INTRODUCTION TO HISTORICAL TEXT REUSE DETECTION

eTRAP: Electronic Text Reuse Acquisition Project

Marco Büchler, Emily Franzini, Greta Franzini & Maria Moritz

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

WHO AM I?

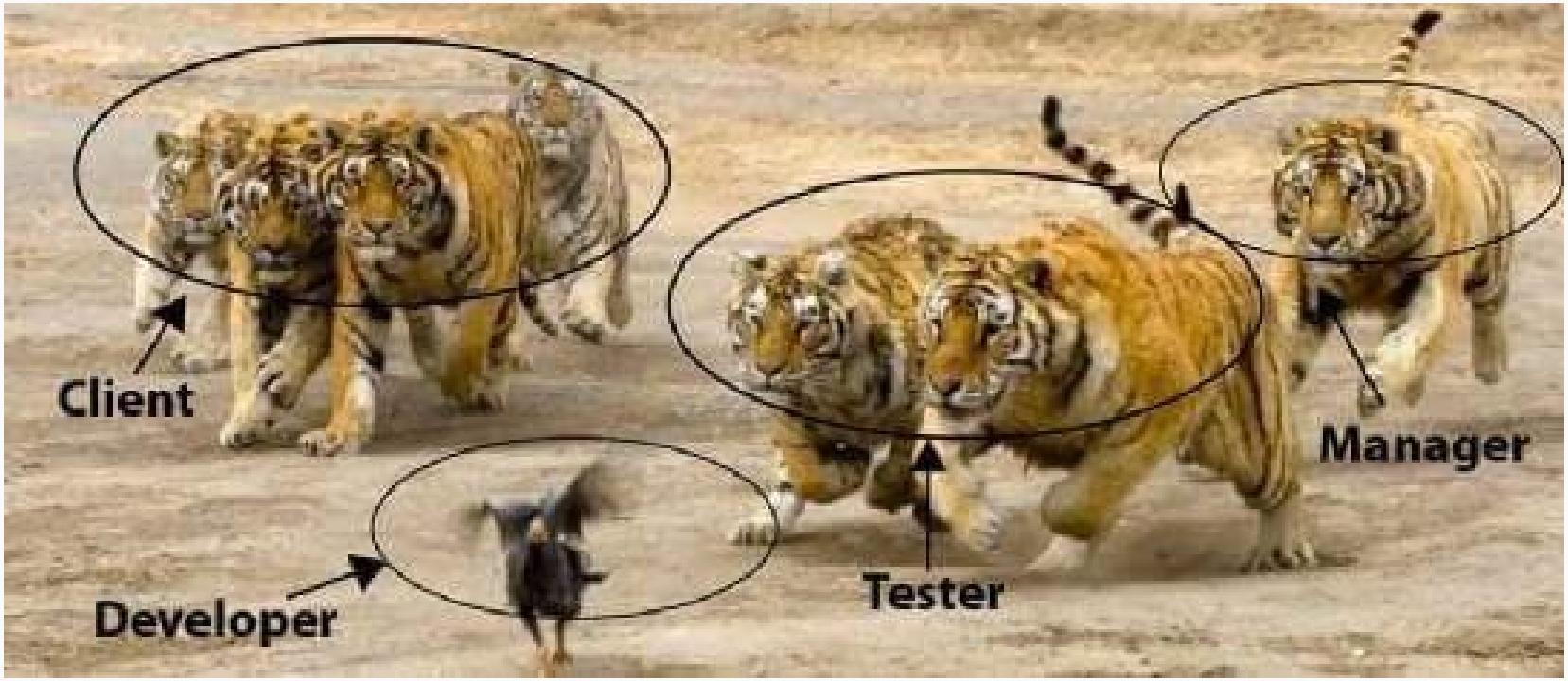
WHO AM I?

- 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.

WHO IS ETRAP?

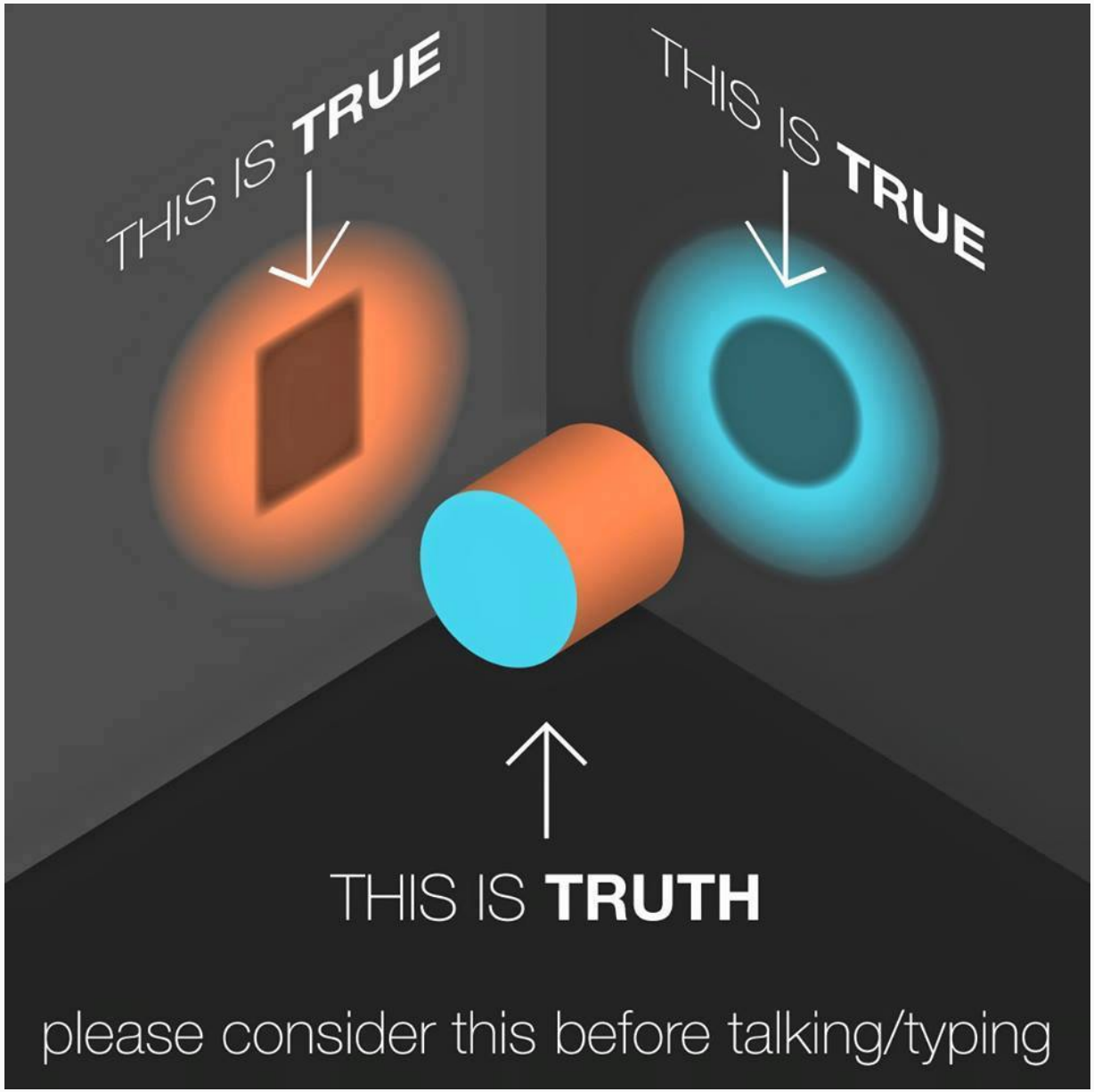
- Interdisciplinary research group
- Funded by the German Ministry of Education and Research
- Project duration: 03/2015 - 02/2019
- Team page: `http://www.etrap.eu/team/`

MY INTERESTS :)



WHAT IS TEXT REUSE?

WHAT DO YOU ASSOCIATE WITH TEXT REUSE AND INTERTEXTUALITY?



TEXT REUSE FOR HUMANITIES AND COMPUTER SCIENCE

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

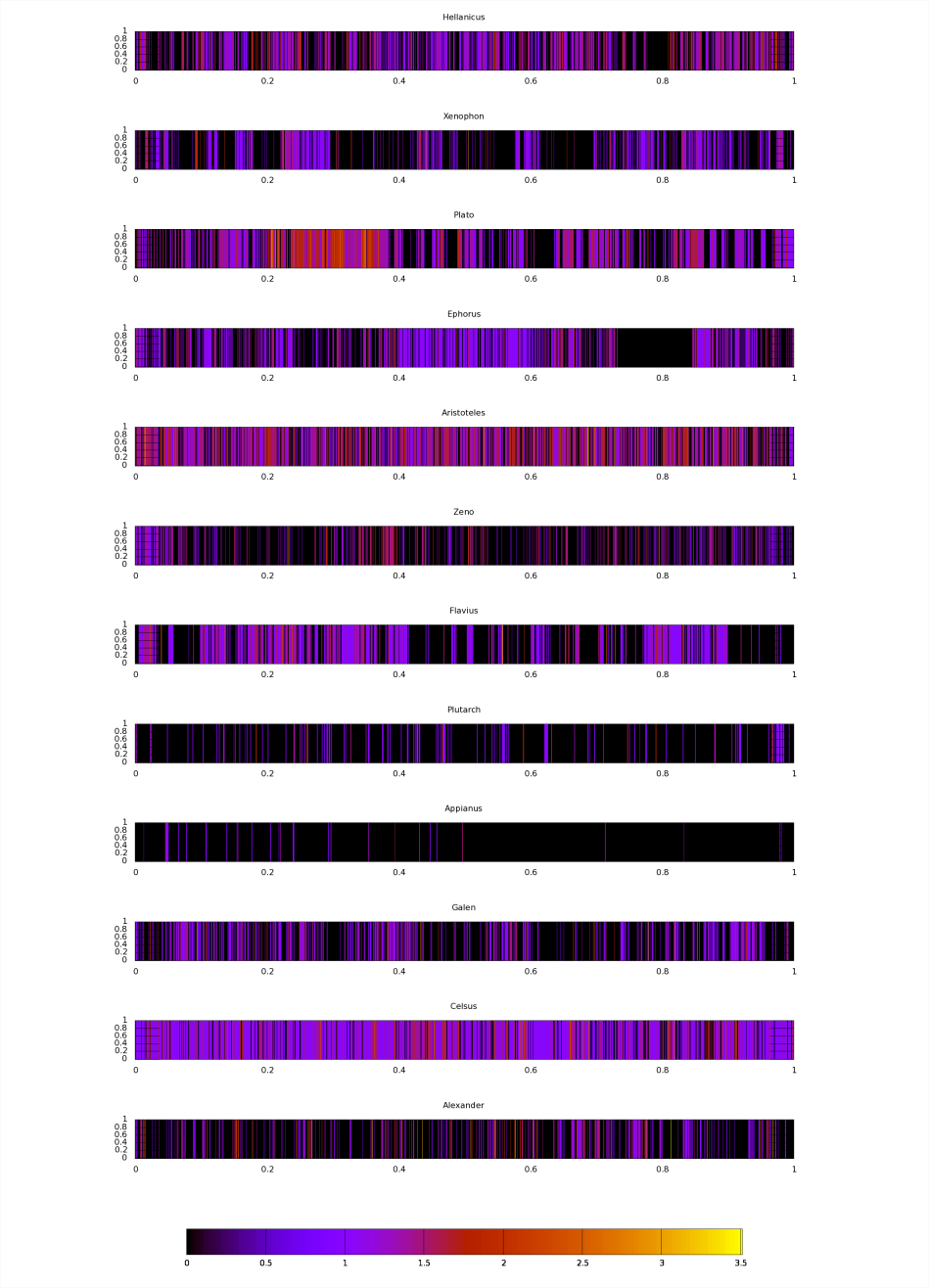
Premise:
The amount of digitally available data is growing exponentially (Big Data).

- 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.

EXPECTATIONS OF A COMPUTER SCIENTIST: OVERSIMPLIFICATION

EXPECTATIONS OF A HUMANIST: OVERSIMPLIFICATION

OPPORTUNITY: TEMPERATURE MAP



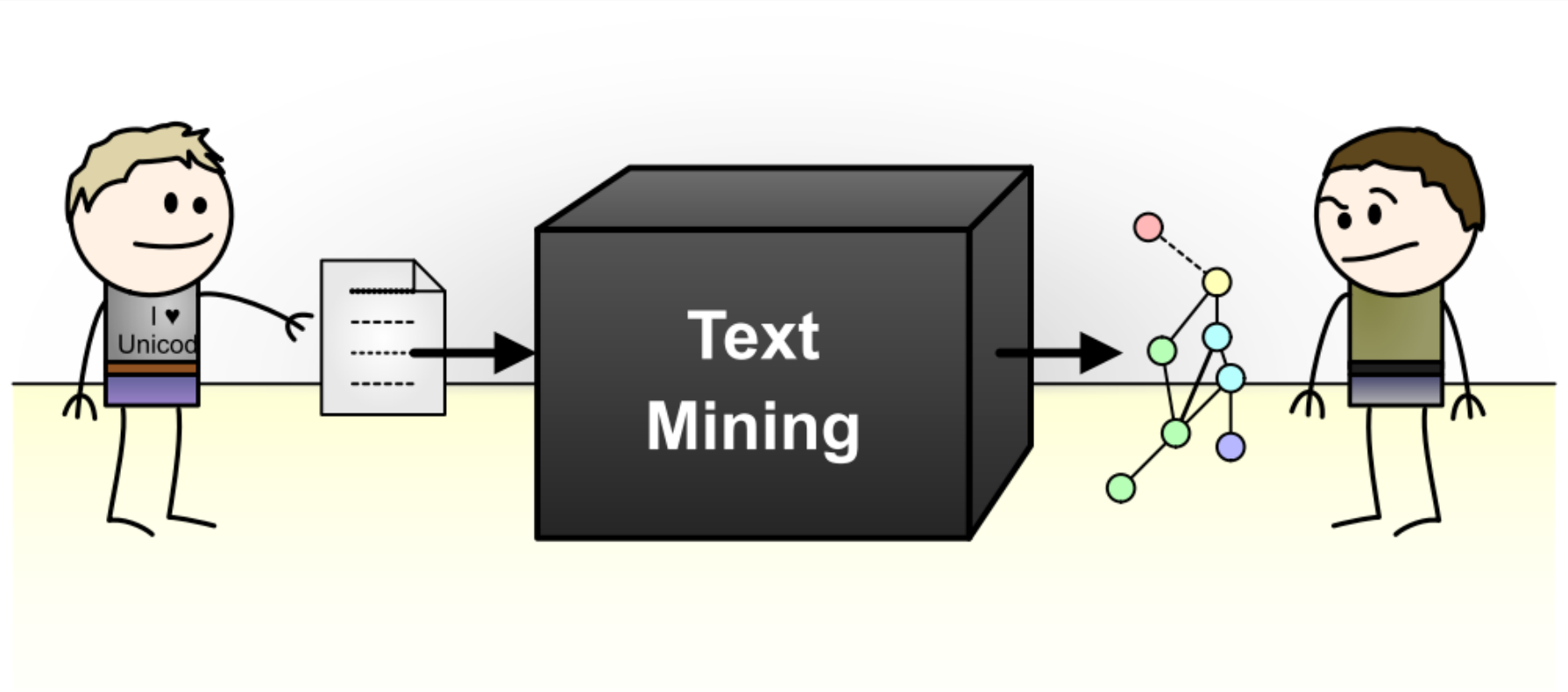
ACID FOR THE DIGITAL HUMANITIES

ACID PARADIGM

ACID for the Digital Humanities:

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

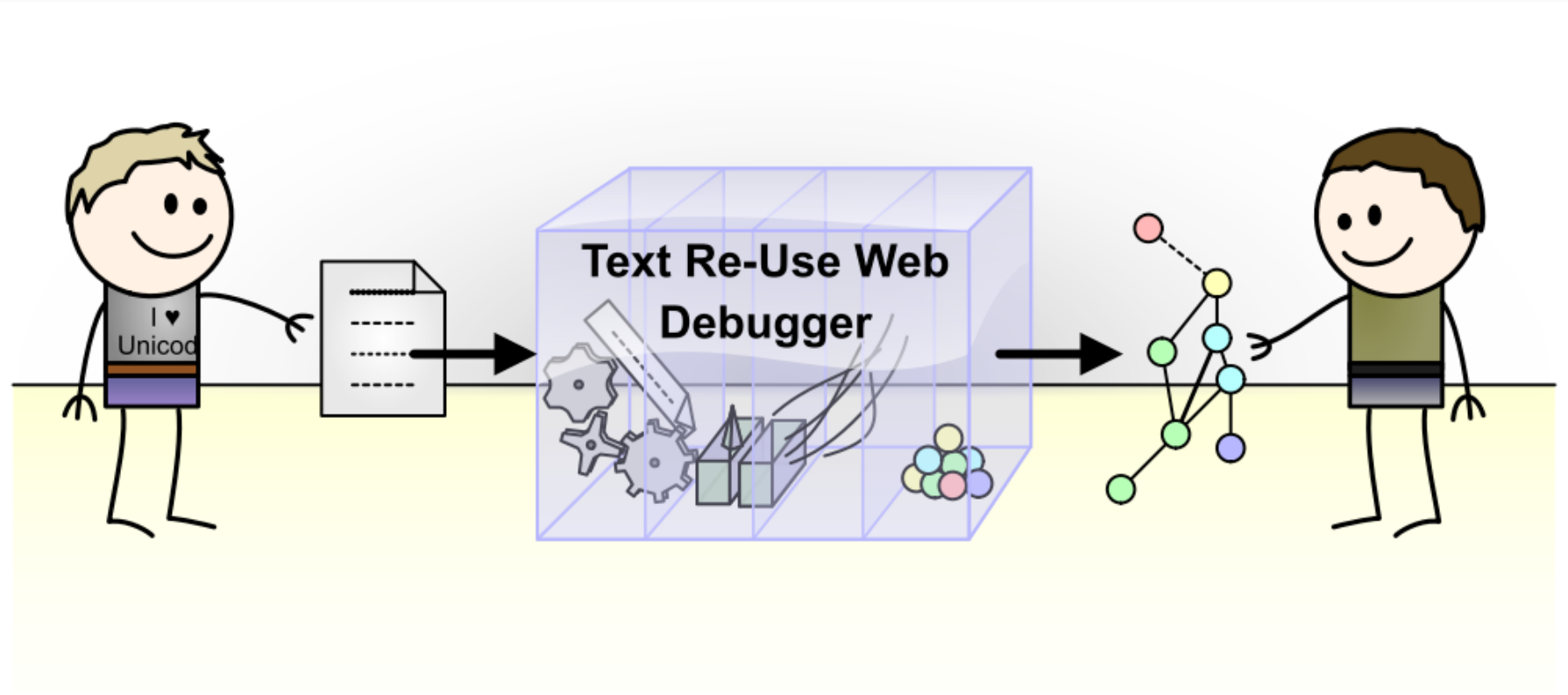
ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE I



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?

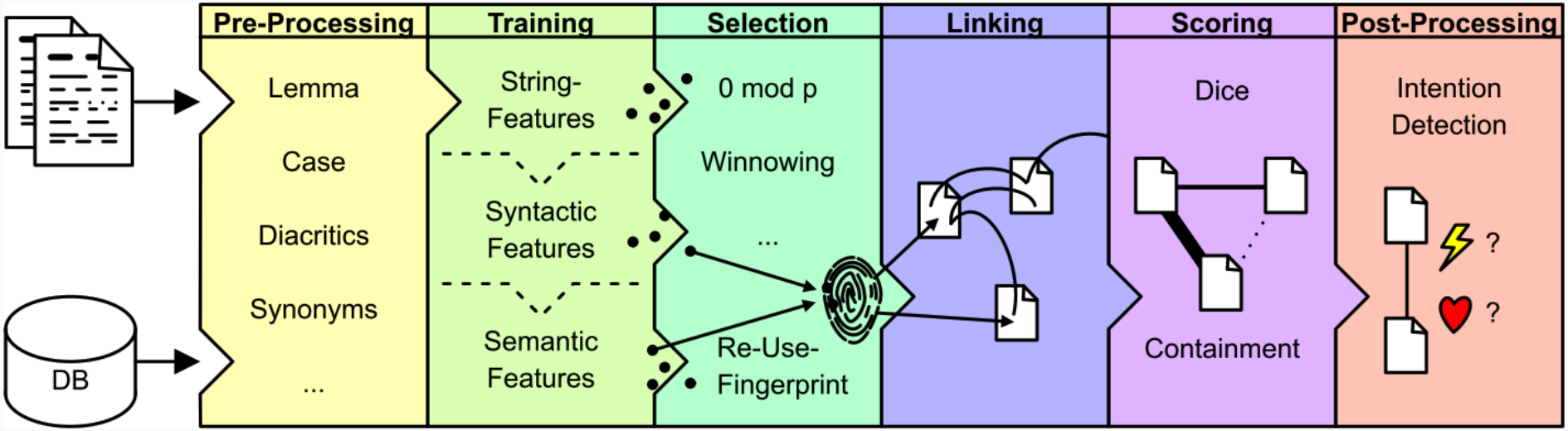
ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE III



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

BIG (HUMANITIES) DATA

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

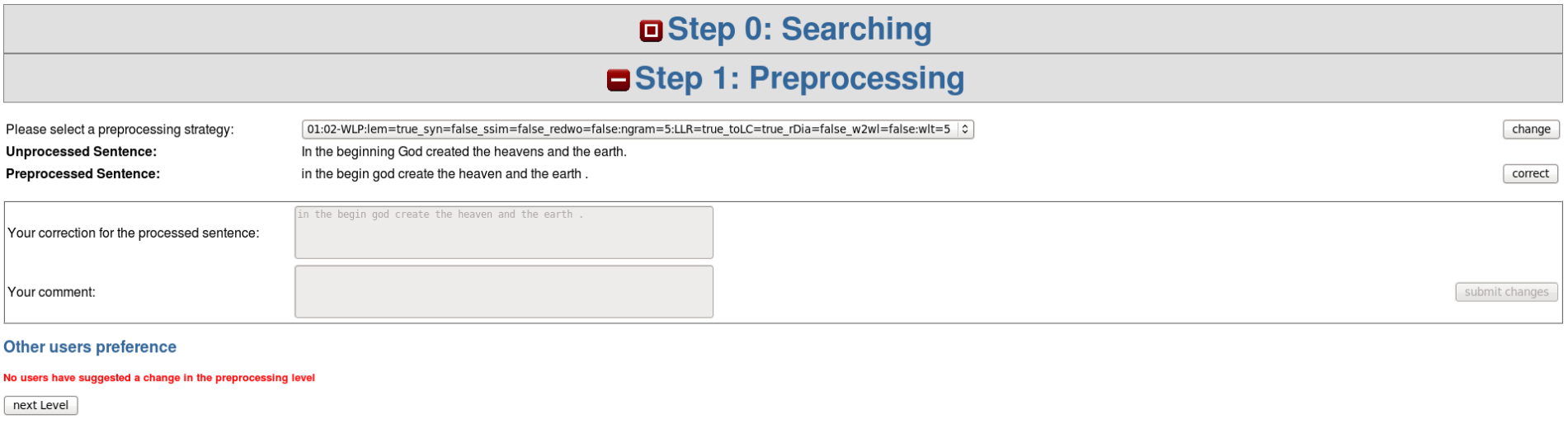
Fields with * are necessary

Trace

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



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 change

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

Position	Feature
0	in the
1	the begin

Position	Feature
2	begin god
3	god create

Position	Feature
4	create the
5	the heaven

Position	Feature
6	heaven and
7	and the

Position	Feature
8	the earth
9	earth .

next Level

ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE VII

Step 3: Selecting

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

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

in the
the begin
god create
the heaven
heaven and
and the
the earth
earth .

begin god
create the

submit changes

next Level

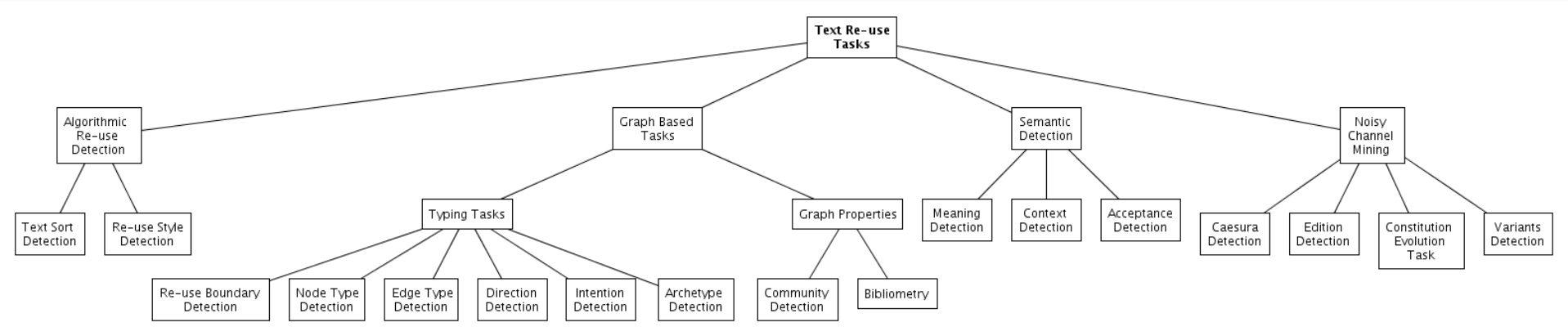
Other users preference

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

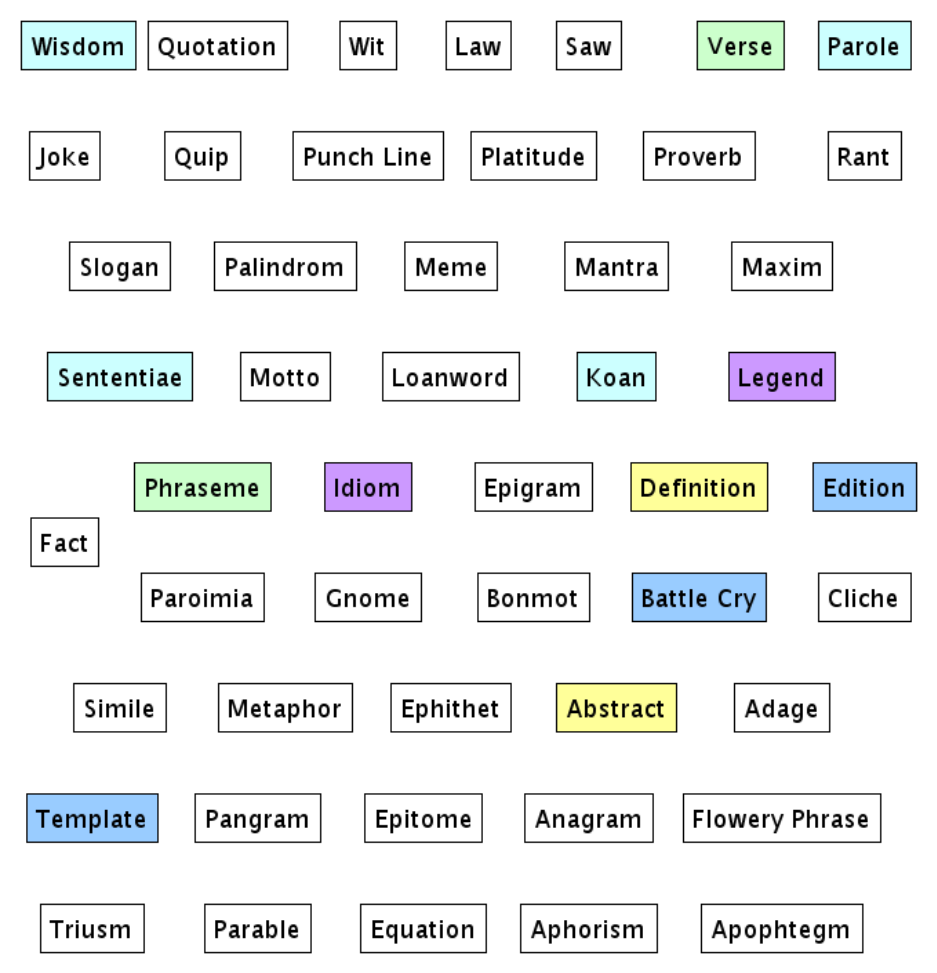
ACID FOR THE DIGITAL HUMANITIES: COMPLEXITY



ACID FOR THE DIGITAL HUMANITIES: INTEROPERABILITY

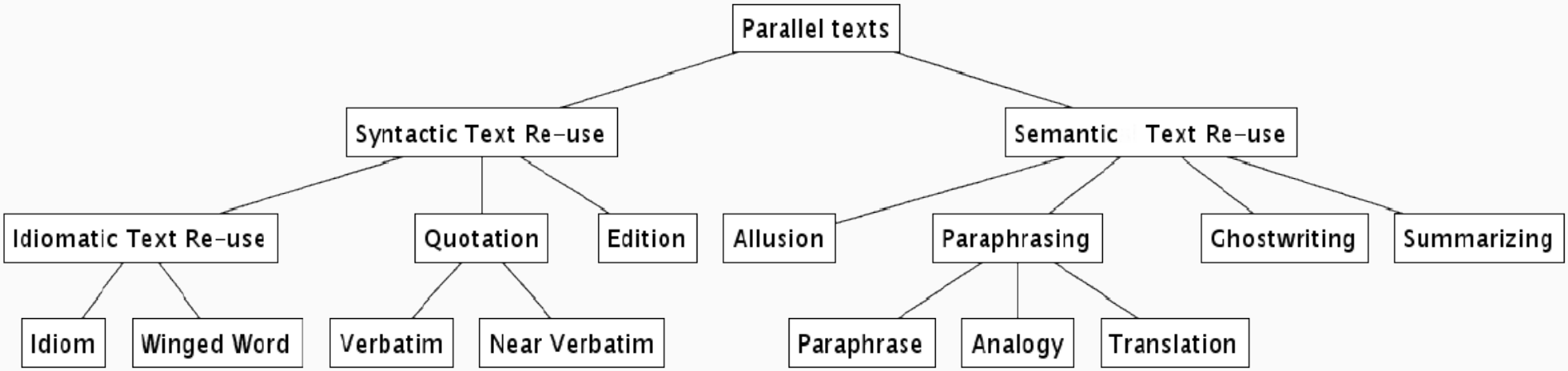
cit-quote-bibl	blockquote	bibl without quote
<cit> <quote> du/o ku/nes a)rgoi\ ei(/ponto </quote> <bibl n="Hom. Od. 2.11"> Od. 2.11 </bibl> </cit>	<quote rend="blockquote"> <line> a)gxou= d' i(stame/nh e)/pea ptero/enta proshu/da <bibl n="Hom. Il. 4.92">Il. 4.92</bibl> </line><line> a)ll' a)/ge nu=n ma/stiga kai\ h(ni/a sigalo/enta <bibl n="Hom. Il. 5.226">Il. 5.226</bibl> </line> </quote>	 [...]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> [...]

ACID FOR THE DIGITAL HUMANITIES: DIVERSITY (REUSE TYPES)



- **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)



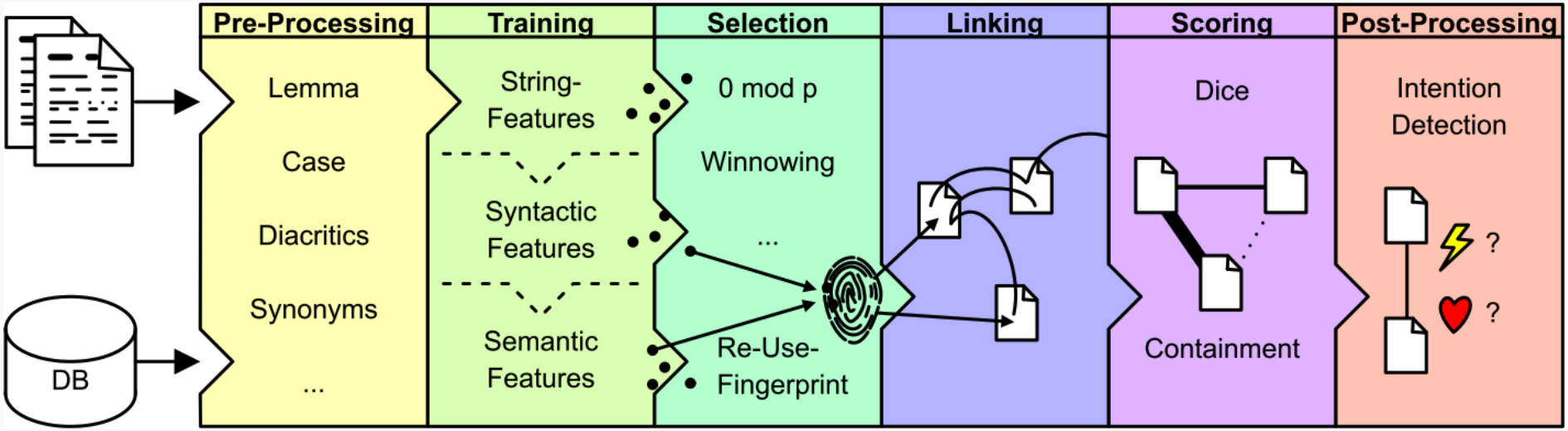
LANGUAGE MODEL

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
...
DB
String-Features
Syntactic Features
Semantic Features
0 mod p
Winnowing
...
Re-Use-Fingerprint
Dice
Containment
Intention Detection
?
?

TRACER MACHINE: A FRAMEWORK FOR THE DETECTION OF HISTORICAL TEXT REUSE

- **Link:** `http://vcs.etrap.eu/tracer-framework/tracer.git`
- **Planned trainings:**
  - **AIUCD 2017** (01/2017): pre-conference workshop, Rome, Italy
  - **DATECH 2017** (05/2017): pre-conference workshop, Göttingen, Germany
  - Three more trainings are still pending until August 2017

CURRENT CHALLENGES

Text reuse challenges:

- Detecting text reuse across languages;
- Detecting text reuse at scale;
- Detecting looser forms of text reuse, e.g. allusion;
- Diversity of historical texts: language evolution, copy errors, etc.

TEXT REUSE AT SCALE

TEXT REUSE AT SCALE

Premise: To trace a motif through space and time you need big data.

Table 1: Google Custom Search vs. Apache Lucene.

Approach	PROs	CONs
Google Custom Search (online)	-Huge data -API	-Not free -Limited result-set (top 100)
Apache Lucene (offline)	-Free -Control over search parameters	-Download & index all docs

Current research on online vs. offline approaches for text reuse detection (German idioms) at scale (Solhdoust, 2016):

- **Google Custom Search** (online): searching in Google Books and the web.
- **Apache Lucene** (offline): searching in Deutsches Textarchiv, zeno.org, Project Gutenberg.

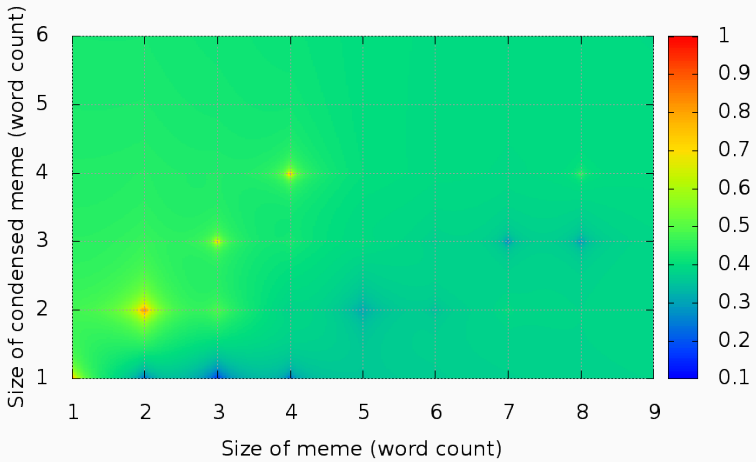


Figure 1: Similarity plot of idiom/meme samples using Google's Custom Search engine (online).

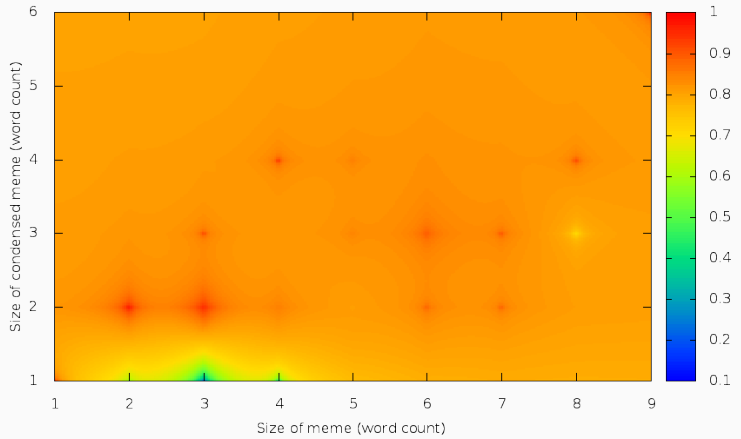


Figure 2: Similarity plot of idiom/meme samples using Apache Lucene (offline).

TEXT REUSE AT SCALE: PARALLELISATION

Current research on Parallelisations (Kirill Bulert, BA topic):

- **User groups:**
  - Humanists with a laptop
  - Small working groups with several desktop computers
  - Groups with access to a computational multi-core server
  - Groups with access to cloud or grid resources
- **Approaches**:
  - laptop
  - Desktop computer: multithreading and -processoring by shared memory systems
  - MPI
  - Shared memory multi-core servers
  - Cloud and grid systems

Parallelisation is not the solution for scaling; improving the algorithms.

RETHINKING THE PRIMITIVES

MOTIFS AS TEXT REUSE FEATURES

Motif: *"1. A minimal thematic unit"* (Prince, 2003, p. 55), a measurable primitive.

Measurable primitives from an interdisciplinary standpoint:

- **Literature**: tracing MOTIFS
- **Cultural Studies**: tracing MEMES
- **Linguistics**: tracing PATTERNS
- **Computer Science**: tracing FEATURES
- **Forensics**: tracing MINUTIAE

DIGITAL BREADCRUMBS OF BROTHERS GRIMM

Project began in October 2015.

Seven editions of *Kinder- und Hausmärchen*: 1812, 1819, 1837, 1840, 1843, 1850, 1857.

Changes in:

- Size: from 156 to 211.
- Content: gruesome to mild.
- Style: Jacob scholarly, Wilhelm figurative.
- Language: Variants, diachronic evolution.

TALES

Tales selected for investigation:

- *Snow White* (AT 709);
- *Puss in Boots* (AT 545B);
- *The Fisherman and his Wife* (AT 555).

EXAMPLE CASE STUDY: SNOW WHITE

Q: How to computationally detect a motif despite its variants?

For example:

- DE [Grimm][1]: ***Schneewittchen und die sieben Zwerge***
- EN [Briggs][2]: ***Snow White and the three robbers***
- IT [Calvino][3]: ***Bella Venezia e i dodici ladroni***
- SQ [von Hahn][4]: ***Schneewittchen und die vierzig Drachen***
- RU [Pushkin][5]: Сказка о мертвой царевне и о семи богатырях
- ...

A: We need to combine Aarne-Thompson (Uther) and Propp approaches. That is, finding the balance between describing a motif (AT specificity) and leaving enough space for variations (Propp typological unity and sequence of events).

Collections and Languages

- **Identified versions**: Albanian, Algerian, Appalachian, Armenian, Breton, Celtic (Scottish), Egyptian, English, Finnish, German, Greek, Italian, Moroccan, Russian, Spanish.
- **Potential others**: African, Australian, Basque, Caribbean, Catalan, Caucasian, Chinese, Danish, Dutch, Estonian, French, Friesian, Georgian, Hawaiian, Icelandic, Indian, Indian-American, Israeli, Japanese, Korean, Latvian, Lithuanian, Macedonian, Mexican, Nepalese, New Zealand, Norwegian, Paraguayan, Persian, Polish, Portuguese, Punjabi, Romansh, Rumanian, Siberian, South-American, Sri Lankan, Swedish, Swiss, Tibetan, Turkish, Uzbek, Yiddish.
- **Does not appear in**: Ladin.

DATA COLLECTION AND CURATION

Tasks: Verify presence of motif in different collections and record its ”base form” as text reuse training data.

ISO Language Codes https://www.loc.gov/standards/iso639-2/php/code_list.php	GER						RUS		ITA	GLA		ARM	ENG			ARA		
Aarne-Thompson: 709	Grimm_1819 VIAF: 187449723	Grimm_1837 VIAF: 187449723	Grimm_1840 VIAF: 187449723	Grimm_1843 VIAF: 187449723	Grimm_1850 VIAF: 187449723	Grimm_1857 VIAF: 187449723	Pushkin_1833 VIAF: 312344013	Tsvetaeva_1911 VIAF: 185088476	Calvino_1956 VIAF: 181208131	Jacobs_1892 VIAF: 315397813	Bruford_1994 VIAF12471835	Hoogasian-Villa_1966 VIAF: 186329063	Campbell_1958 VIAF: 25969242	Taylor_1823 VIAF: 59071527	Briggs_1970 VIAF: 46803237	El-Shamy_1999 VIAF: 276573319	El Koudia_2003 VIAF: 5206198	Jason_1977 VIAF 9970253
D1300-D1379. Magic objects effect changes in persons																		
D1364. Object causes magic sleep	x	x	x	x	x	x	x	null	x	x	x	x	x	x	x	x	x	x
D1364.4. Fruit causes magic sleep	x	x	x	x	x	x	x	null	null	null	null	null	x	x	x	null	null	null
D1364.4.1. Apple causes magic sleep	x	x	x	x	x	x	x	null	null	null	null	null	x	x	x	null	null	null
D1364.9. Comb causes magic sleep	x	x	x	x	x	x	null	null	null	null	null	null	x	x	null	null	null	null
D1364.13. Cloth causes magic sleep	x	x	x	x	x	x	null	null	null	null	null	null	null	x	null	null	null	null
D1364.13.1. Lace causes magic sleep	x	x	x	x	x	x	null	null	null	null	null	null	null	x	null	null	null	null

Figure 3: Microsoft Excel matrix of motifs. Left column lists AT motifs in *Snow White* (AT 709); top row lists languages and collections covered.

Q400-Q599. Kinds of punishment	
Q411. Death as punishment	zu todt tanzen
Q414. Punishment: burning alive	glühende Pantoffeln, zu todt tanzen
Q414.4. Punishment: dancing to death in red-hot shoes	eiserne Pantoffeln, Feuer, glühend, anziehen, tanzen, Füße jämmerlich verbrannt, nicht aufhören, zu todt tanzen

Figure 4: Grimm motifs reduced to keywords.

MINING THE CHANGES

APPROACH

Inspired by Shannon's noisy-channel (Shannon, 1949) & Kolmogorov Complexity (Li and Vitáni, 2008), we study Greek and Latin text reuse to understand how text is transferred.

- We identify operations that characterize word changes.
- We show how linguistic resources can help detecting non-literal reuse.
- We complement the automated approach with a manual analysis.

DATASETS - ANCIENT GREEK AND LATIN DATASET

"Salvation for the Rich"
Clement of Alexandria
Christian theologian, 2nd cent.

- Known for his retelling of biblical excerpts
- Reuse annotated upfront by Biblindex team (Mellerin, 2014; Mellerin, 2016)
- We obtain 199 verse-reuse-pairs
- Pointing to 15 Bible books

Extracts from 12 works & 2 collections
Bernard of Clairvaux
French abbot, 12th cent.

- Known for his influence to the Cistercian order and his work in biblical studies
- Reuse extracted upfront by Biblindex team (Mellerin, 2014; Mellerin, 2016)
- We obtain 162 verse-reuse-pairs
- Pointing to 31 Bible books

TRANFORMATION OPERATIONS

Table 2: Operation list for the automated approach

operation	description	example
NOP(reuse_word, orig_word)	Original and reuse word are equal.	*NOP(maledictus,maledictus)*
upper(reuse_word, orig_word)	Word is lowercase in reuse and uppercase in original.	*upper(kai,Kai)* - in Greek
lower(reuse_word, orig_word)	Word is uppercase in reuse and lowercase in original.	*lower(Gloriam,gloriam)*
lem(reuse_word, orig_word)	Lemmatization leads to equality of reuse and original.	*lem(penetrat,penetrabit)*
repl_syn(reuse_word, orig_word)	Reuse word replaced with a synonym to match original word.	*repl_syn(magnificavit,glorificavit)*
repl_hyper(reuse_word, orig_word)	Word in bible verse is a hyperonym of the reused word.	*hyper(cupit,habens)*
repl_hypo(reuse_word, orig_word)	Word in bible verse is a hyponym of the reused word.	*hypo(dederit,tollet)*
repl_co-hypo(reuse_word, orig_word)	Reused word and original have the same hyperonym.	*repl_co-hypo(magnificavit,fecit)*
NOPmorph(reuse_tags, orig_tags)	Case or PoS did not change between reused and original word.	*NOPmorph(na,na)*
repl_pos(reuse_tag, orig_tag)	Reuse and original contain the same cognate, but PoS changed.	*repl_pos(n,a)*
repl_case(reuse_tag, orig_tag)	Reuse and original have the same cognate, but the case changed	*repl_case(g,d)* - cases genitive, dative
lemma_missing(reuse_word, orig_word)	Lemma unknown for reuse or original word	*lemma_missing(tentari, inlectus)*
no_rel_found(reuse_wword, orig_word)	Relation for reuse or original word not found in AGWN	*no_rel_found(gloria,arguitur)*

LITERAL SHARE OF THE REUSE

What is the extent of non-literal reuse in our datasets?

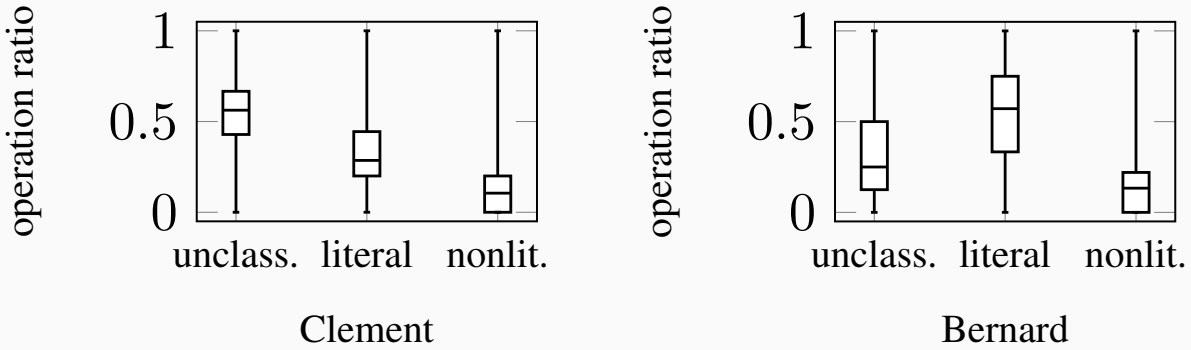


Figure 5: Ratios of operations in reuse instances. literal: NOP, lem, lower, etc.; nonlit: syn, hyper, etc.

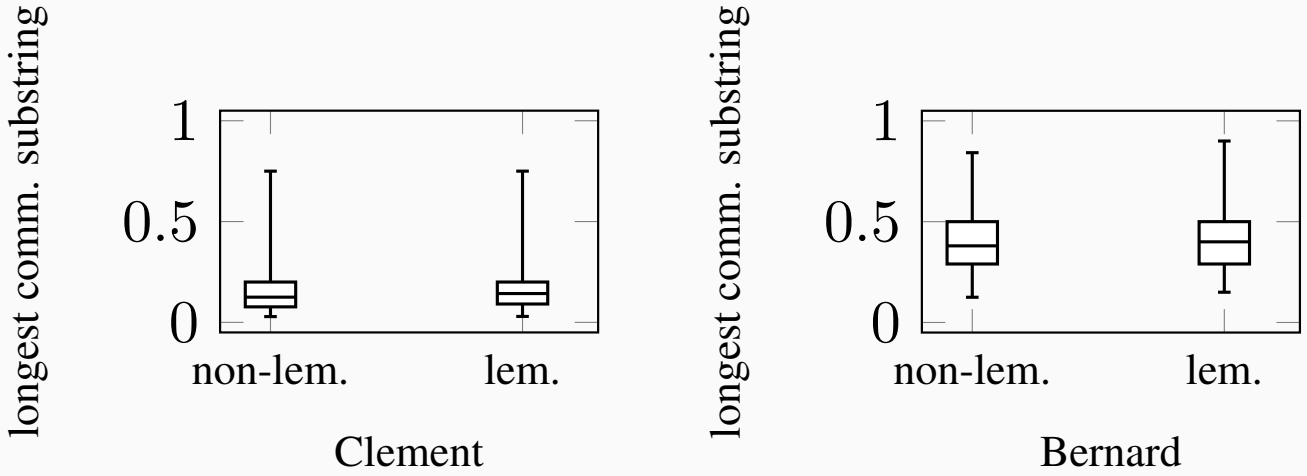


Figure 6: Ratios of literal overlap between reuse instances and originals

How is the non-literally reused text modified in our datasets?
How can linguistic resources support the discovery of non-literal reuse?

Table 3: Absolute numbers of operations identified automatically

	literal				nonliteral				unclassified		
	NOP	upper	lower	lem	syn	hyper	hypo	co-hypo	no_rel_found	lem_missing	total
Greek	337	6	0	356	153	20	14	101	563	639	2189
Latin	587	0	44	102	60	14	28	68	347	85	1335

SOME OUTLINES

SPECIAL ISSUE ON INTERTEXTUALITY COMING OUT SOON

Special Issue Call for Contribution

Computer-Aided Processing of Intertextuality in Ancient Languages

http://jdmdh.episciences.org/page/call-for-contribution-special-issue

Questions & comments to: **mbuechler@gcdh.de**

DATECH 2017: A DIGITAL HUMANITIES CONFERENCE ON DIGITAL TRANSFORMATION

- **Link:** `http://ddays.digitisation.eu/datech-2016/`
- **Topics:**
  - Improved OCR and special OCR techniques for historical documents
  - Innovative views and tools for the exploitation of digital content by both experts and non-expert communities in the humanities
  - Advanced tools for a higher productivity and quality in the creation of useful digital content
  - Improved treatment of historical languages (diachronic language development) and multilingualism
  - New mining techniques on historical text collections (addressing, e.g., historical text re-use, or person and event detection).

CONTACT

Team

Marco Büchler, Greta Franzini, Emily Franzini and Maria Moritz.

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

LICENCE