ETRAP: ELECTRONIC TEXT REUSE ACQUISITION PROJECT

On the primitives and the process of text reuse

Emily Franzini & Marco Büchler (with contributions from Greta Franzini & Maria Moritz)

eTRAP
Electronic Text Reuse Acquisition Project

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

WHO ARE WE?

WHO AM I?

- 2008-2011: BA Latin & Ancient Greek at University College London
- 2011-2012: MSc Management Science & Innovation at University College London
- 2012-2013: Liaison Officer and Administration for the preservation of cultural assets at FAI
- 2013-2014: Research Associate at University of Leipzig (Chair for Digital Humanities)
- 2014-2016: Research Associate at University of Göttingen (Digital Humanities in Dept. Computer Science)

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)

Early Career Research Group funded by German Ministry of Education & Research (BMBF).

Budget: €1.6M.
Duration: January 2015 - February 2019. Research since October 2015.
Team: 4 core staff; 5-9 research & student assistants and several BA and MA thesis students.

- Interdisciplinary: Classics, Computer Science, German Philology, Mathematics, Philosophy, Software Engineering. Cognitive Literature
- International: Recently from nine different nationalities.

WHAT IS TEXT REUSE?

Text reuse = spoken and written repetition of text across time and space.

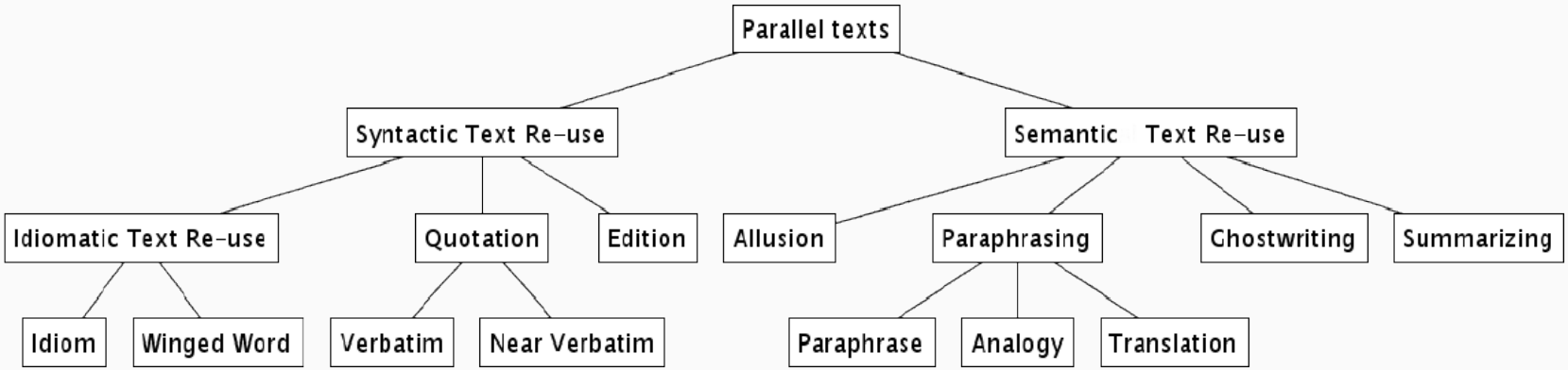


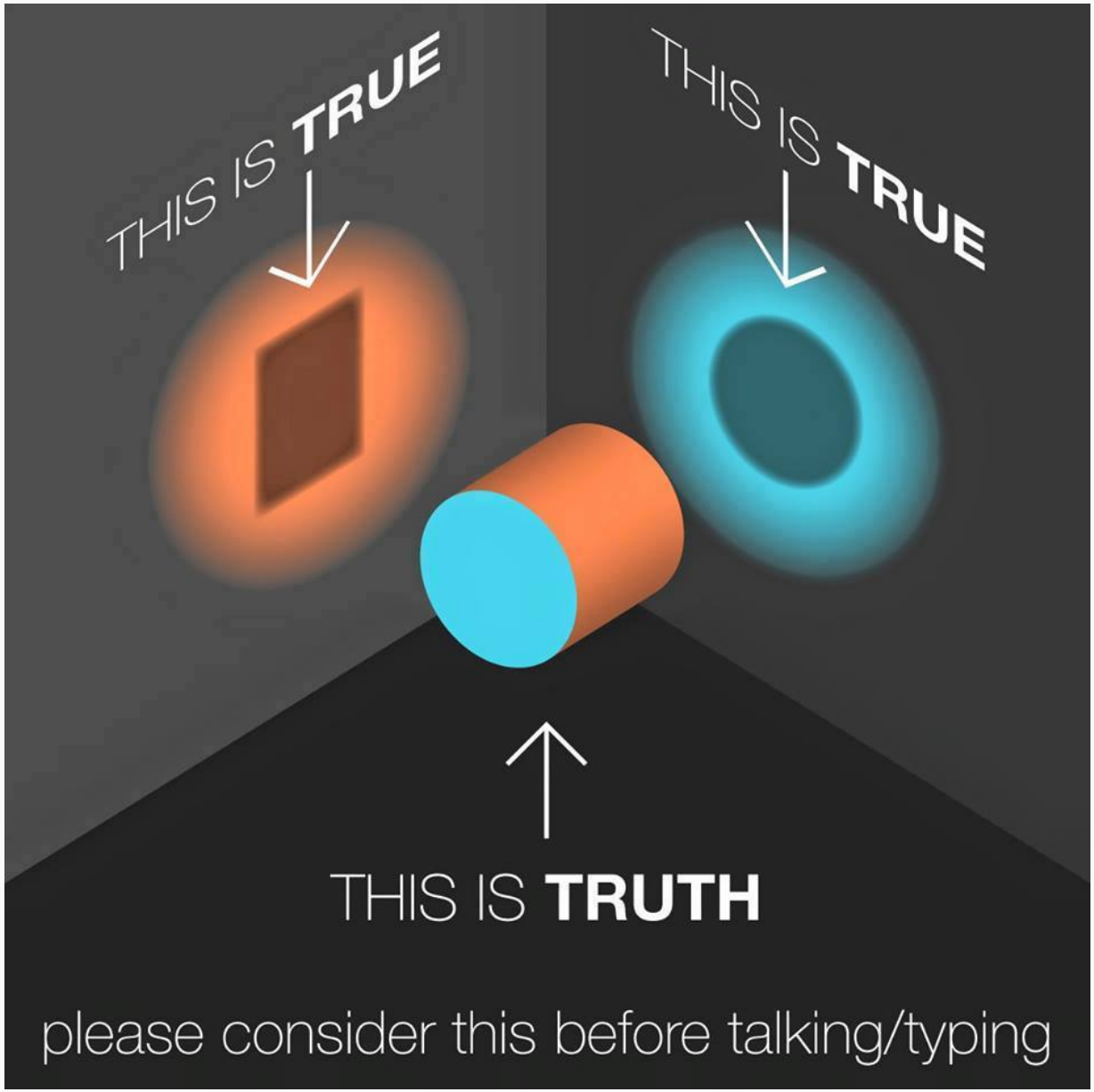
Figure 1: Text reuse styles [Author: Marco Büchler].

TEXT AS A MOSAIC OF QUOTATIONS

"[...] a text is [...] a multidimensional space in which a variety of writings, none of them original, blend and clash. The text is a tissue of quotations drawn from the innumerable centres of culture... the writer can only imitate a gesture that is always anterior, never original. His only power is to mix writings [...]." (Barthes, 1977, pp. 146-47)

"[...] any text is constructed as a mosaic of quotations [...]." (Kristeva, 1980, p.66)

WHAT DO YOU ASSOCIATE WITH TEXT REUSE AND INTERTEXTUALITY?



EXPECTATIONS OF A COMPUTER SCIENTIST: OVERSIMPLIFICATION

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.

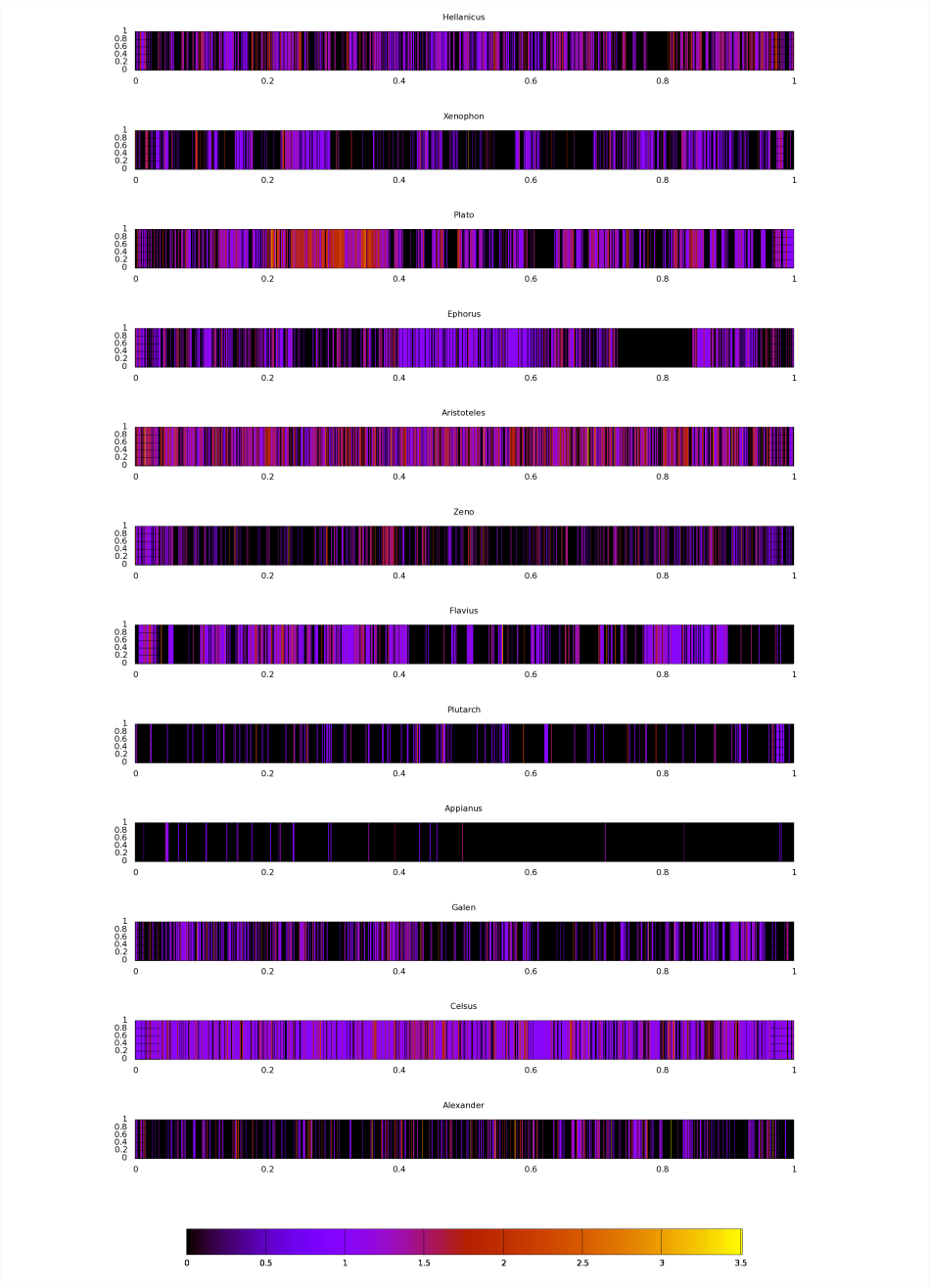
BIG (HUMANITIES) DATA

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).

TEMPERATURE MAP

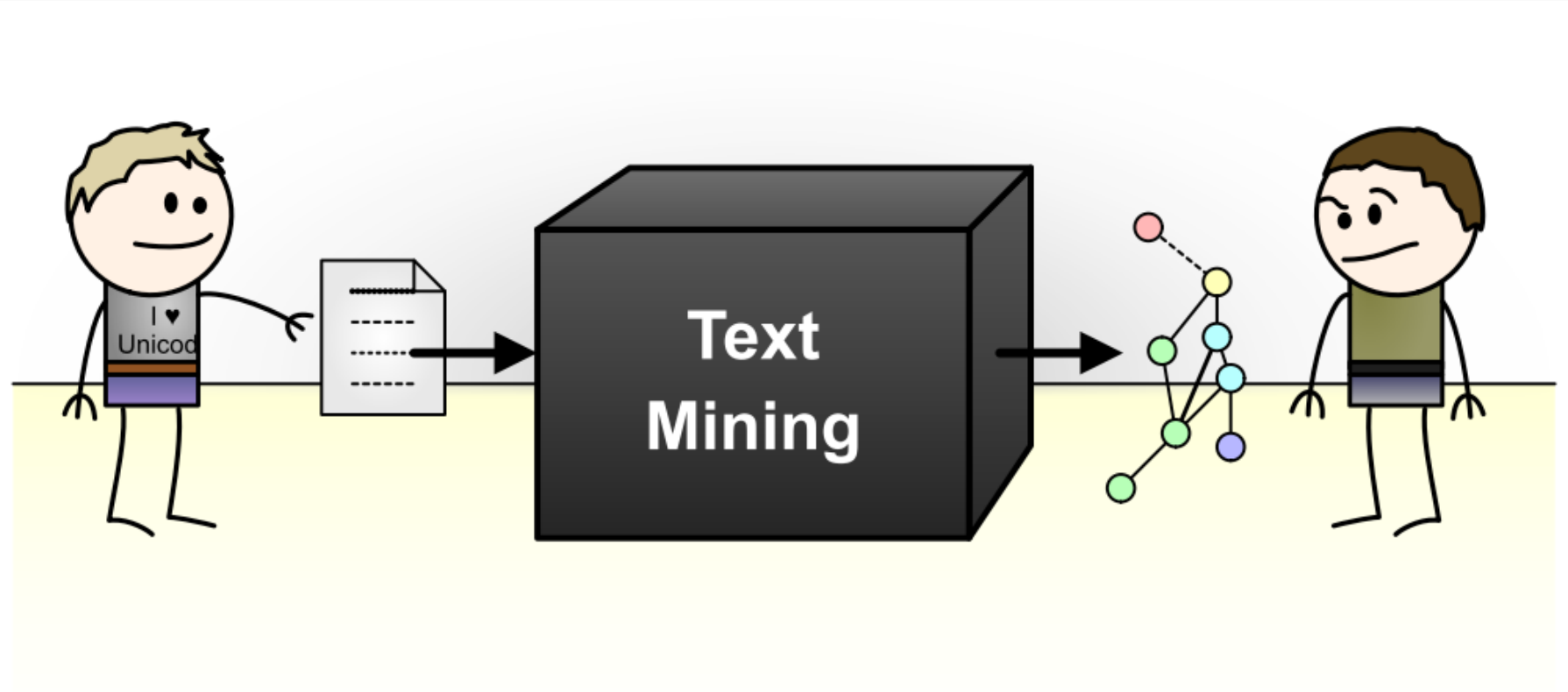


ACID PARADIGM

ACID for the Digital Humanities:

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

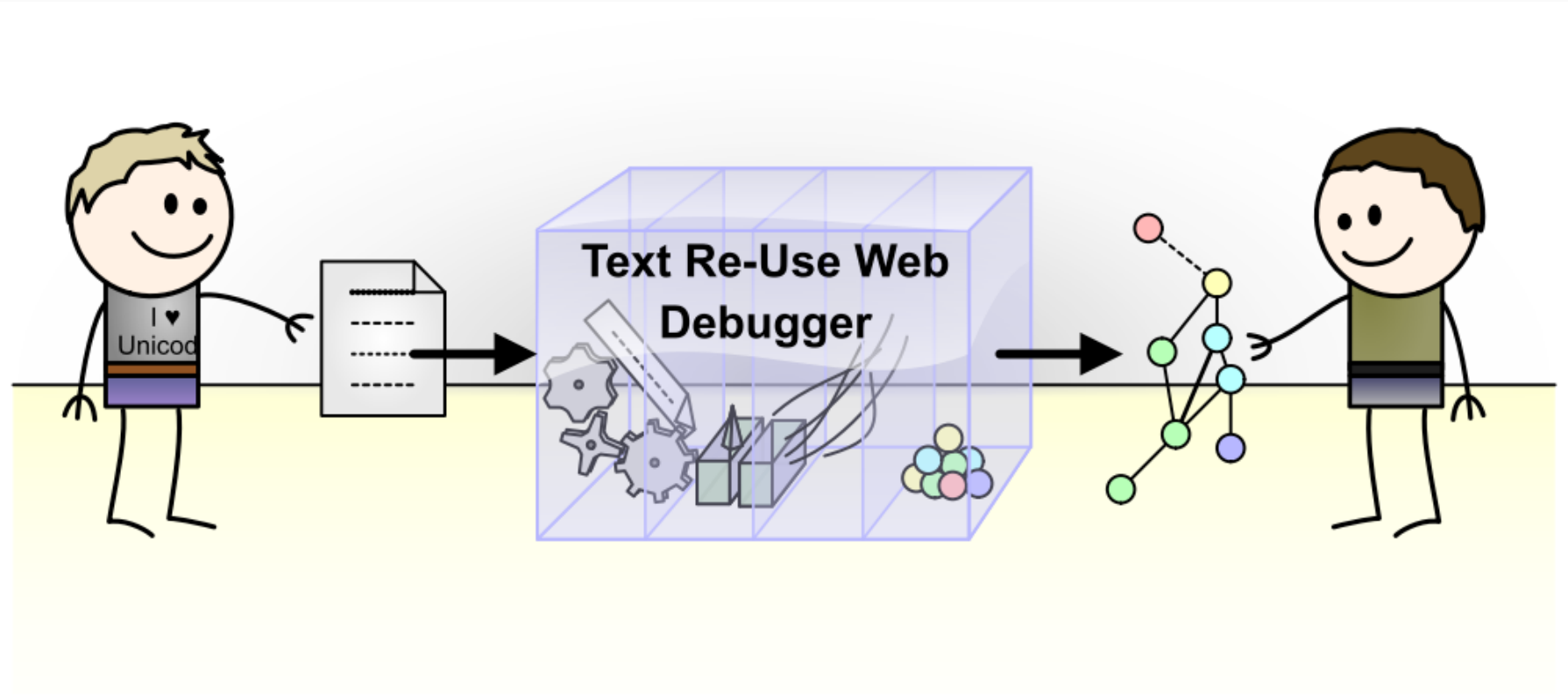
ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE I



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?

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

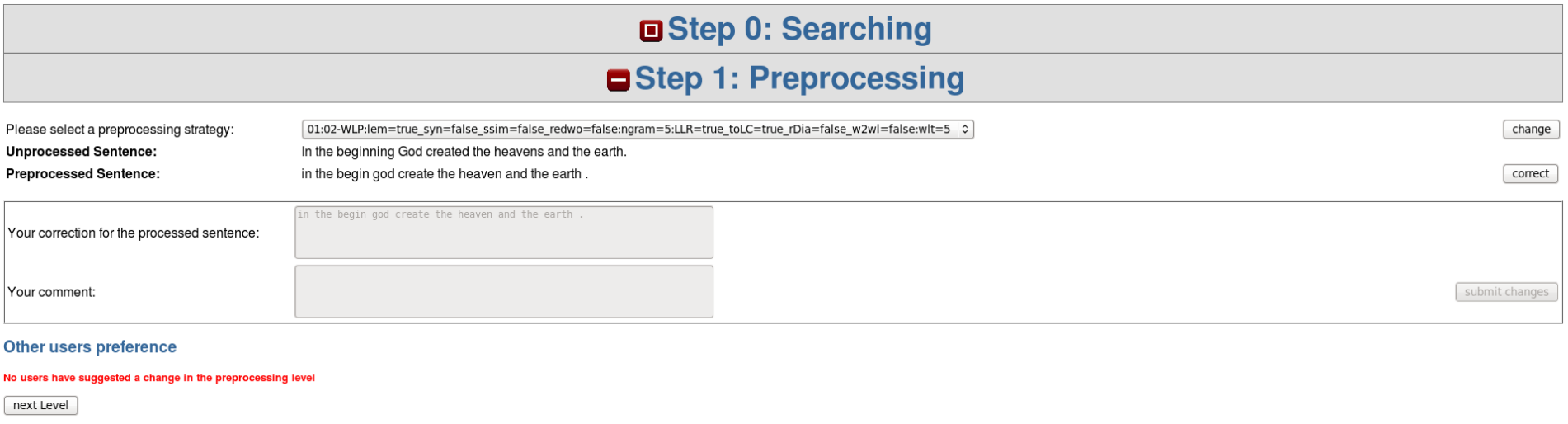
Fields with * are necessary

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

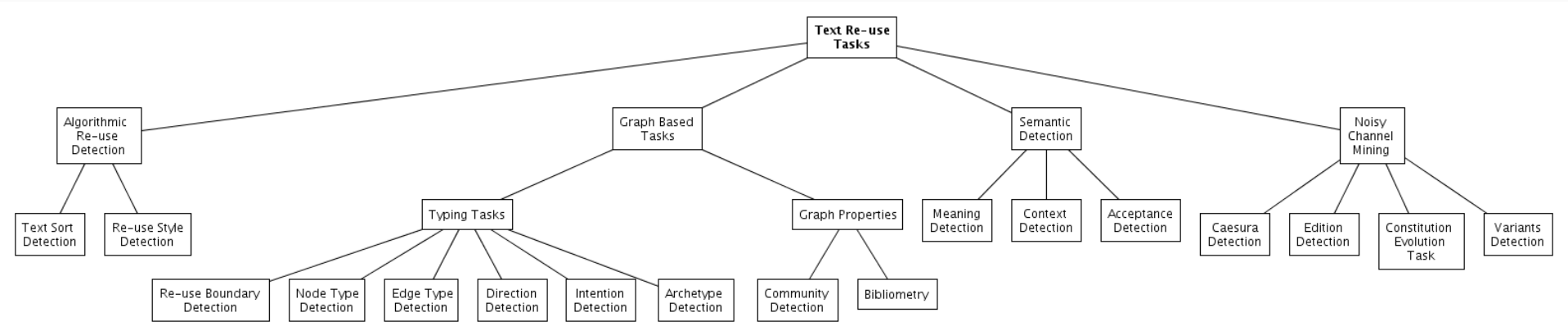
Other users preference

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

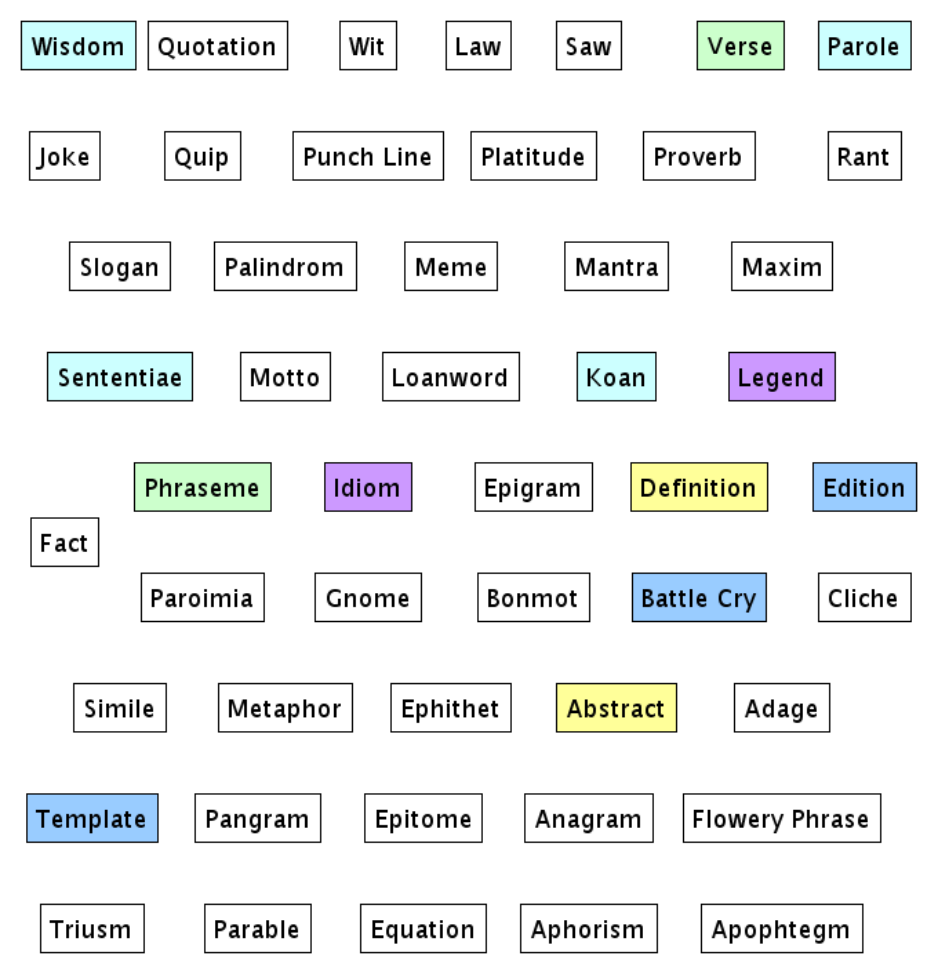
ACID FOR THE DIGITAL HUMANITIES: COMPLEXITY



ACID FOR THE DIGITAL HUMANITIES: INTEROPERABILITY

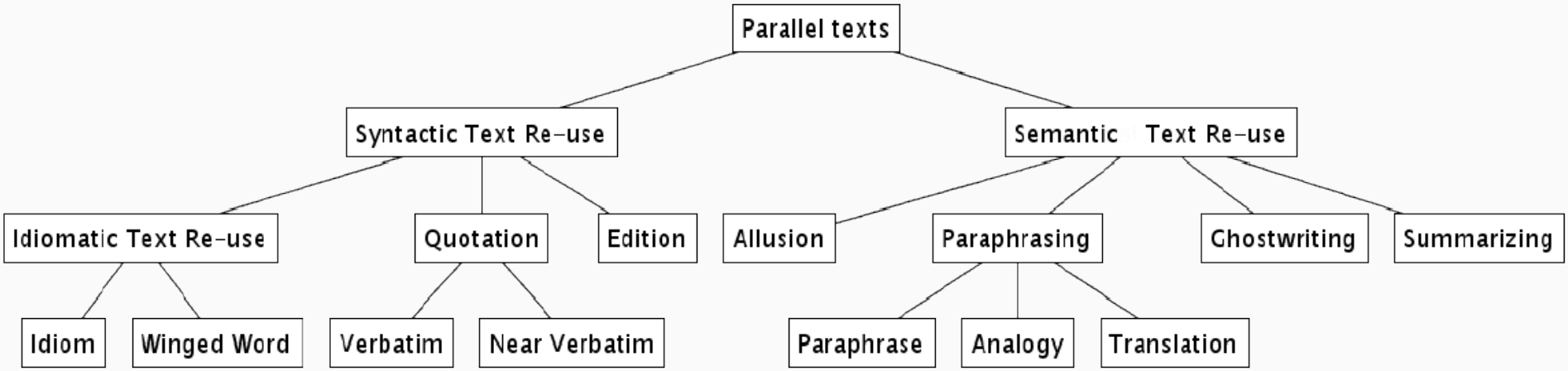
cit-quote-bibl	blockquote	bibl without quote
<cit> <quote> du/o ku/nes a)rgoi\ ei(/ponto </quote> <bibl n="Hom. Od. 2.11"> Od. 2.11 </bibl> </cit>	<quote rend="blockquote"> <line> a)gxou= d' i(stame/nh e)/pea ptero/enta proshu/da <bibl n="Hom. Il. 4.92">Il. 4.92</bibl> </line><line> a)ll' a)/ge nu=n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

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)



KEY PROBLEM

Question:

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

TRACER: suite of 700 algorithms; developed by Marco Büchler.

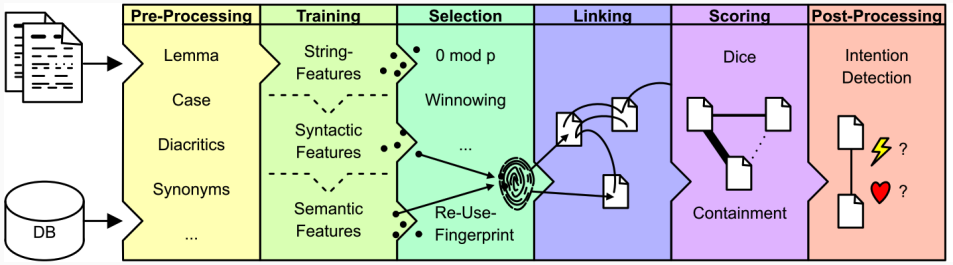


Figure 2: TRACER steps. More than 1M permutations of implementations of different levels are possible.

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

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

MOTIF DATABASE

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.

COMPUTATIONAL FOLKLORISTICS

> *"Over the course of the past decade [...] the size and scope of digital archives of folklore have exploded, and the magnitude of digital materials available for folkloristic consideration has increased exponentially." (Tangherlini, 2016, p. 5).*

> *"We are in the **very early days of working computationally** with rich folklore resources [...]." (Tangherlini, 2016, p. 10).*

Tangherlini (2013) outlines four areas of research in computational folkloristics: (1) collecting **and archiving**, (2) **indexing and classifying**, (3) **visualization and navigation**, and (4) analysis.

GRIMMS MÄRCHEN: MOTIVATION

Motivation:

- Impact on society
- Global scope
- Big Data
- Interdisciplinary

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.

MOTIFS AS SCHOLARLY PRIMITIVES

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

GOAL

The collection and automatic detection of folktale motifs as text reuse units at scale and across languages.

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.

MOTIF DATABASE RATIONALE

Why build the database?

- Investigate & record primitives and their changes;
- Improve algorithms to sharpen our understanding of why and how a text is reused.

Premise: To trace a reuse 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.

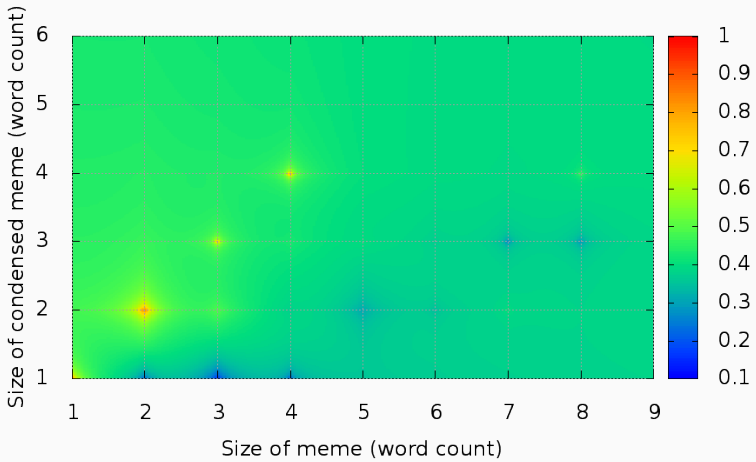


Figure 5: Similarity plot of idiom/meme samples using Google's Custom Search engine (online).

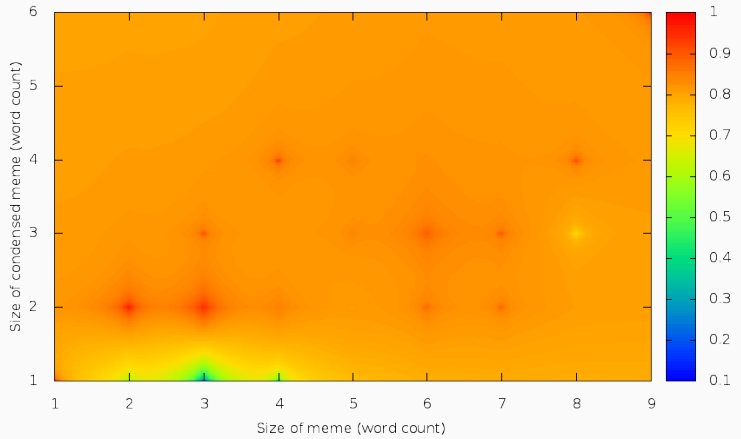


Figure 6: Similarity plot of idiom/meme samples using Apache Lucene (offline).

Thompson Motif Index (TMI) ontology (OWL/RDF), by Antónia Koštová, Thierry Declerck and Tyler Klement (Declerck et al., 2016).

```
<http://www.semanticweb.org/tonka/ontologies/2015/5/tmi-atu-ontology#T11.6>
  rdf:type :Motif ;
  rdf:type :T11 ;
  rdf:type owl:NamedIndividual ;
  rdfs:comment "\"Terminal motif T11.6\""@en ;
  rdfs:label "\"Wish for wife red as blood, white as snow, black as raven.\""@en ;
.
```

Figure 7: Representation of a motif in the TMI ontology. Image reproduced with permission of Thierry Declerck.

CONCLUSION

Contribution so far:

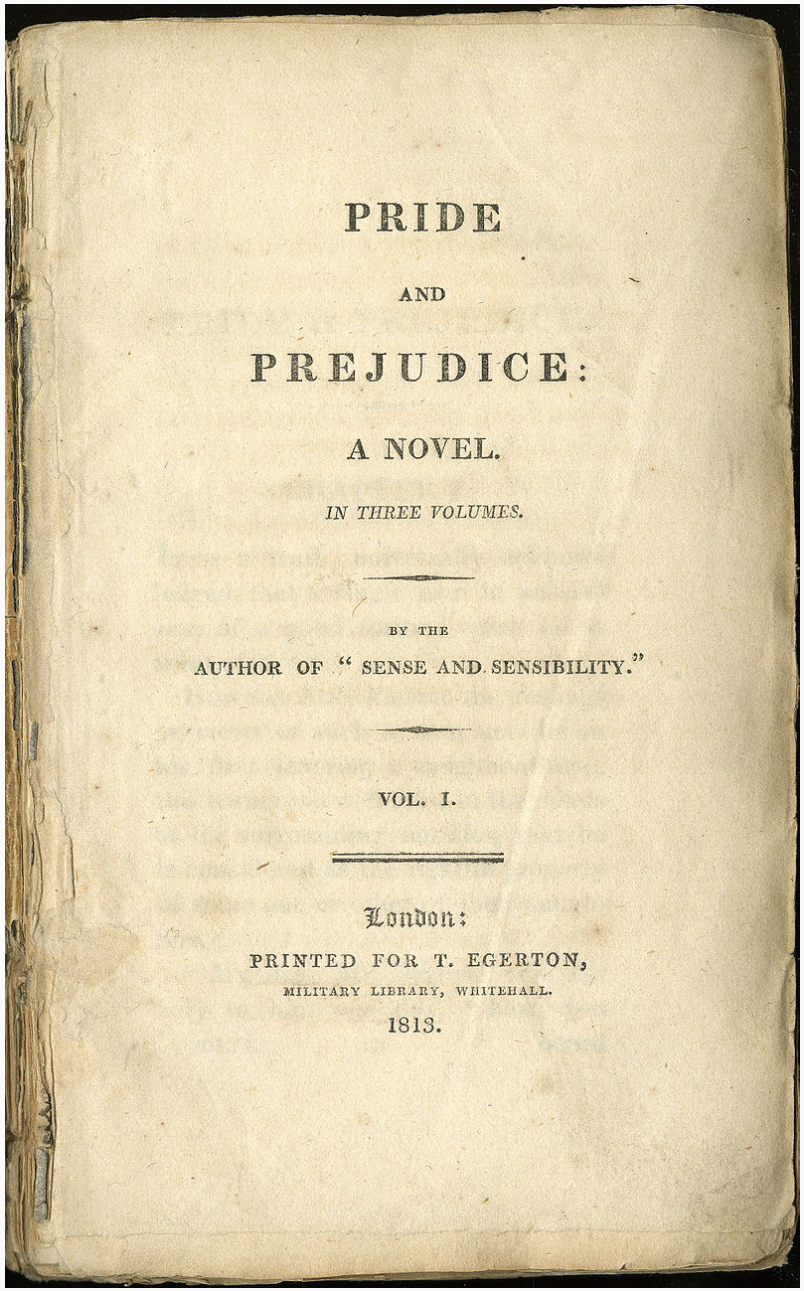
- Multilingual, curated dataset (not openly available yet);
- Results for online vs. offline text reuse detection at scale.

Short-term objectives:

- Run computational analyses on collected folktale data and study the results;
- Release multilingual dataset in SKOS XL for integration with existing ontological resources;
- Extend dataset to more languages and collections.

ANALYSING THE PROCESS OF TEXT REUSE: A CASE STUDY OF JANE AUSTEN

JANE AUSTEN'S PRIDE & PREJUDICE



PRIDE

AND

PREJUDICE:

A NOVEL.

IN THREE VOLUMES.

BY THE

AUTHOR OF "SENSE AND SENSIBILITY."

VOL. I.

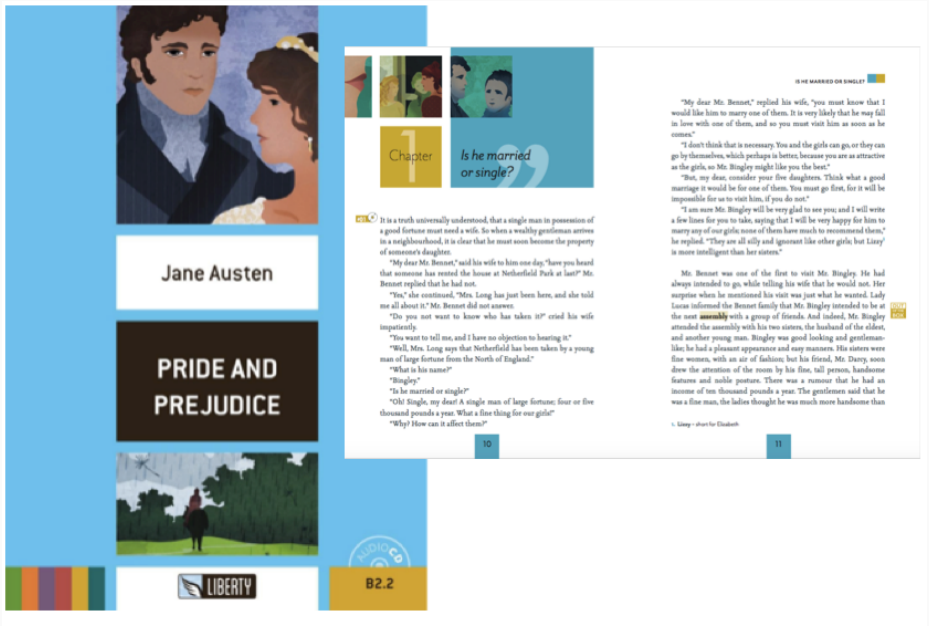
London:

PRINTED FOR T. EGERTON,

MILITARY LIBRARY, WHITEHALL.

1813.

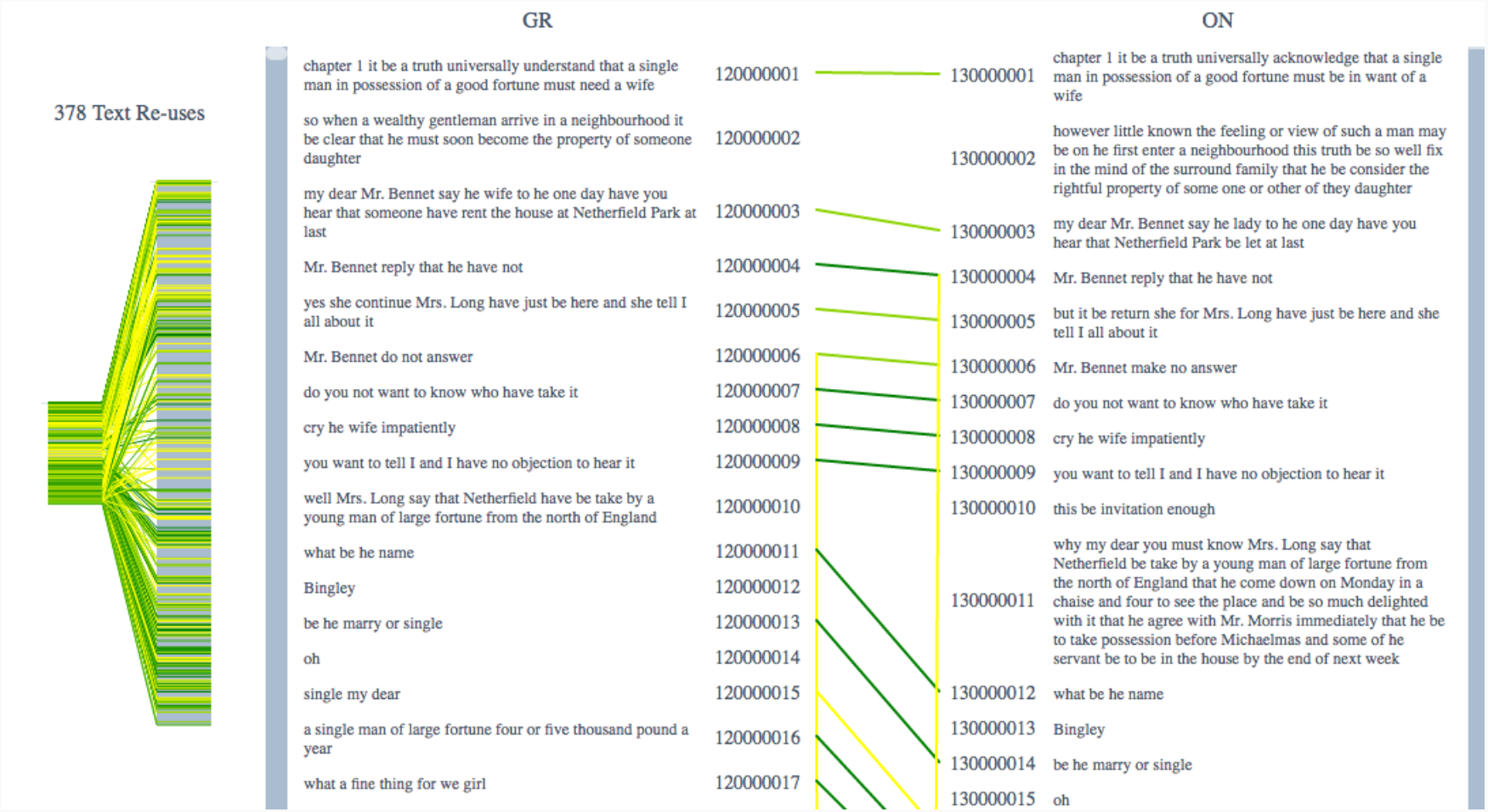
GRADED READER



Defintion:

Graded readers are ”simplified books written at varying levels of difficulty for second language learners”, which ”cover a huge range of genres ranging from adaptation of classic works of literature to original stories, to factual materials such as biographies, reports and so on” [Wallace 2012].

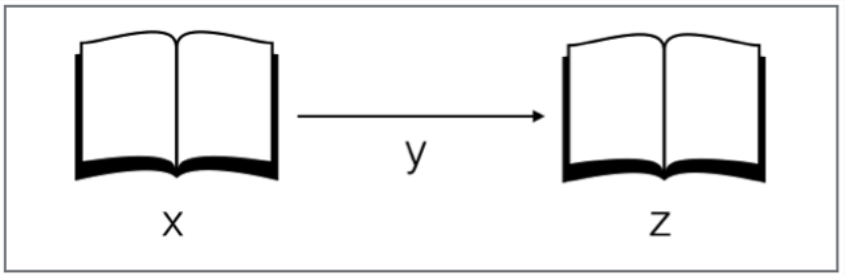
AUTOMATIC ALIGNMENT OF ORIGINAL NOVEL WITH GRADED READER



RESEARCH OBJECTIVES

To computationally analyse the process Y and classifying the changes:

- Do the changes follow strict rules?
- Do they form patterns?
- Can they be computationally reproduced?



Categories of changes:

- Cognitive
- Structural
- Cognitive and structural

TYPES OF CHANGES

Structural changes:

- Elizabeth is exceedingly handsome.
- Elizabeth is very beautiful.

Congnitive changes:

- ... Soon after this event, Elizabeth received a visit...

Structural & congnitive changes:

- Elizabeth is exceedingly beautiful.

TESTING THE SIMPLIFICATION WITH READABILITY TESTS

Readability tests aim to classify texts by their degree of complexity and understandability. Measured primitives are sentence length and difficulty of the words.

Two tests, the ARI score and the Dale-Chall-Index have been selected:

The ARI score is based on the word length and the sentence length:

$$R_{ARI} = 4.71\left(\frac{characters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43 \quad (1)$$

The Dale-Chall-Index is based on the word frequency (3000 most frequent words) and the sentence length:

$$R_{DCI} = 0.1579\left(\frac{difficult\ words}{words} * 100\right) + 0.0496\left(\frac{words}{sentences}\right) \quad (2)$$

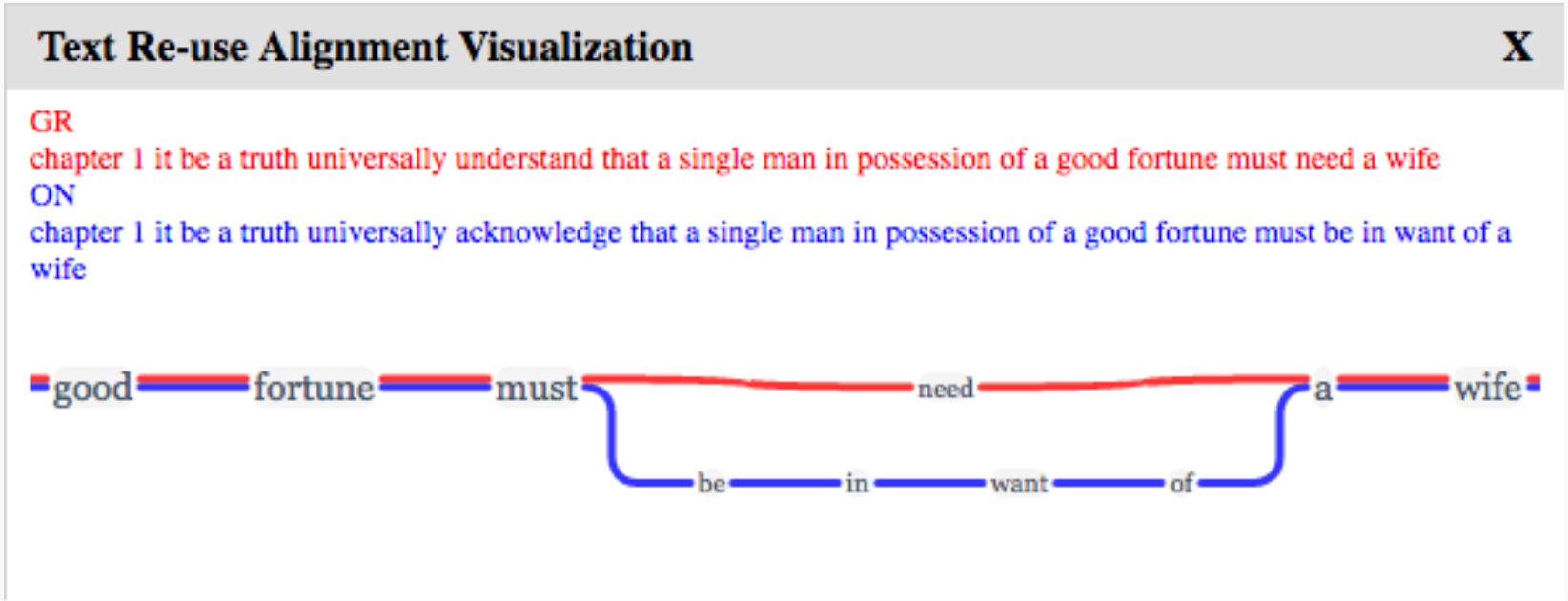
RESULTS OF THE SIMPLIFICATION WITH READABILITY TESTS

Readability test result matrix:

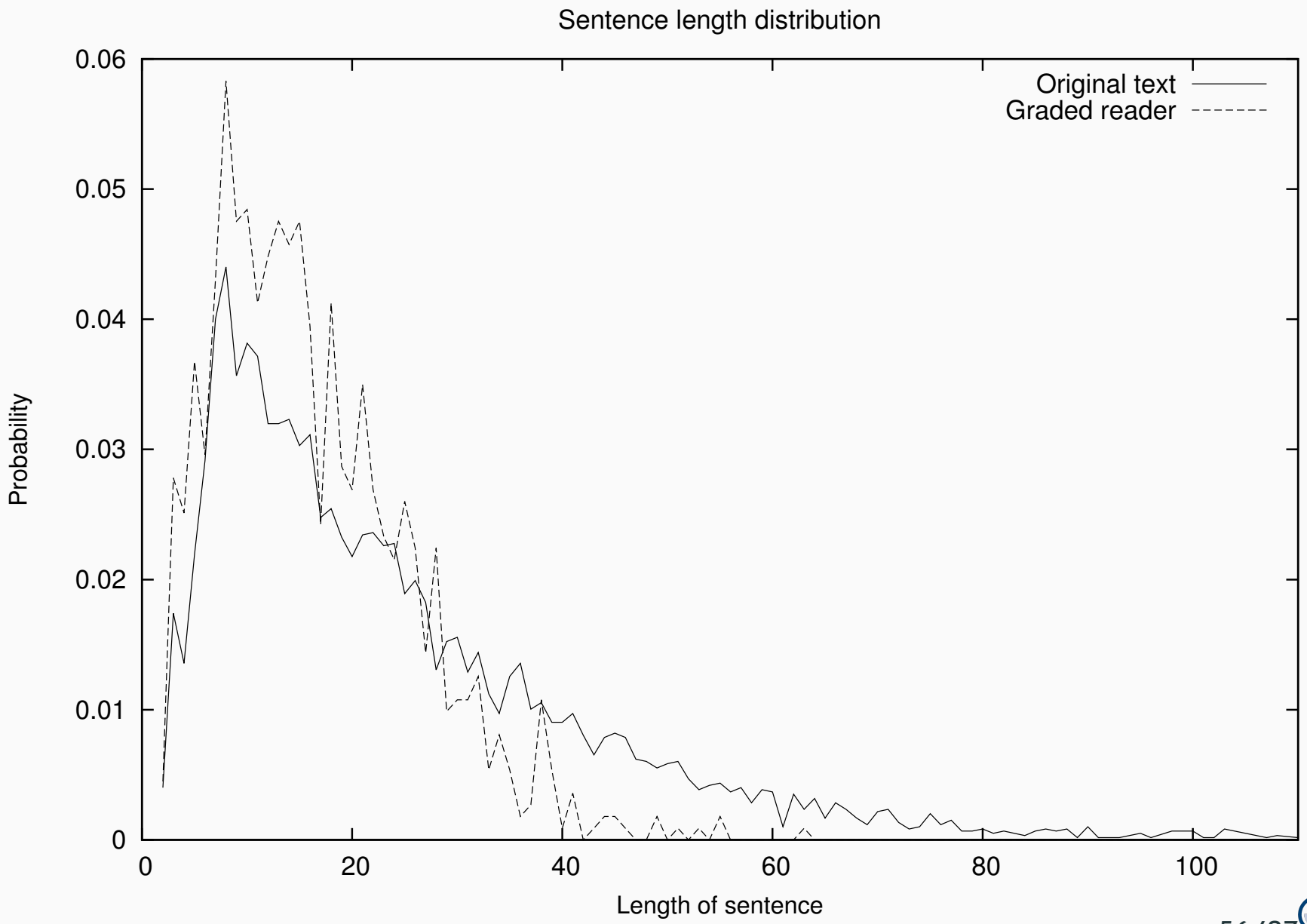
	ARI	Dale-Chall
Original Novel	14-15 year olds	14-16 year olds
Graded Reader	11-12 year olds	11-13 year olds

SIMPLIFICATION & SENTENCE LENGTH

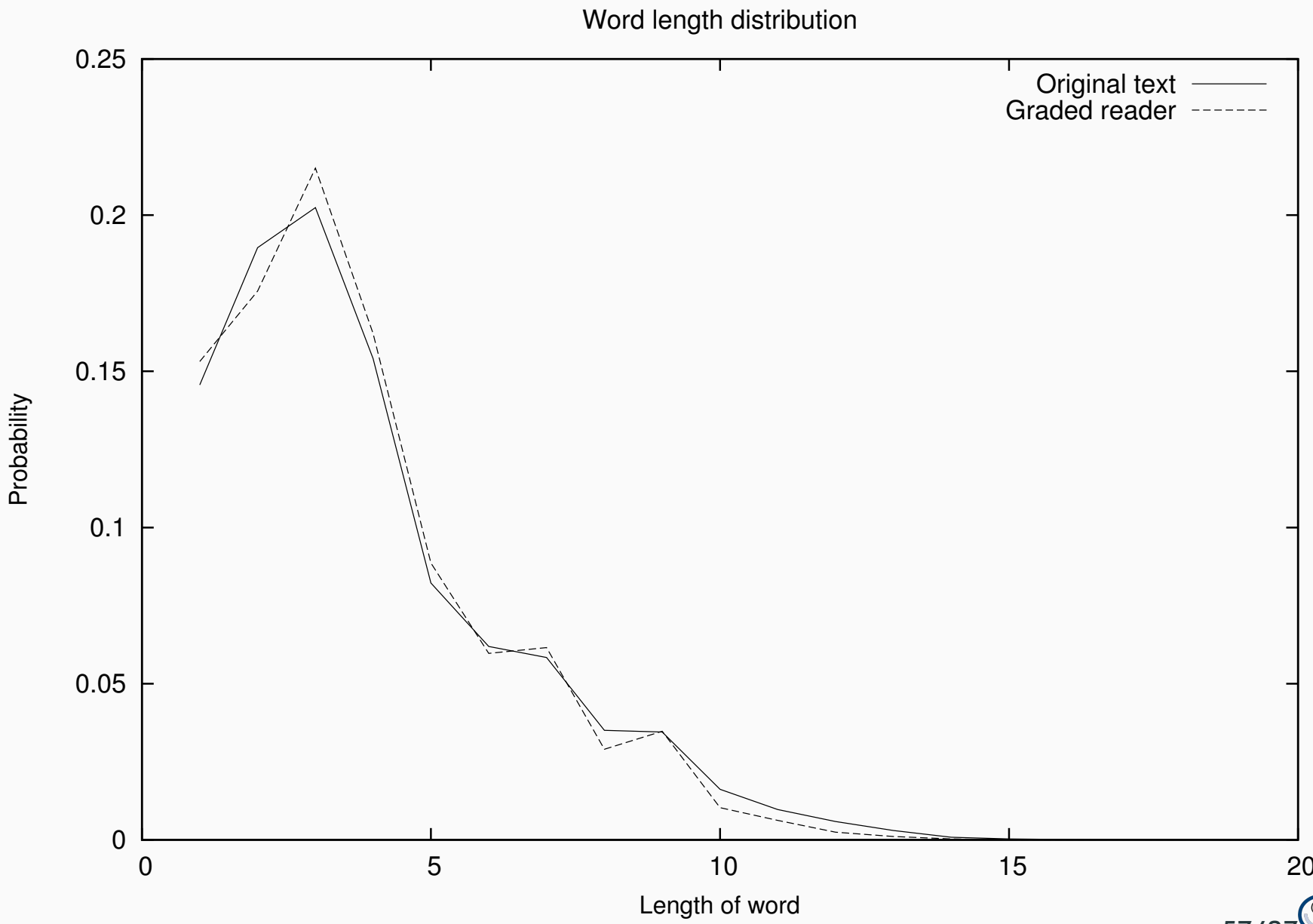
An example of a structural text simplification > many-to-one.



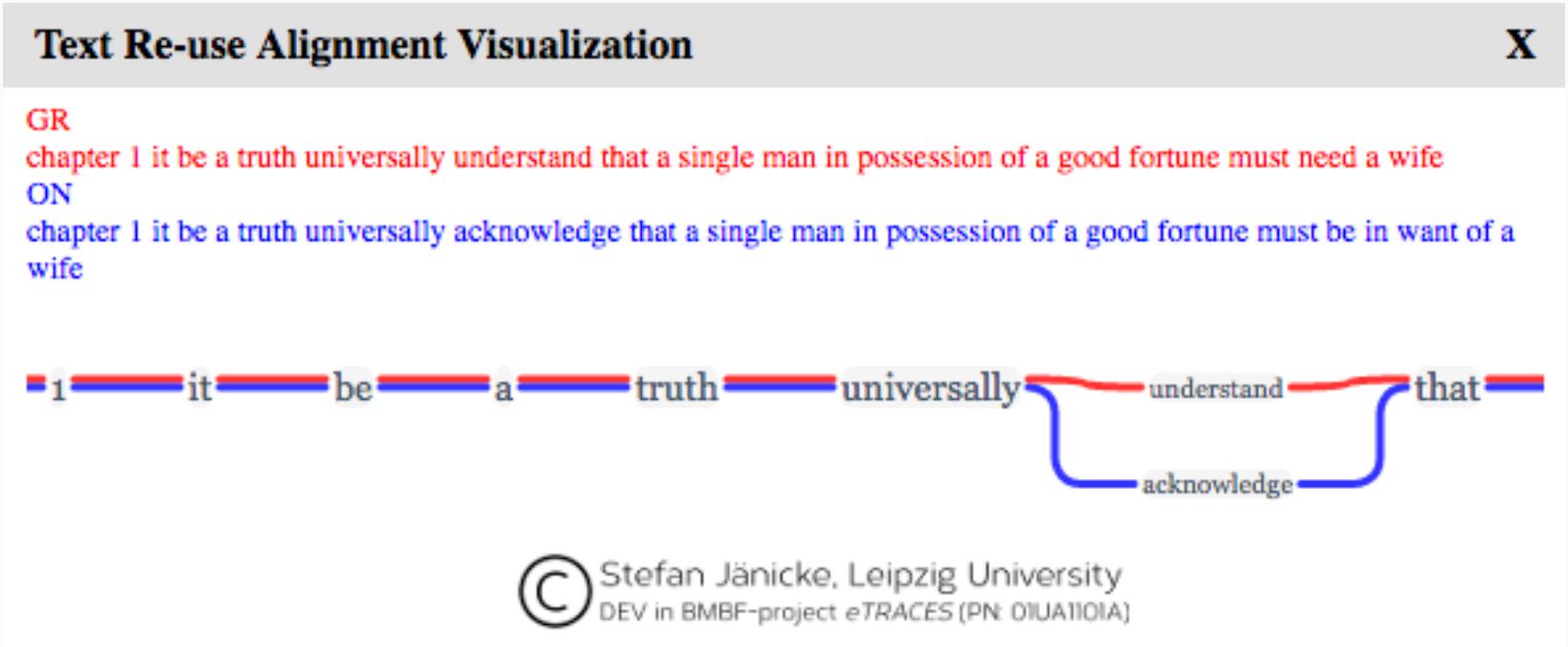
COMPARISON OF SENTENCE LENGTH



COMPARISON OF WORD LENGTH



EXAMPLE OF WORD REPLACEMENT



Conclusion: The simplification of words is provided by using easier and more frequent words instead of shortened words.

DIFFERENCE ANALYSIS: WORDS APPEARING ONLY IN THE ORIGINAL

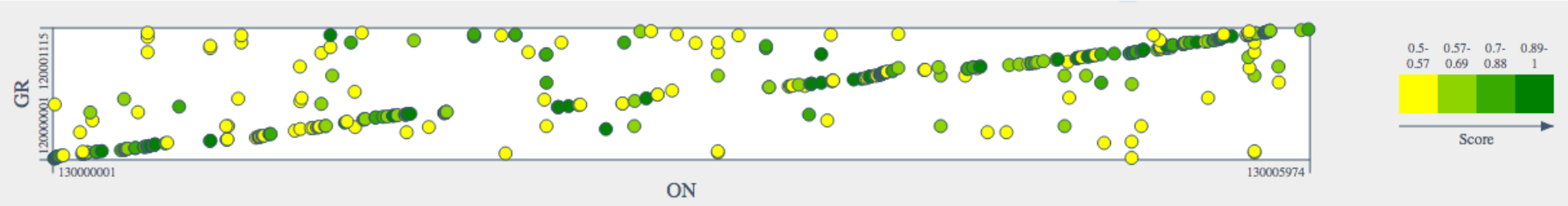
Word	Frequency	Word	Frequency
upon	75	table	31
least	65	astonishment	30
acquaintance	63	fancy	30
either	59	attempt	29
whose	59	dine	29
dare	53	beg	28
regard	53	depend	28
determine	47	highly	28
scarcely	45	satisfaction	28
ladyship	42	acknowledge	27
former	38	credit	27
put	36	thus	27
amiable	35	disposition	26
deal	34	exceedingly	26
design	32	praise	26

DIFFERENCE ANALYSIS FOR PART-OF-SPEECH TAGS

PoS	More frequent in ON	Similar frquency	More frequent in GR
JJS adjective, superlative	X		
JJR adjective, comparative	X		
PDT predeterminer	X		
RBS adverb, superlative	X		
WDT WH-determiner	X		
FW foreign word	X		
: colon	X		
WP$ WH-pronoun, possessive	X		
NNPS noun, proper, plural	X		
SYM symbol	X		
RP particle		X	
RB adverb		X	
VB verb, base form		X	
TO 'to' as preposition		X	
JJ adjective or numeral, ordinal		X	
NNS noun, proper, singular		X	
CC conjunction, coordinating		X	
PRP$ prounoun, possessive		X	
NN noun, common, singular		X	
MD modal auxiliary		X	
IN preposition or conjuction, subordinating		X	
DT determiner		X	
VBN verb, past participle		X	
VBG verb, present participle		X	
POS genitive marker		X	
RBR adverb, comparative		X	
EX existential 'there'		X	
UH interjection			X
NNP noun, proper, plural			X
WRB WH-adverb			X
VBD verb, past tense			X
VBP verb, present tense, not 3rd person singular			X
VBZ verb, present tense, 3rd person singular			X
WP WH-pronoun			X
CD numeral, cardinal			X
PRP prounoun, personal			X

MACRO SCALE: VISUALISATION OF THE SELECTION PROCESS

The Dotplot view of original novel against the graded reader on a sentence-wise segmentation uncovers which passages were taken over in the graded reader and which not:



ANALYTIC MINING OF 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?

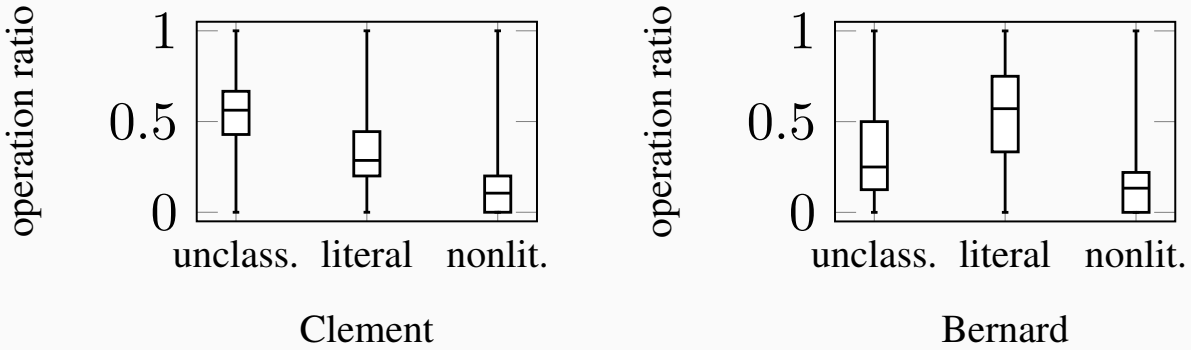


Figure 8: Ratios of operations in reuse instances. literal: NOP, lem, lower, etc.; nonlit: syn, hyper, etc.

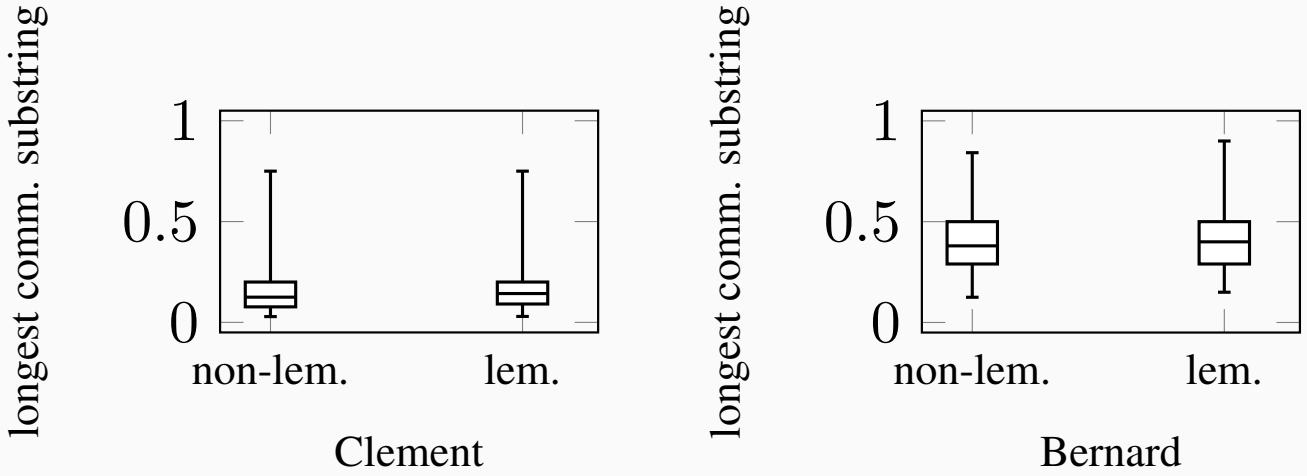


Figure 9: Ratios of literal overlap between reuse instances and originals

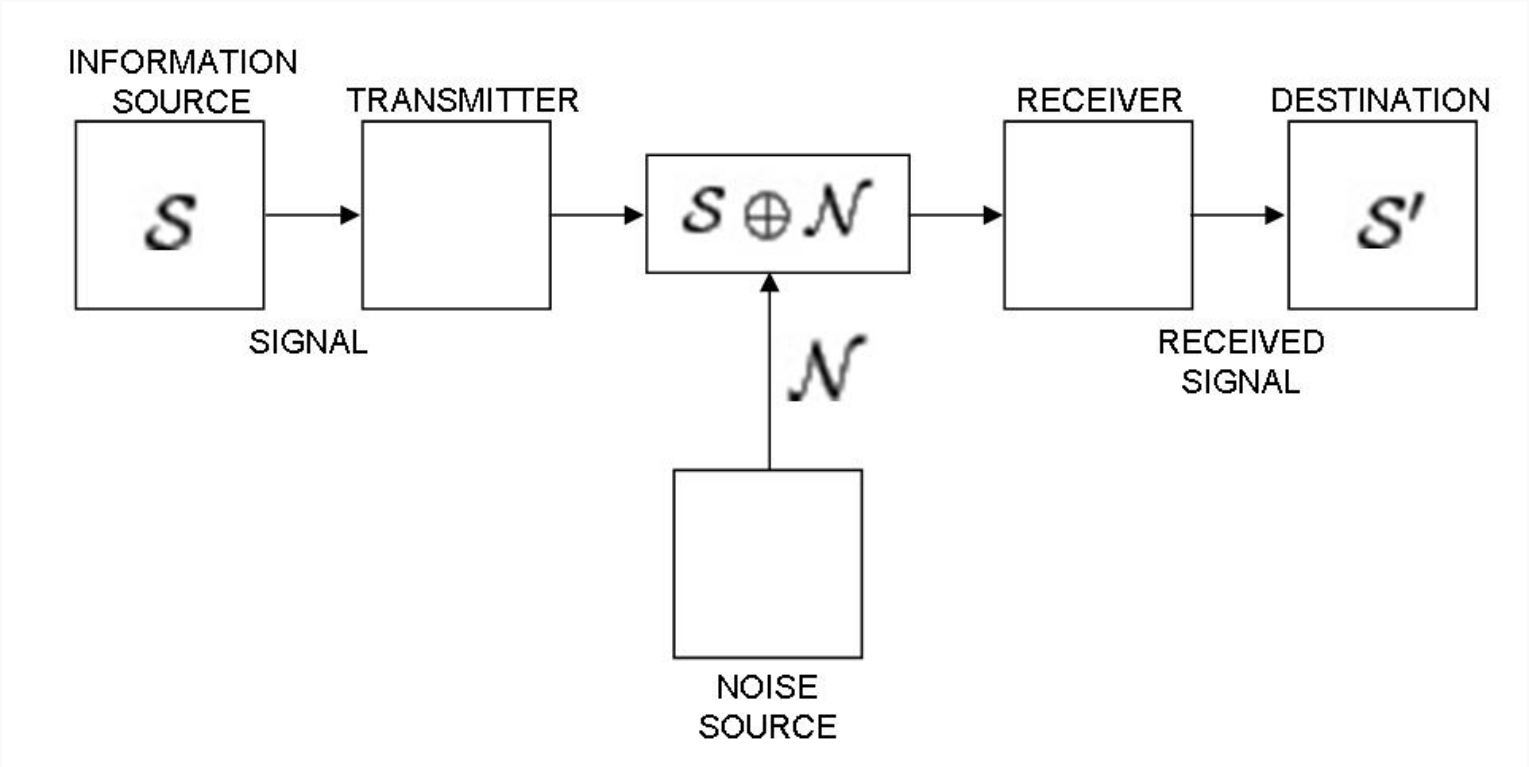
How is the non-literally reused text modified in our datasets?
How can linguistic resources support the discovery of non-literal reuse?

Table 3: Absolute numbers of operations identified automatically

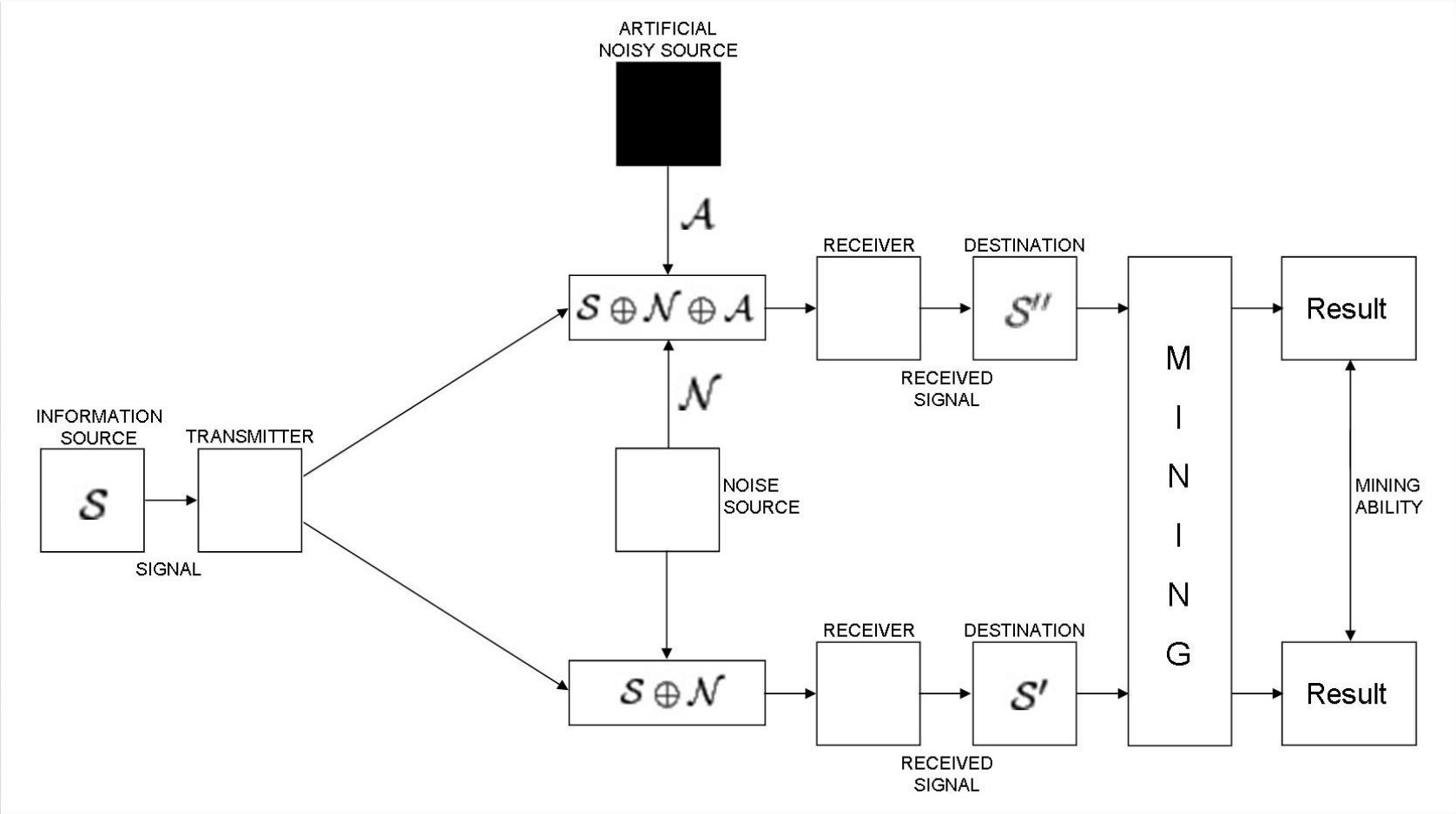
	literal				nonliteral				unclassified		
	NOP	upper	lower	lem	syn	hyper	hypo	co-hypo	no_rel_found	lem_missing	total
Greek	337	6	0	356	153	20	14	101	563	639	2189
Latin	587	0	44	102	60	14	28	68	347	85	1335

EVALUATION OF TEXT REUSE

Basic idea: Embed historical text reuse in Shannon's **Noisy Channel** theorem.



METHODOLOGY: NOISY CHANNEL EVALUATION I



Hint: The results are ALWAYS compared between the natural texts and the randomised texts as a whole.

METHODOLOGY: NOISY CHANNEL EVALUATION II

Signal-Noise-Ratio *adapted* from signal- and satellite techniques:

$$SNR = \frac{P_{signal}}{P_{noise}}$$

Signal-Noise-Ratio *scaled*, unit is dB:

$$SNR_{db} = 10.log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$$

Mining Ability (in dB): The Mining Ability describes the power of a method to make distinctions between natural-language structures/patterns and random noise given a model with the same parameters.

$$L_{Quant}(\Theta) = 10.log_{10}\frac{|E_{D_s,\phi_\Theta}|}{max(1,|E_{D_s^m},\phi_\Theta|)}dB$$

METHODOLOGY: NOISY CHANNEL EVALUATION III

Motivation for randomisation by **Word Shuffling**:

1. Syntax and distributional semantics are randomised and ”destroyed”.
2. Distributions of words and sentence lengths remain unchanged; changes JUST and ONLY depend on destruction of 1) and are not induced by changes of distributions.
3. Easy measurement of ”randomness” of the randomising method with the entropy test:

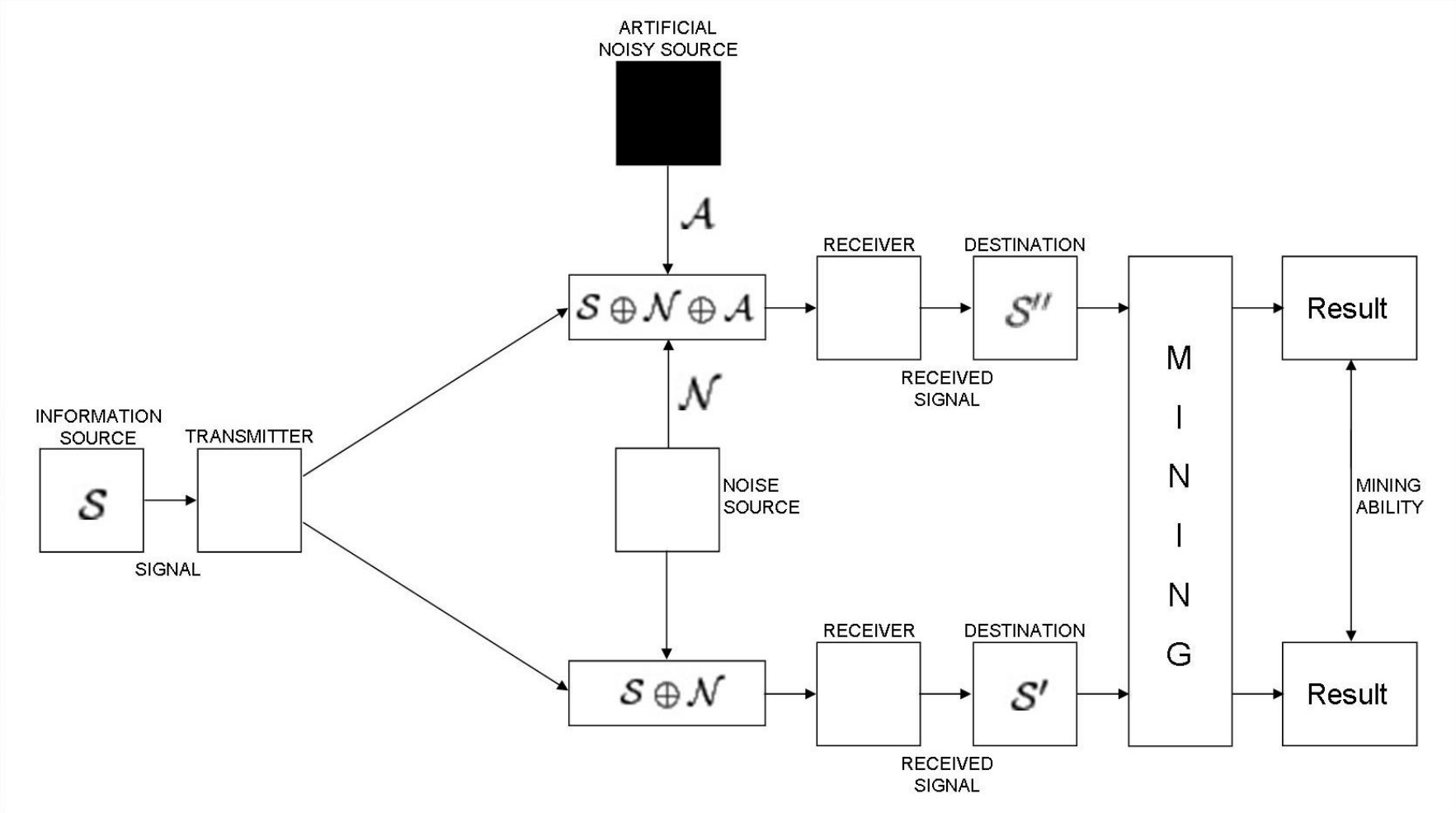
$$\Delta H^n = H_{max} - H^n$$

Die Wahl von $n \in [180, 183]$ sichert eine Genauigkeit von $\Delta H^n \leq 10^{-3}$ Bit für den Entropietest.

METHODOLOGY: TEXT REUSE COMPRESSION

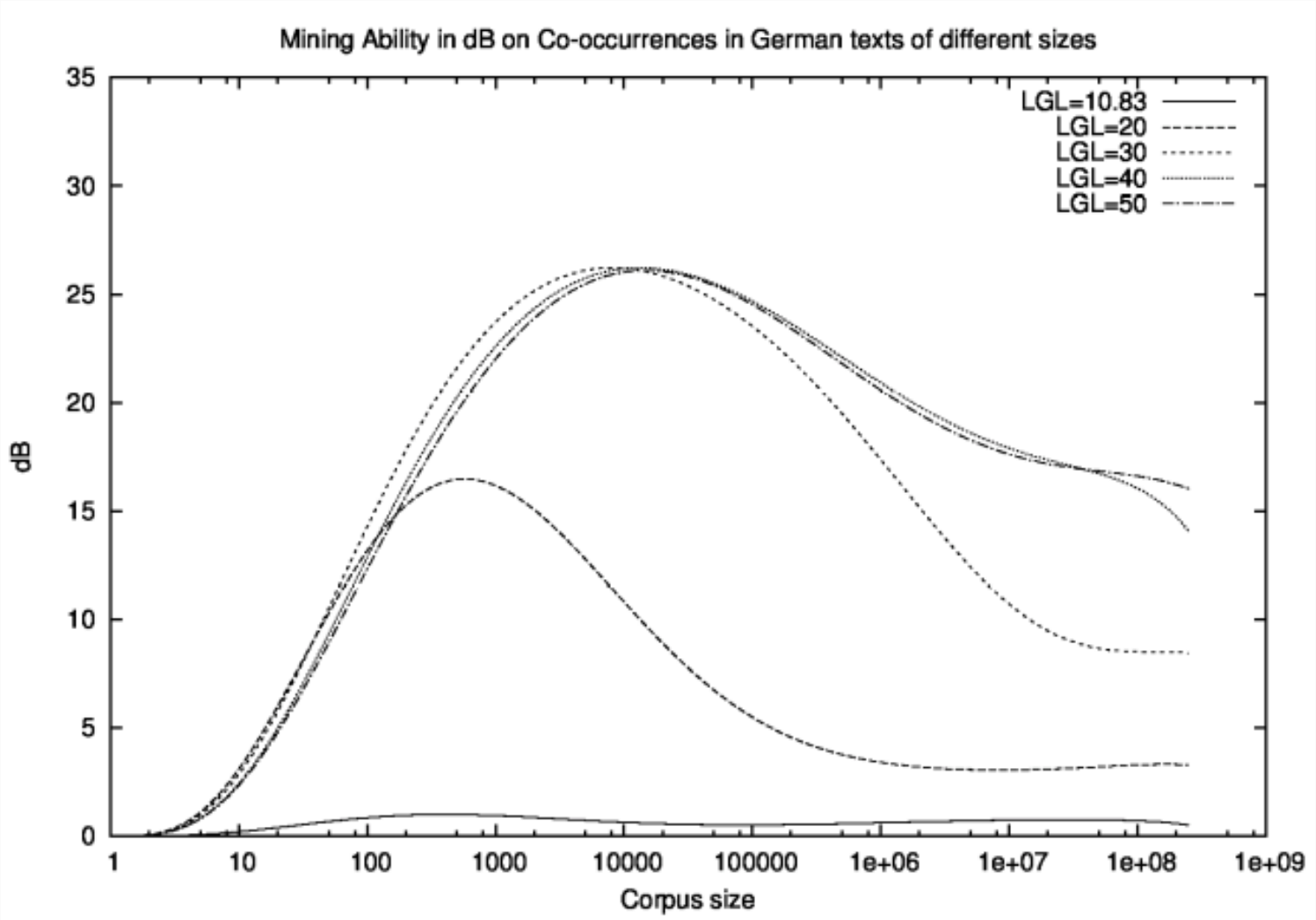
$$C_{\Theta} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} \theta_{\Theta}(S_i, S_j)}{n.m}$$

RANDOMNESS & STRUCTURE



Question: Why is the result of a randomised Digital Library typically not empty?

RANDOMNESS & STRUCTURE: IMPACTS



Corpus size in sentences (average sentence length is ca. 18 words). LGL is the threshold for the Log-Likelihood-Ratio.

TEXT REUSE IN ENGLISH BIBLE VERSIONS: SETUP

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.

TEXT REUSE IN ENGLISH BIBLE VERSIONS: RESULTS – RECALL

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.09	0.10	0.11	0.12
ASV vs. DBY	0.16	0.17	0.17	0.17	0.28	0.30	0.30	0.31	0.70	0.72	0.73	0.74
ASV vs. KJV	0.36	0.38	0.37	0.38	0.53	0.56	0.55	0.56	0.86	0.88	0.88	0.88
ASV vs. WEB	0.32	0.34	0.32	0.33	0.46	0.48	0.47	0.47	0.76	0.79	0.77	0.77
ASV vs. WBS	0.27	0.29	0.28	0.29	0.44	0.46	0.46	0.46	0.82	0.84	0.84	0.85
ASV vs. YLT	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.18	0.21	0.25	0.26

TEXT REUSE IN ENGLISH BIBLE VERSIONS: RECALL VS. TEXT REUSE COMPRESSION

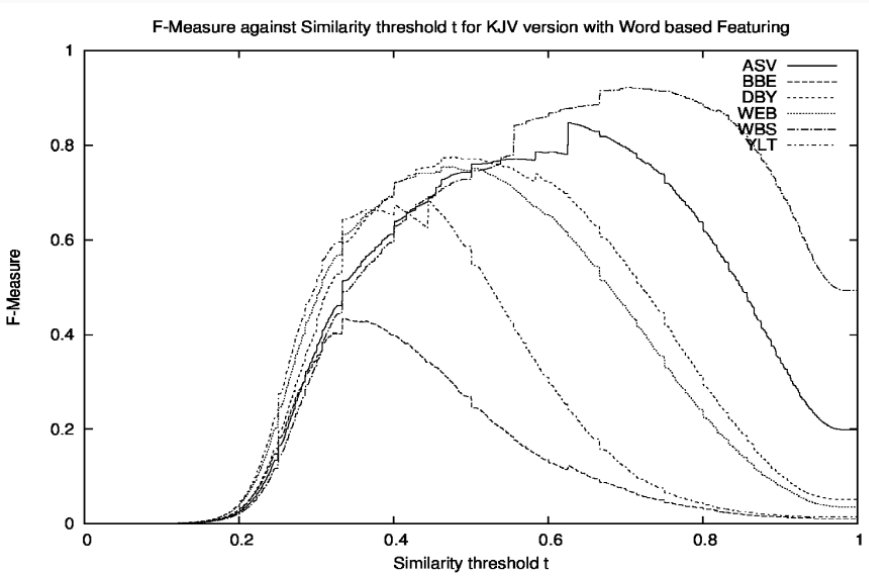
With

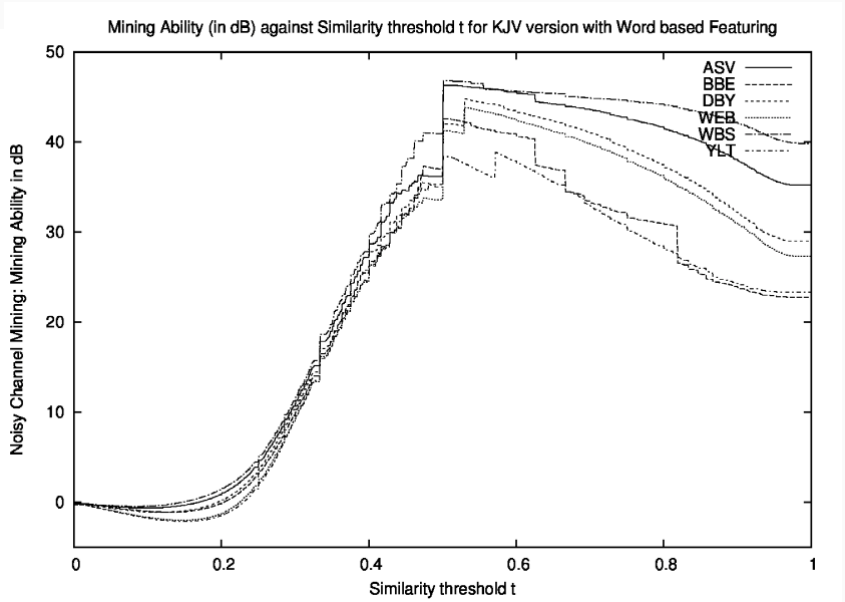
	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.09	0.10	0.11	0.12
ASV vs. DBY	0.16	0.17	0.17	0.17	0.28	0.30	0.30	0.31	0.70	0.72	0.73	0.74
ASV vs. KJV	0.36	0.38	0.37	0.38	0.53	0.56	0.55	0.56	0.86	0.88	0.88	0.88
ASV vs. WEB	0.32	0.34	0.32	0.33	0.46	0.48	0.47	0.47	0.76	0.79	0.77	0.77
ASV vs. WBS	0.27	0.29	0.28	0.29	0.44	0.46	0.46	0.46	0.82	0.84	0.84	0.85
ASV vs. YLT	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.18	0.21	0.25	0.26
BBE vs. ASV	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.09	0.10	0.11	0.12
BBE vs. DBY	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.07	0.08	0.08	0.10
BBE vs. KJV	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.08	0.09	0.10	0.11
BBE vs. WEB	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.11	0.12	0.13	0.15
BBE vs. WBS	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.10	0.11	0.13
BBE vs. YLT	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.04
DBY vs. ASV	0.16	0.17	0.17	0.17	0.28	0.30	0.30	0.31	0.70	0.72	0.73	0.74
DBY vs. BBE	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.07	0.08	0.08	0.10
DBY vs. KJV	0.12	0.13	0.12	0.13	0.22	0.24	0.23	0.24	0.62	0.65	0.65	0.66
DBY vs. WEB	0.07	0.08	0.07	0.08	0.14	0.15	0.14	0.15	0.46	0.49	0.49	0.51
DBY vs. WBS	0.12	0.13	0.12	0.13	0.22	0.24	0.23	0.24	0.64	0.67	0.67	0.68
DBY vs. YLT	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.18	0.21	0.26	0.27
KJV vs. ASV	0.36	0.38	0.37	0.38	0.53	0.56	0.55	0.56	0.86	0.88	0.88	0.88
KJV vs. BBE	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.08	0.09	0.10	0.11
KJV vs. DBY	0.12	0.13	0.12	0.13	0.22	0.24	0.23	0.24	0.62	0.65	0.65	0.66
KJV vs. WEB	0.10	0.11	0.10	0.10	0.18	0.20	0.19	0.19	0.51	0.55	0.53	0.55
KJV vs. WBS	0.75	0.78	0.76	0.77	0.89	0.91	0.90	0.90	0.99	0.99	0.99	0.99
KJV vs. YLT	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.14	0.16	0.19	0.20
WEB vs. ASV	0.32	0.34	0.32	0.33	0.46	0.48	0.47	0.47	0.76	0.79	0.77	0.77
WEB vs. BBE	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.11	0.12	0.13	0.15
WEB vs. DBY	0.07	0.08	0.07	0.08	0.14	0.15	0.14	0.15	0.46	0.49	0.49	0.51
WEB vs. KJV	0.10	0.11	0.10	0.10	0.18	0.20	0.19	0.19	0.51	0.55	0.53	0.55
WEB vs. WBS	0.11	0.12	0.11	0.12	0.20	0.22	0.21	0.21	0.56	0.60	0.59	0.60
WEB vs. YLT	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.12	0.15	0.16
WBS vs. ASV	0.27	0.29	0.28	0.29	0.44	0.46	0.46	0.46	0.82	0.84	0.84	0.85
WBS vs. BBE	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.10	0.11	0.13
WBS vs. DBY	0.12	0.13	0.12	0.13	0.22	0.24	0.23	0.24	0.64	0.67	0.67	0.68
WBS vs. KJV	0.75	0.78	0.76	0.77	0.89	0.91	0.90	0.90	0.99	0.99	0.99	0.99
WBS vs. WEB	0.11	0.12	0.11	0.12	0.20	0.22	0.21	0.21	0.56	0.60	0.59	0.60
WBS vs. YLT	0.01	0.02	0.02	0.01	0.02	0.03	0.03	0.03	0.15	0.17	0.21	0.22
YLT vs. ASV	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.18	0.21	0.25	0.26
YLT vs. BBE	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.04
YLT vs. DBY	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.18	0.21	0.26	0.27
YLT vs. KJV	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.14	0.16	0.19	0.20
YLT vs. WEB	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.12	0.15	0.16
YLT vs. WBS	0.01	0.02	0.02	0.01	0.02	0.03	0.03	0.03	0.15	0.17	0.21	0.22

Without

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	6.16	6.15	6.16	6.18	6.02	6.01	6.01	5.99	5.42	5.39	5.37	5.33
ASV vs. DBY	5.22	5.19	5.20	5.19	4.98	4.96	4.97	4.95	4.60	4.58	4.58	4.57
ASV vs. KJV	4.97	4.95	4.96	4.95	4.80	4.78	4.79	4.78	4.49	4.47	4.47	4.47
ASV vs. WEB	5.03	5.00	5.02	5.02	4.86	4.84	4.86	4.86	4.60	4.59	4.59	4.59
ASV vs. WBS	5.10	5.07	5.08	5.08	4.89	4.87	4.88	4.87	4.58	4.56	4.56	4.56
ASV vs. YLT	6.34	6.26	6.30	6.29	6.08	6.01	6.05	6.03	5.00	4.95	4.92	4.91
BBE vs. ASV	6.16	6.15	6.16	6.18	6.02	6.01	6.01	5.99	5.42	5.39	5.37	5.33
BBE vs. DBY	6.42	6.36	6.41	6.41	6.24	6.20	6.22	6.20	5.51	5.47	5.44	5.42
BBE vs. KJV	6.35	6.30	6.34	6.32	6.00	5.97	5.99	5.97	5.26	5.23	5.00	4.98
BBE vs. WEB	6.17	6.16	6.17	6.18	6.01	6.00	6.00	6.01	5.30	5.27	5.26	5.22
BBE vs. WBS	5.75	5.74	5.75	5.74	5.55	5.54	5.55	5.54	4.94	4.93	4.83	4.82
BBE vs. YLT	6.86	6.77	6.84	6.85	6.68	6.62	6.66	6.66	5.99	5.94	5.92	5.92
DBY vs. ASV	5.22	5.19	5.20	5.19	4.98	4.96	4.97	4.95	4.60	4.58	4.58	4.57
DBY vs. BBE	6.42	6.36	6.41	6.41	6.24	6.20	6.22	6.20	5.51	5.47	5.44	5.42
DBY vs. KJV	5.49	5.45	5.46	5.44	5.21	5.18	5.19	5.18	4.72	4.70	4.70	4.69
DBY vs. WEB	5.69	5.65	5.67	5.65	5.42	5.39	5.40	5.38	4.85	4.82	4.82	4.80
DBY vs. WBS	5.49	5.45	5.46	5.44	5.21	5.17	5.18	5.17	4.63	4.61	4.61	4.60
DBY vs. YLT	6.38	6.31	6.33	6.32	6.15	6.08	6.09	6.07	5.26	5.19	5.13	5.10
KJV vs. ASV	4.97	4.95	4.96	4.95	4.80	4.78	4.79	4.78	4.49	4.47	4.47	4.47
KJV vs. BBE	6.35	6.30	6.34	6.32	6.00	5.97	5.99	5.97	5.26	5.23	5.00	4.98
KJV vs. DBY	5.49	5.45	5.46	5.44	5.21	5.18	5.19	5.18	4.72	4.70	4.70	4.69
KJV vs. WEB	5.57	5.52	5.55	5.55	5.31	5.27	5.29	5.28	4.81	4.78	4.79	4.78
KJV vs. WBS	4.63	4.61	4.63	4.62	4.55	4.53	4.54	4.54	4.41	4.41	4.41	4.41
KJV vs. YLT	6.39	6.33	6.39	6.39	6.16	6.09	6.15	6.14	5.41	5.33	5.28	5.26
WEB vs. ASV	5.03	5.00	5.02	5.02	4.86	4.84	4.86	4.86	4.60	4.59	4.59	4.59
WEB vs. BBE	6.17	6.16	6.17	6.18	6.01	6.00	6.00	6.01	5.30	5.27	5.26	5.22
WEB vs. DBY	5.69	5.65	5.67	5.65	5.42	5.39	5.40	5.38	4.85	4.82	4.82	4.80
WEB vs. KJV	5.57	5.52	5.55	5.55	5.31	5.27	5.29	5.28	4.81	4.78	4.79	4.78
WEB vs. WBS	5.52	5.48	5.51	5.50	5.26	5.22	5.24	5.23	4.75	4.72	4.73	4.72
WEB vs. YLT	6.38	6.30	6.34	6.33	6.23	6.16	6.17	6.15	5.51	5.44	5.36	5.33
WBS vs. ASV	5.10	5.07	5.08	5.08	4.89	4.87	4.88	4.87	4.58	4.56	4.56	4.56
WBS vs. BBE	5.75	5.74	5.75	5.74	5.55	5.54	5.55	5.54	4.94	4.93	4.83	4.82
WBS vs. DBY	5.49	5.45	5.46	5.44	5.21	5.17	5.18	5.17	4.63	4.61	4.61	4.60
WBS vs. KJV	4.63	4.61	4.63	4.62	4.55	4.53	4.54	4.54	4.41	4.41	4.41	4.41
WBS vs. WEB	5.52	5.48	5.51	5.50	5.26	5.22	5.24	5.23	4.75	4.72	4.73	4.72
WBS vs. YLT	6.25	6.22	6.24	6.34	6.06	6.02	6.04	6.08	5.35	5.29	5.23	5.21
YLT vs. ASV	6.34	6.26	6.30	6.29	6.08	6.01	6.05	6.03	5.00	4.95	4.92	4.91
YLT vs. BBE	6.86	6.77	6.84	6.85	6.68	6.62	6.66	6.66	5.99	5.94	5.92	5.92
YLT vs. DBY	6.38	6.31	6.33	6.32	6.15	6.08	6.09	6.07	5.26	5.19	5.13	5.10
YLT vs. KJV	6.39	6.33	6.39	6.39	6.16	6.09	6.15	6.14	5.41	5.33	5.28	5.26
YLT vs. WEB	6.38	6.30	6.34	6.33	6.23	6.16	6.17	6.15	5.51	5.44	5.36	5.33
YLT vs. WBS	6.25	6.22	6.24	6.34	6.06	6.02	6.04	6.08	5.35	5.29	5.23	5.21

TEXT REUSE IN ENGLISH BIBLE VERSIONS: F-MEASURE VS. NOISY CHANNEL EVAL. I





F-Measure: WBS, ASV, DBY, WEB, YLT, BBE

NCE: WBS, ASV, DBY, WEB, BBE, YLT

INTERDISCIPLINARY CONCEPT OF ETRAP

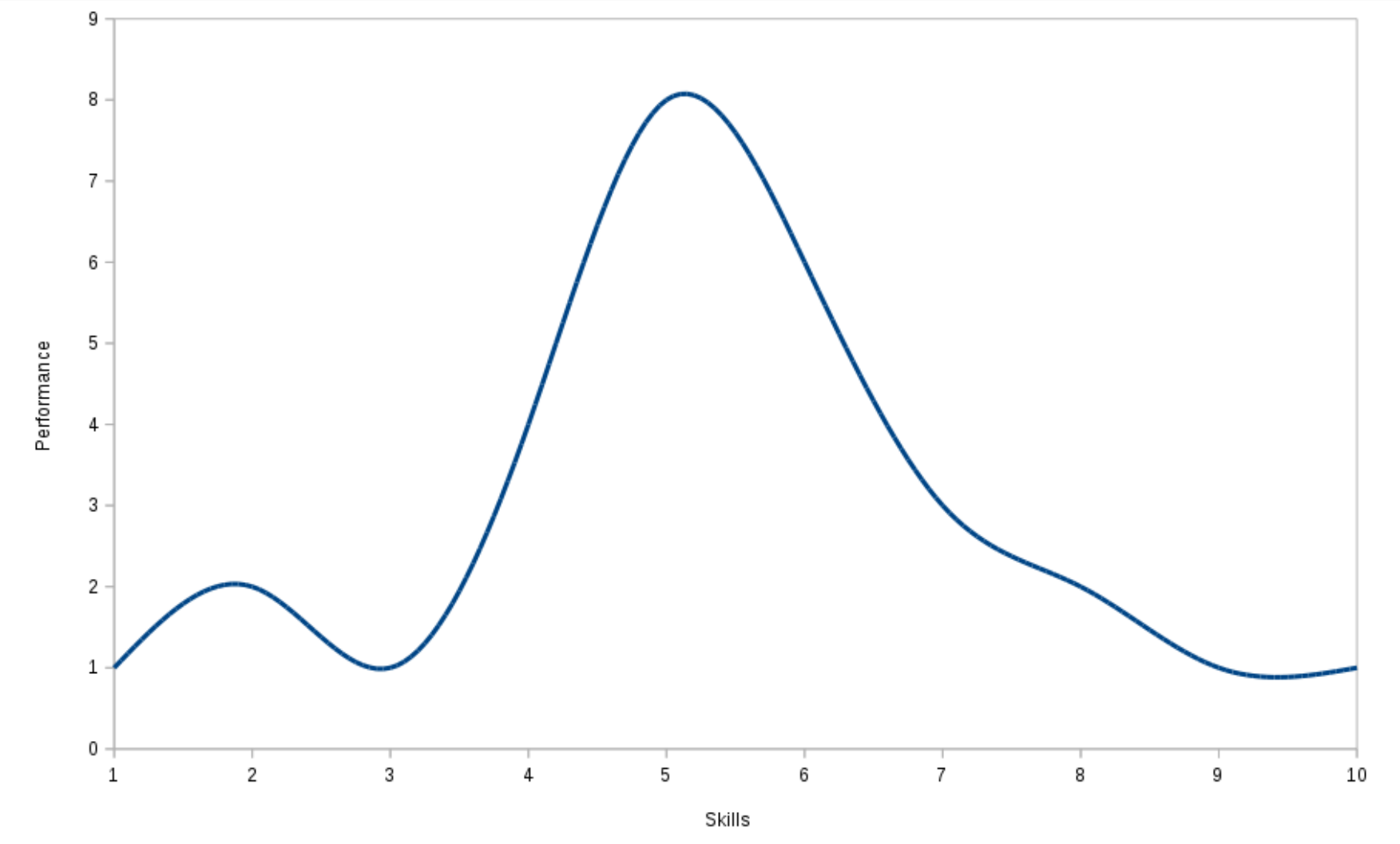
PROFESSIONAL TEAM COACHING OF ETRAP

Professional team coaching for effective group dynamic:

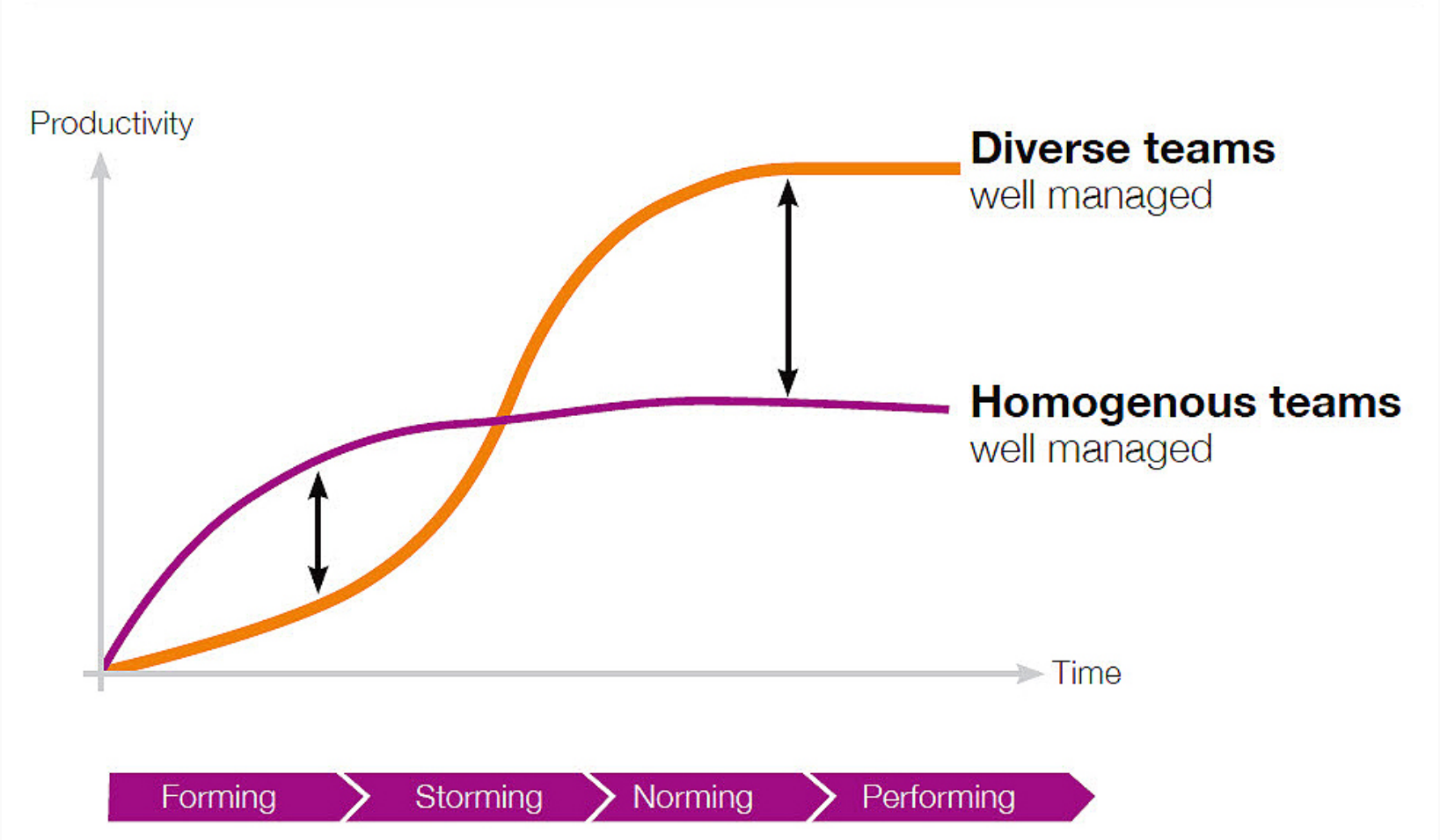
- Effective communication;
- Making the most of strengths;
- Effective delegation.

coach evolve lead
inspeyer

STRENGTHEN YOUR STRENGTHS OR YOUR WEAKNESSES?

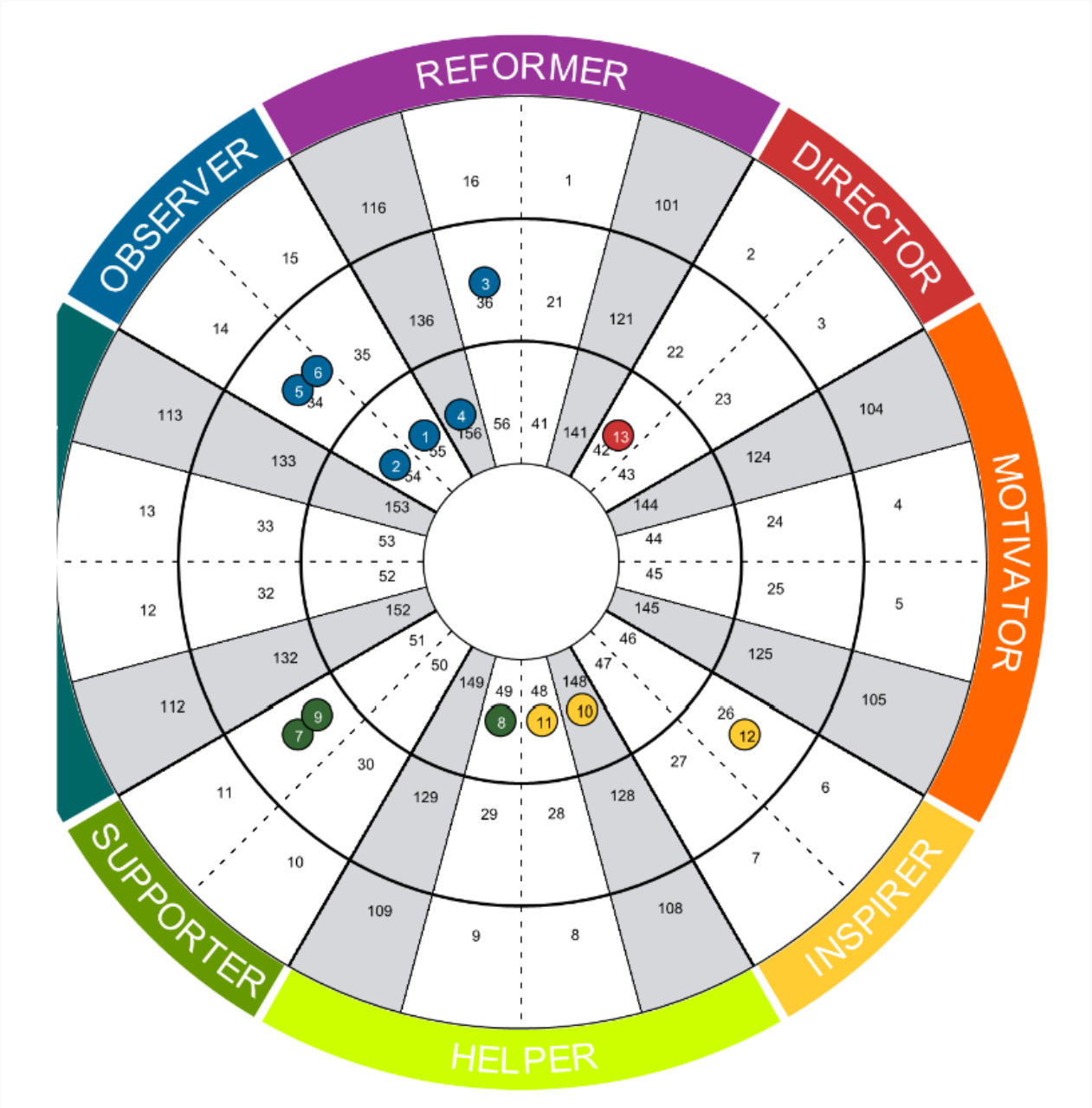


BUILDING A HIGH PERFORMANCE TEAM



TEAM TRAINING WITH PERSONALITY PROFILES

BUILDING A HIGH PERFORMANCE TEAM BY DIVERSITY OF SKILLS



CONTACT

Speaker

Emily Franzini & Marco Büchler.

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

eTRAP
Eleotronio Text Reuse Aoquisition Projeot
INSTITUTE OF COMPUTER SCIENCE
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

LICENCE