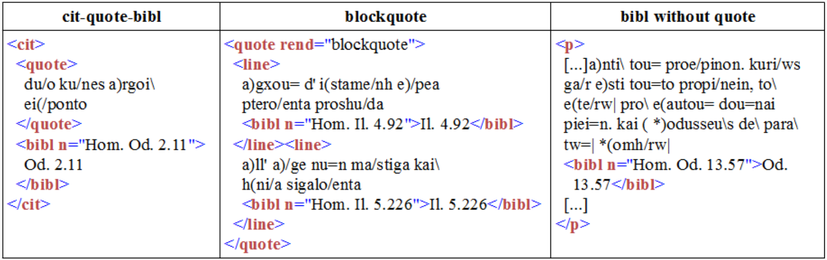
Statistics

Feature	Selected Features	Total number of features
in the	**27114**	**32227**
the begin	**470**	**480**
begin god	**0**	**5**
god create	**27**	**45**
create the	**17**	**38**
the heaven	**1624**	**1695**
heaven and	**389**	**398**
and the	**31808**	**40850**
the earth	**4776**	**5222**
earth .	**1030**	**1040**

ACID FOR THE DIGITAL HUMANITIES: COMPLEXITY

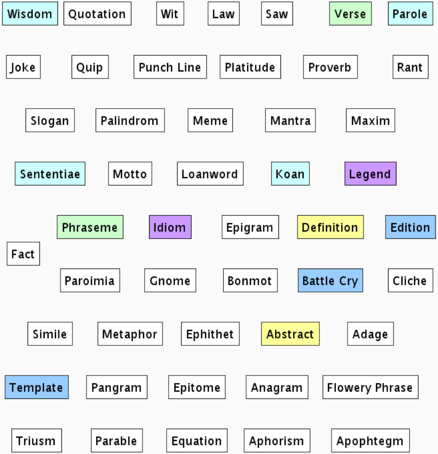


ACID FOR THE DIGITAL HUMANITIES: INTEROPERABILITY



cit-quote-bibl	blockquote	bibl without quote
<cit> <quote> du/o ku/nes a)rgoi\ ei(/ponto </quote> <bibl n="Hom. Od. 2.11"> Od. 2.11 </bibl> </cit>	<quote rend="blockquote"> <line> a)gxou= d' i(stame/nh e)/pea ptero/enta proshu/da <bibl n="Hom. Il. 4.92">Il. 4.92</bibl> </line><line> a)ll' a)/ge nu=n ma/stiga kai\ h(ni/a sigalo/enta <bibl n="Hom. Il. 5.226">Il. 5.226</bibl> </line> </quote>	<p> [...]a)nti\ tou= proe/pinon. kuri/ws ga/r e)sti tou=to propi/nein, to\ e(te/rw| pro\ e(autou= dou=nai piei=n. kai (*)odusseu\s de\ para\ tw=| *(omh/rw| <bibl n="Hom. Od. 13.57">Od. 13.57</bibl> [...] </p>

ACID FOR THE DIGITAL HUMANITIES: DIVERSITY (REUSE TYPES)



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

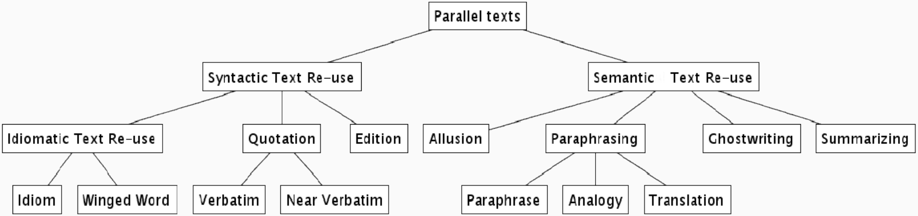
Simile Metaphor Ephithet Abstract Adage

Template Pangram Epitome Anagram Flowery Phrase

Triusm Parable Equation Aphorism Apophtegm

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

ACID FOR THE DIGITAL HUMANITIES: DIVERSITY (REUSE STYLES)



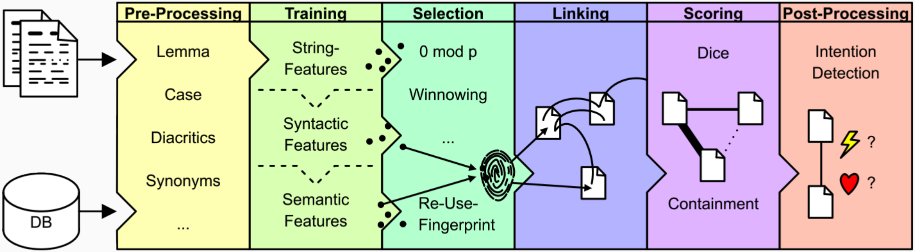
LANGUAGE MODEL

KEY PROBLEM

Question:

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

OUTLINE


Pre-Processing
Training
Selection
Linking
Scoring
Post-Processing
Lemma
Case
Diacritics
Synonyms
...
String-Features
Syntactic Features
Semantic Features
0 mod p
Winnowing
...
Re-Use-Fingerprint
Dice
Containment
Intention Detection
?
?
DB

Finito!

CONTACT

Team

Marco Büchler, Greta Franzini and Emily Franzini.

Visit us

🌐 `http://www.etrap.eu`

✉ `contact@etrap.eu`

TRACER
TEXT REUSE DETECTION MACHINE

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

SPONSORED BY THE

Federal Ministry
of Education
and Research

LICENCE