

# DETECTING HISTORICAL TEXT REUSE

## FROM A RESEARCH QUESTION TO THE RIGHT MODEL FOR DETECTING HISTORICAL TEXT REUSE

---

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



# 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.



## 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 IS TEXT REUSE?**

---

## Text Reuse:

- spoken and written repetition of text across time and space.

## For example:

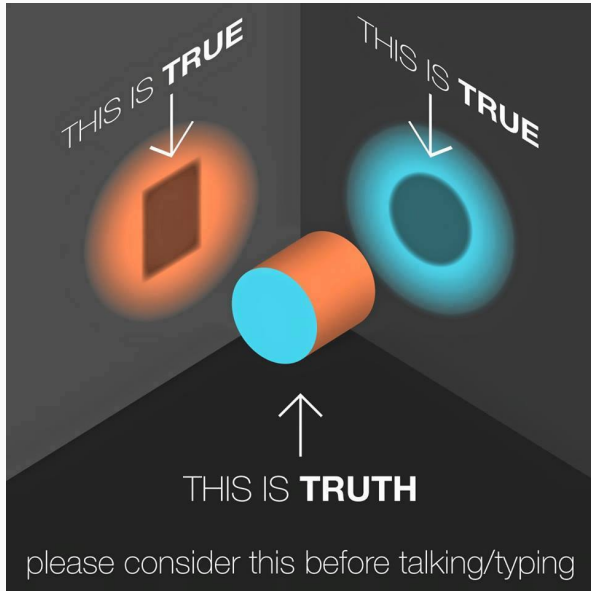
- citations, allusions, translations.

Detection methods are needed to support scholarly work.

- E.g. they help to ensure clean libraries or identify fragmentary authors.

Text is often modified during the reuse process.

# WHAT DO YOU ASSOCIATE WITH TEXT REUSE AND INTERTEXTUALITY?



# EXPECTATIONS OF A COMPUTER SCIENTIST: OVERSIMPLIFICATION



## EXPECTATIONS OF A HUMANIST: OVERSIMPLIFICATION



## Question:

Why is text reuse detection relevant for Humanities and Computer Science?

- **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.
  - Text Mining, Corpus Linguistics.

## “REUSE FROM SAME SOURCE”: COMMONALITIES & DIFFERENCES





# WITTGENSTEIN'S "FAMILY RESEMBLANCE"

Family resemblance is an equivalence relation that clusters common objects of similar and not identical characteristics together.

Family resemblance is hierarchical such as in the examples before "Greta", "Franzini", "Human", "creature".

# ETRAP'S OBJECTIVE

**Title:** eTRAP - electronic Text Reuse Acquisition Project

**Premise:** Language is a changing system. Compared to biometry the volatility is much higher.

- Research on the **characteristics**
  - What are **good characteristics**?
  - Which characteristics are **stable** and which are **volatile** and therefore not helpful in the detection process?
- Research on the **reuse process**
  - Begins with: **Why** do we quote what we quote?
  - Passes by: If changes in the **reuse process** happen, why do they happen and what is the model behind (if one exists)?
  - Ends with: **Understanding** paraphrases and allusions

## 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.

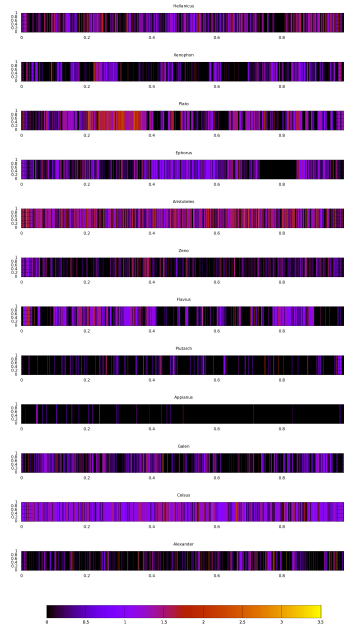
Ulrike Rieß (*Big Data bestimmt die IT-Welt*):

- **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;
- **Heterogeneous and distributed** data across resources.

**Information overload** = large amounts of data (Big Data).

**Information poverty** = noisy, fragmentary (Humanities Data).

# TEMPERATURE MAP



# RESEARCH ON THE CHARACTERIS- TICS

---

**Motif:** "1. A minimal thematic unit" (Prince, 2003, p. 55), set of **core elements**.

Core elements from an **interdisciplinary** standpoint:

- Literature: tracing **MOTIFS**
- **Cultural Studies**: tracing **MEMES**
- **Linguistics**: tracing **PATTERNS**
- **Computer Science**: tracing **FEATURES**
- **Forensics**: tracing **MINUTIAE**
- **Cognitive Psychology & Literature Studies**: tracing **FIGURES OF MEMORY**



# DATA COLLECTION AND CURATION

**Tasks:** Verify presence of motifs in different collections and record their “base form” as text reuse **training data**.

ISO Language Codes <a href="https://www.loc.gov/standards/iso639-2/php/code_list.php">https://www.loc.gov/standards/iso639-2/php/code_list.php</a>	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: 185098476	Calvino_1956 VIAF: 181208131	Jacobs_1892 VIAF: 315397813	Bruford_1994 VIAF: 12471835	Hooqasian- Villa_1966 VIAF: 186329063	Campbell_1958 VIAF: 25969242	Taylor_1823 VIAF: 59071527	Briggs_1970 VIAF: 46803237	El-Shamy_1969 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 1:** 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 2:** Grimm motifs reduced to keywords.



Train an (adapted) **Named Entity Recognition** (NER) tagger, ideally as language-independent as possible, to **automatically annotate** further fairy tales and texts.

## EXAMPLE CASE STUDY: SNOW WHITE

**RQ:** How to computationally **detect** a motif despite its **variants**?

For example:

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

**A:** We **strike a balance between precision and recall**. That is, finding the balance between a specific motif (Aarne-Thompson-Uther index) and its ontological root (Propp's typological unity).

**HOW?**

## The NRC (National Research Council Canada) Emotion Lexicon:

- The Roget Thesaurus
- 14,182 words types

### Emotions: (Plutchik, 1980)

anger  
anticipation  
disgust  
fear  
joy  
sadness  
surprise  
trust

### Sentiments:

negative emotions  
positive emotions

## Classroom Questionnaires



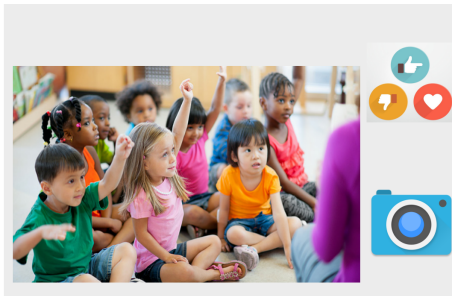
- Empathy
- Identification
- Transportation



- Six- and ten-year-old children
- Y-Labor



- Data set



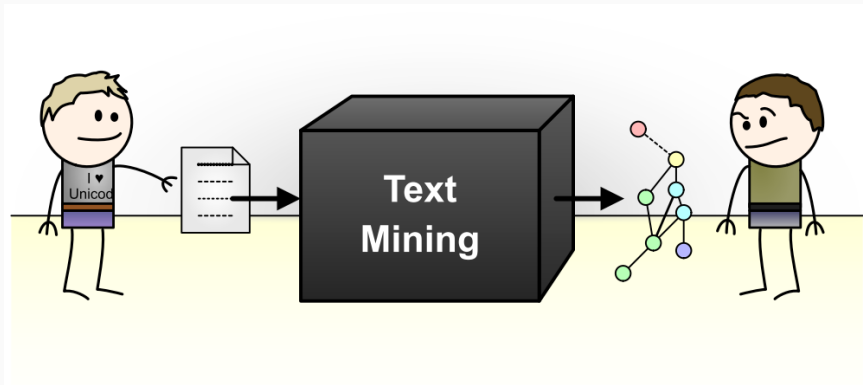
## ACID PARADIGM

---

## ACID for the Digital Humanities:

- Acceptance
- Complexity
- Interoperability
- Diversity

# ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE I





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





**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.

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.

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

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

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

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

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.

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

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.

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

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.

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.

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:

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.

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

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.

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.

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

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

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](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

[Trace](#)

prev 0 1 2 3 4 5 6 ... 1546 next

# ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE V

## Step 0: Searching

## Step 1: Preprocessing

Please select a preprocessing strategy:

01:02-WLPitem=true\_syn=false\_ssim=false\_redwo=false\_gram=5:LLR=true\_toLC=true\_rDia=false\_w2wl=false\_wlt=5

change

**Unprocessed Sentence:**

In the beginning God created the heavens and the earth.

**Preprocessed Sentence:**

in the begin god create the heaven and the earth .

correct

Your correction for the processed sentence:

in the begin god create the heaven and the earth .

Your comment:

submit changes

### Other users preference

No users have suggested a change in the preprocessing level

next level

# ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE VI

▣ Step 0: Searching

▣ Step 1: Preprocessing

▣ Step 2: Featurizing

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

next Level

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 .

# ACID FOR THE DIGITAL HUMANITIES: ACCEPTANCE VII

## Step 3: Selecting

Please select a selecting strategy:

### 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

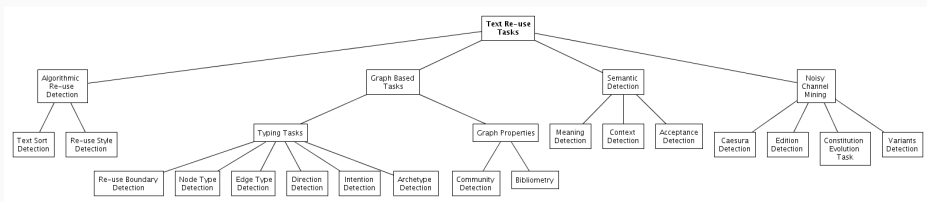
### Other users preference

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

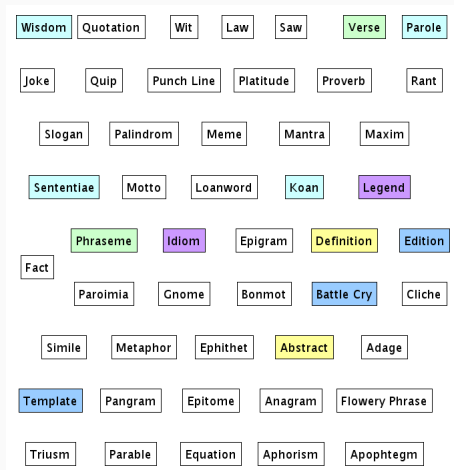
Feature	Selected Features	Total number of features
in the	27114	32227
the begin	470	480
begin god	0	5
god create	27	45
create the	17	38
the heaven	1624	1695
heaven and	389	396
and the	31908	40650
the earth	4776	5222
earth .	1030	1040

# ACID FOR THE DIGITAL HUMANITIES: COMPLEXITY



cit-quote-bibl	blockquote	bibl without quote
<pre> &lt;cit&gt;   &lt;quote&gt;     du/o ku/nes a)rgoi\     ei(/ponto   &lt;/quote&gt;   &lt;bibl n="Hom. Od. 2.11"&gt;     Od. 2.11   &lt;/bibl&gt; &lt;/cit&gt; </pre>	<pre> &lt;quote rend="blockquote"&gt;   &lt;line&gt;     a)gxou= d' i(stame/nh e)/pea     ptero/enta proshu/da   &lt;bibl n="Hom. Il. 4.92"&gt;Il. 4.92&lt;/bibl&gt;   &lt;/line&gt;&lt;line&gt;     a)ll' a)/ge nu=n ma/stiga kai\     h(ni/a sigalo/enta   &lt;bibl n="Hom. Il. 5.226"&gt;Il. 5.226&lt;/bibl&gt;   &lt;/line&gt; &lt;/quote&gt; </pre>	<pre> &lt;p&gt;   [...]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    &lt;bibl n="Hom. Od. 13.57"&gt;Od.   13.57&lt;/bibl&gt;   [...] &lt;/p&gt; </pre>

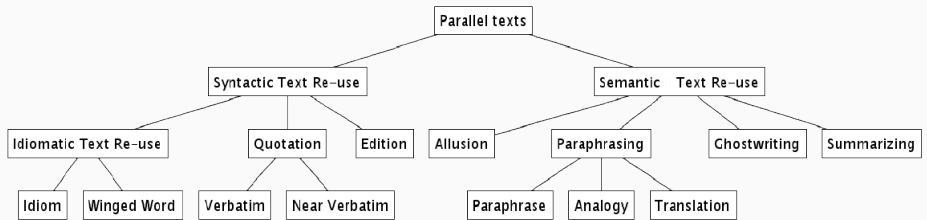
# DIVERSITY (REUSE TYPES)



- **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?

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

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

**Upcoming tutorials:**

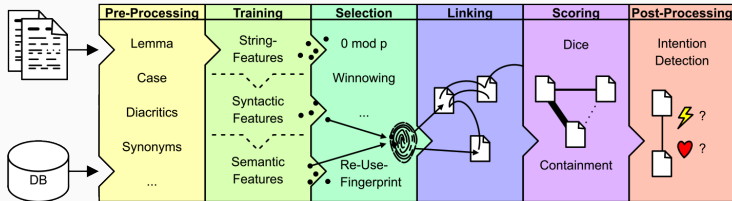
- **DATECH 2017** (May 2017): pre-conference workshop, Göttingen, Germany.
- Three more tutorials in 2017 pending confirmation.

## **COMPARISON OF LUKE & MARK**

---

# TRACER: OVERVIEW

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



**Figure 3:** 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.

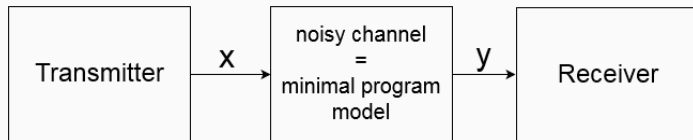
## **REUSE PROCESS**

---

Paraphrasing and non-literal reuse challenges many approaches:

- Alzahrani et al. (2012)
  - study n-gram-, syntax-, and semantic-based detection approaches;
  - they find: as soon as reuse is slightly modified (words changed) most approaches fail.
- Barrón-Cedeño et al. (2013)
  - experiment with paraphrasing to improve plagiarism detection;
  - they found that complex paraphrasing with a high density challenges plagiarism detection, and
  - that lexical substitution is the most frequent plagiarism technique.

- Inspired by
  - **Noisy channel model**: given a “scrambled” word or sentence, guess the intended version of that sentence (Brill, 2000),
  - **Kolmogorov Complexity**: describes the length of the shortest program that produces an output string (Li and Vitáni, 2008),
- we study Ancient text reuse to understand how text is transferred.
  - **Identify** operations to characterize morphological & semantic changes
  - **Design** an algorithm which applies these OPs to our datasets
  - **Transform** one text excerpt into another by a minimum OP set





# DATA-SETS - ANCIENT GREEK AND LATIN DATA-SET

“Salvation for the Rich”

**Clement of Alexandria**

Christian theologian, 2nd cent.

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

The data was tokenized and punctuation was kept but ignored in the analyses.

Extracts from 12 works & 2 collections

**Bernard of Clairvaux**

French abbot, 12th cent.

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

# BIBLICAL REUSE EXAMPLES

more literal	Bible verse	Bernard reuse
Proverbs 18 3	<b>impius cum in profundum venerit</b> peccatorum <b>contemnit</b> sed sequitur eum ignominia et obprobrium ( <i>When the wicked man is come into the depth of sins, also contempt comes but ignominy and reproach follow him</i> )	<b>Impius , cum venerit in profundum</b> malorum , <b>contemnit</b> ( <i>When the wicked man is come into the depth of evil</i> )
less literal	Bible verse	Clement reuse
1Cor 13 13	νυνὶ δὲ μένει πίστις , ἐλπίς , ἀγάπη , τὰ τρία ταῦτα μείζων δὲ τούτων ἡ ἀγάπη ( <i>And now remain faith, hope, love, these three; but the greatest of those is love.</i> )	πίστει καὶ ἐλπίδι καὶ ἀγάπῃ ( <i>faith, and hope, and love - in dative case</i> ) ἀγάπῃν , πίστιν , ἐλπίδα ( <i>love, faith, hope - in accusative case</i> ) μένει δὲ τὰ τρία ταῦτα , πίστις , ἐλπίς , ἀγάπη · μείζων δὲ ἐν τούτοις ἡ ἀγάπη ( <i>and remain these three, faith, hope, love; but the greatest among them is love</i> )
non-literal	Bible verse	Clement reuse
Mt 12 35	ὁ ἀγαθὸς ἄνθρωπος ἐκ τοῦ ἀγαθοῦ θησαυροῦ ἐκβάλλει ἀγαθὰ , καὶ ὁ <b>πονηρὸς</b> ἄνθρωπος ἐκ τοῦ πονηροῦ θησαυροῦ ἐκβάλλει <b>πονηρά</b> . ( <i>A good man out of good storage brings out good things , and an evil man out of the evil storage brings evil things .</i> )	Ψυχῆς , τὰ δὲ ἐκτός , κἂν μὲν ἡ ψυχὴ χρητῇ καλῶς , καλὰ καὶ ταῦτα δοκεῖ , ἐὰν δὲ <b>πονηρῶς</b> , <b>πονηρά</b> , ὁ κελεύων ἀπαλλοτριοῦν τὰ ὑπάρχοντα ( <i>[are whithin the] soul, and some are out, and if the soul uses them good, those things are also thought of as good, but if [they are used as] bad, [they are thought of as] bad; he who commands the renouncement of possessions</i> )

We aggregate:

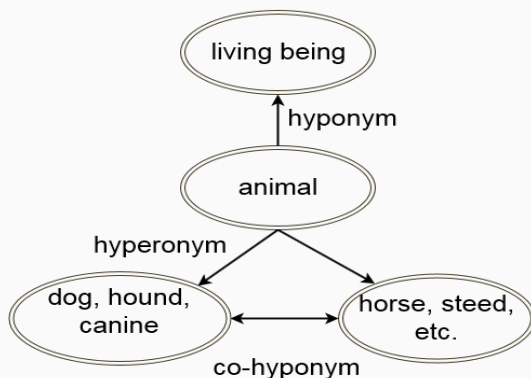
- **Bibindex' Lemma Lists**
  - 65,537 Biblical Greek entries
  - 315,021 Latin entries
- **Classical Language Tool Kit (CLTK)** (Johnson et al., 2014)
  - 953,907 Ancient Greek words
  - 270,228 Latin words
- **Greek New Testament of the Society of Biblical Literature<sup>1</sup> & Septuaginta** (Rahlfs, 1935a; UPenn) 59,510 word-lemma-pairs

---

<sup>1</sup> Logos Bible Software <http://sblgnt.com/about/>

99K synsets

of which 33K contain Ancient Greek and 27K Latin words  
(Bizzoni et al., 2014; Minozzi, 2009)



**Table 1:** Operation list for the automated approach

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

We manually analyze:

- 60 Ancient Greek & 100 Latin instances
- 192 & 224 replacements
- Using `ins(word)`, `del(word)` and replacements:
  - NOP, `lem`, `repl_syn`,  
`repl_hyper`, `repl_hypo`,  
`repl_co-hypo`
- We assign morphological categories from Perseus' tag-set (Bamman and Crane 2011)
  - E.g., `repl_case_a_g`  
`repl_num_s_p`

Table 2: Excerpt from Perseus' tag-set

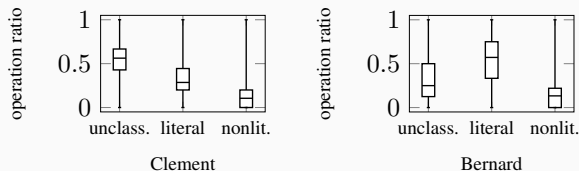
Category	Value	Tag
person	first person	1
	second person	2
	third person	3
number	singular	s
	plural	p
	dual	d
tense	present	p
	imperfect	i
	perfect	r
	pluperfect	l
	future perfect	t
	future	f
	aorist	a

## RESULTS

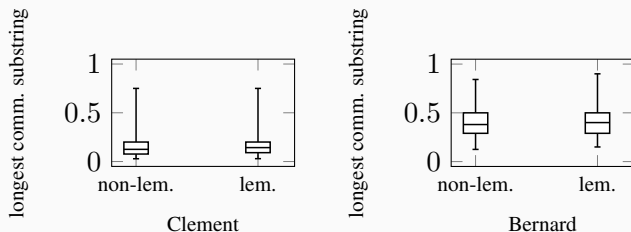
---

# LITERAL SHARE OF THE REUSE (RQ1)

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



**Figure 4:** Ratios of operations in reuse instances. **literal:** NOP, lem, lower, etc.; **nonlit:** syn, hyper, etc.



**Figure 5:** Ratios of literal overlap between reuse instances and originals.



How is the non-literally reused text modified in our datasets? (RQ2)

How can linguistic resources support the discovery of non-literal reuse?  
(RQ2.1)

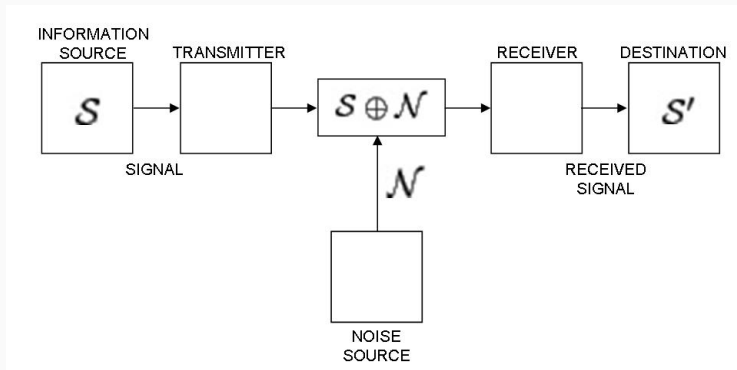
**Table 3:** Absolute numbers of operations identified automatically.

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

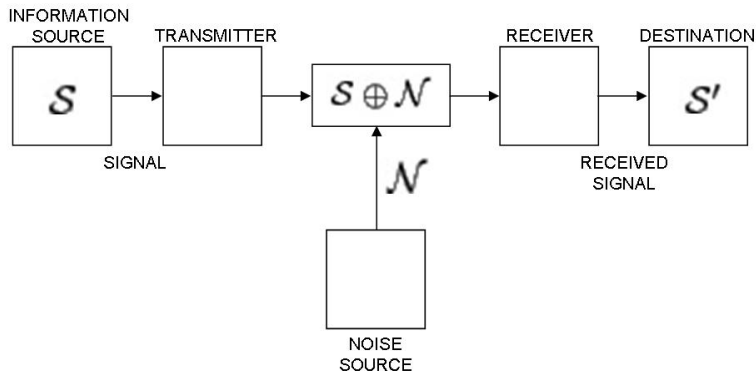
# **AUTOMATIC EVALUATION**

---

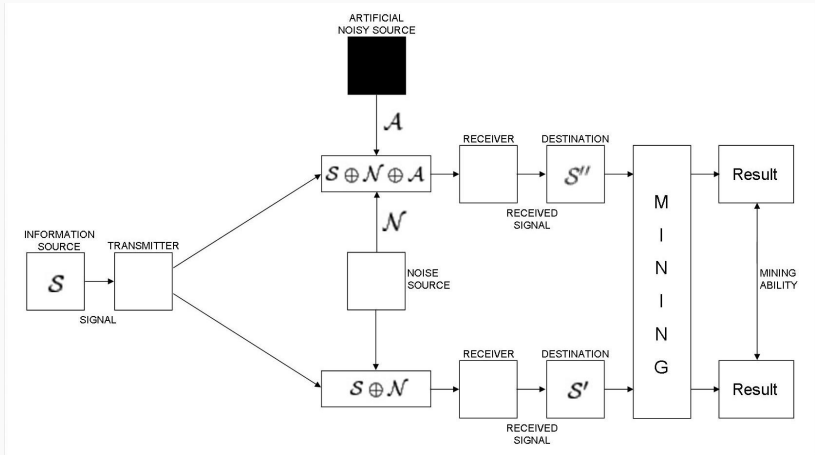
**Basic idea:** Embed historical text reuse in Shannon's **Noisy Channel** theorem.



**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.

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

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

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

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

**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.

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

## 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.

1. eTRAP works on text reuse.
2. eTRAP works on text reuse.
3. eTRAP works on text reuse.
4. eTRAP works on text reuse.
5. eTRAP works on text reuse.
6. ...

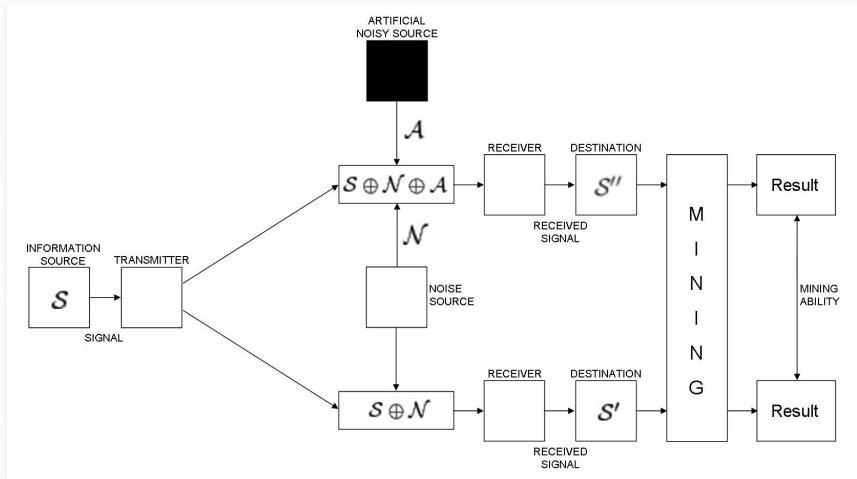
	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$
$s_1$	0.00	1.00	1.00	1.00	1.00
$s_2$	1.00	0.00	1.00	1.00	1.00
$s_3$	1.00	1.00	0.00	1.00	1.00
$s_4$	1.00	1.00	1.00	0.00	1.00
$s_5$	1.00	1.00	1.00	1.00	0.00

$$C_{\Theta} = \frac{n \cdot (n - 1)}{n^2} = 1 - \frac{1}{n}$$

$$C_{\Theta} = \frac{\sum_{j=1}^m \sum_{i=1}^n \theta_{\Theta}(s_i, s_j)}{n * m}$$

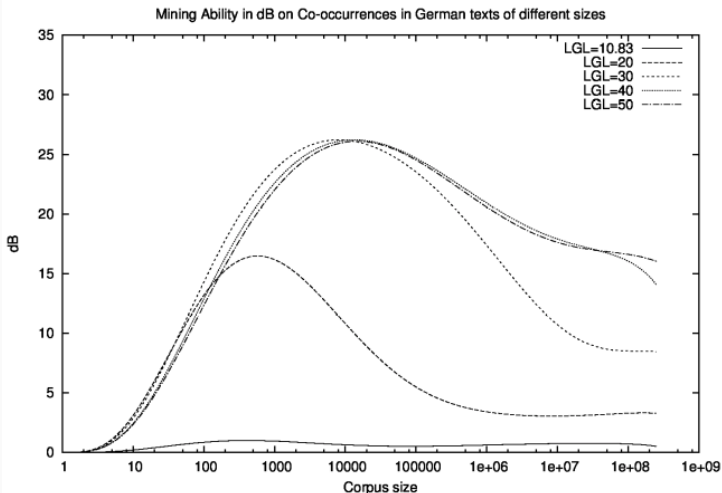


# RANDOMNESS & STRUCTURE



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

# RANDOMNESS & STRUCTURE: IMPACT



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

**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

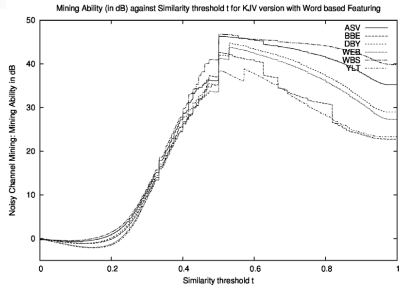
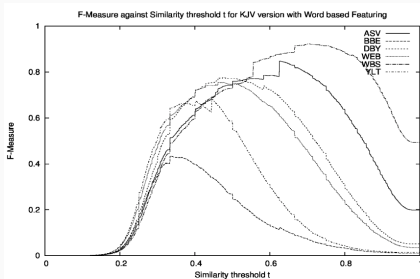
With

	Trigram Shingling				Bigram Shingling				Word based Featurings			
	S <sub>11</sub>	S <sub>12</sub>	S <sub>13</sub>	S <sub>14</sub>	S <sub>21</sub>	S <sub>22</sub>	S <sub>23</sub>	S <sub>24</sub>	S <sub>31</sub>	S <sub>32</sub>	S <sub>33</sub>	S <sub>34</sub>
ASV vs. BBE	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.09	0.30	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	0.72	0.74
ASV vs. KJV	0.36	0.38	0.37	0.38	0.51	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	0.70	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	0.01	0.03	0.03	0.03	0.18	0.21	0.21	0.22
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	0.72	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	0.40	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	0.21	0.22
KJV vs. ASV	0.36	0.38	0.37	0.38	0.51	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	0.55	0.55
KJV vs. WBS	0.75	0.78	0.76	0.77	0.80	0.81	0.80	0.80	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	0.70	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	0.40	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	0.55	0.55
WEB vs. WBS	0.11	0.12	0.11	0.12	0.20	0.22	0.21	0.21	0.59	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	0.13	0.15
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	0.80	0.81	0.80	0.80	0.99	0.99	0.99	0.99
WBS vs. WEB	0.21	0.22	0.21	0.22	0.25	0.26	0.25	0.25	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	0.01	0.03	0.03	0.03	0.18	0.21	0.21	0.22
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	0.21	0.22
YLT vs. KJV	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.10	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 Featurings			
	S <sub>11</sub>	S <sub>12</sub>	S <sub>13</sub>	S <sub>14</sub>	S <sub>21</sub>	S <sub>22</sub>	S <sub>23</sub>	S <sub>24</sub>	S <sub>31</sub>	S <sub>32</sub>	S <sub>33</sub>	S <sub>34</sub>
ASV vs. BBE	0.01	0.15	0.36	0.18	0.02	0.01	0.01	0.01	5.99	5.42	5.39	5.33
ASV vs. DBY	5.23	5.19	5.20	5.19	4.98	4.96	4.97	4.96	4.99	4.98	4.98	4.98
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	5.01	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	5.48	4.56	4.56	4.56
ASV vs. YLT	0.34	0.26	0.30	0.29	0.08	0.01	0.05	0.01	5.98	4.95	4.92	4.91
BBE vs. ASV	0.16	0.15	0.16	0.18	0.02	0.01	0.01	0.01	5.99	5.42	5.39	5.33
BBE vs. DBY	0.42	0.36	0.41	0.41	0.24	0.20	0.22	0.20	5.51	5.47	5.44	5.42
BBE vs. KJV	0.35	0.30	0.34	0.32	0.00	0.97	0.99	0.97	5.26	5.23	5.00	4.98
BBE vs. WEB	0.17	0.16	0.17	0.18	0.01	0.00	0.60	0.01	5.30	5.27	5.26	5.22
BBE vs. WBS	0.75	0.74	0.75	0.74	0.55	0.54	0.55	0.54	4.94	4.93	4.83	4.82
BBE vs. YLT	0.86	0.77	0.84	0.85	0.68	0.62	0.66	0.66	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	4.96	4.99	4.98	4.98	4.98
DBY vs. BBE	0.42	0.36	0.41	0.41	0.24	0.20	0.22	0.20	5.51	5.47	5.44	5.42
DBY vs. KJV	0.49	0.45	0.46	0.44	0.21	0.18	0.19	0.18	4.72	4.70	4.70	4.69
DBY vs. WEB	0.69	0.65	0.67	0.65	0.42	0.39	0.40	0.38	4.85	4.82	4.82	4.80
DBY vs. WBS	0.49	0.45	0.46	0.44	0.21	0.18	0.19	0.18	4.72	4.70	4.70	4.69
DBY vs. YLT	0.38	0.31	0.33	0.32	0.15	0.08	0.69	0.07	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	0.35	0.30	0.34	0.32	0.00	0.97	0.99	0.97	5.26	5.23	5.00	4.98
KJV vs. DBY	0.49	0.45	0.46	0.44	0.21	0.18	0.19	0.18	4.72	4.70	4.70	4.69
KJV vs. WEB	0.57	0.52	0.55	0.55	0.31	0.27	0.29	0.28	4.81	4.78	4.78	4.78
KJV vs. WBS	4.61	4.61	4.63	4.62	4.55	4.51	4.54	4.54	4.41	4.41	4.41	4.41
KJV vs. YLT	0.39	0.33	0.39	0.39	0.16	0.09	0.15	0.14	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	0.17	0.16	0.17	0.18	0.01	0.00	0.60	0.01	5.30	5.27	5.26	5.22
WEB vs. DBY	0.69	0.65	0.67	0.65	0.42	0.39	0.40	0.38	4.85	4.82	4.82	4.80
WEB vs. KJV	0.57	0.52	0.55	0.55	0.31	0.27	0.29	0.28	4.81	4.78	4.78	4.78
WEB vs. WBS	0.52	0.48	0.51	0.50	0.26	0.22	0.24	0.23	4.75	4.72	4.72	4.72
WEB vs. YLT	0.38	0.30	0.34	0.33	0.23	0.16	0.17	0.16	5.31	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	0.75	0.74	0.75	0.74	0.55	0.54	0.55	0.54	4.94	4.93	4.83	4.82
WBS vs. DBY	0.49	0.45	0.46	0.44	0.21	0.18	0.19	0.18	4.72	4.70	4.70	4.69
WBS vs. KJV	4.61	4.61	4.63	4.62	4.55	4.51	4.54	4.54	4.41	4.41	4.41	4.41
WBS vs. WEB	0.52	0.48	0.51	0.50	0.26	0.22	0.24	0.23	4.75	4.72	4.72	4.72
WBS vs. YLT	0.25	0.22	0.24	0.24	0.06	0.02	0.04	0.08	5.35	5.29	5.23	5.21
YLT vs. ASV	0.34	0.26	0.30	0.29	0.08	0.01	0.05	0.01	5.98	4.95	4.92	4.91
YLT vs. BBE	0.86	0.77	0.84	0.85	0.68	0.62	0.66	0.66	5.99	5.94	5.92	5.92
YLT vs. DBY	0.38	0.31	0.33	0.32	0.15	0.08	0.69	0.07	5.26	5.19	5.13	5.10
YLT vs. KJV	0.39	0.33	0.39	0.39	0.16	0.09	0.15	0.14	5.41	5.33	5.28	5.26
YLT vs. WEB	0.38	0.30	0.34	0.33	0.23	0.16	0.17	0.16	5.31	5.44	5.36	5.33
YLT vs. WBS	0.25	0.22	0.24	0.24	0.06	0.02	0.04	0.08	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**

---

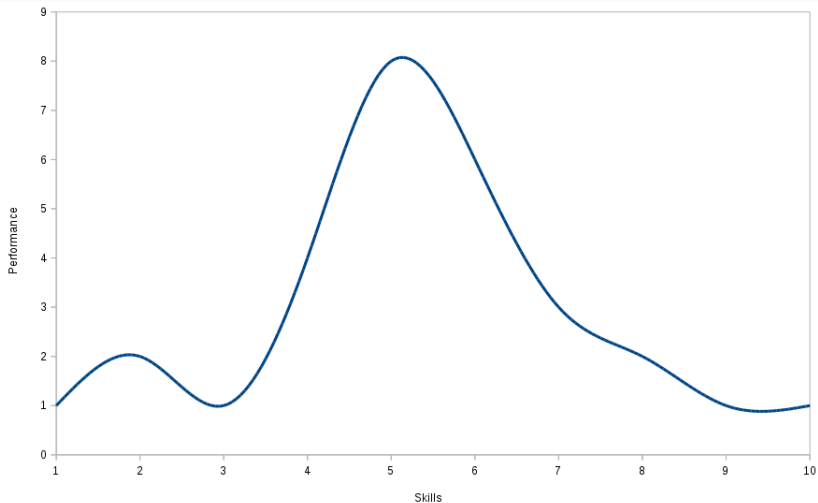
Professional team coaching for **effective group dynamic**:

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

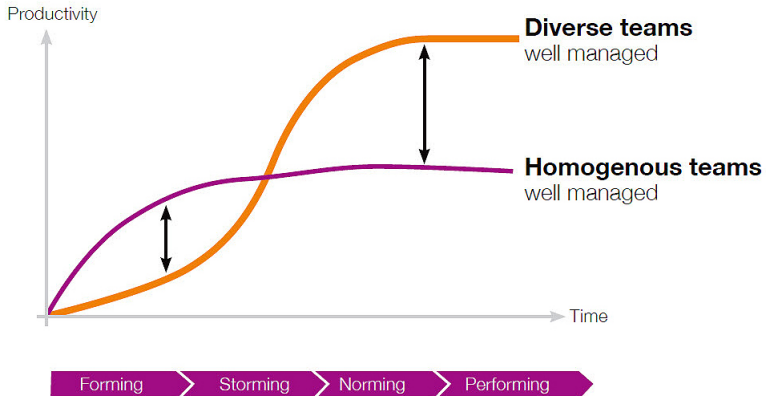




# STRENGTHEN YOUR STRENGTHS OR YOUR WEAKNESSES?



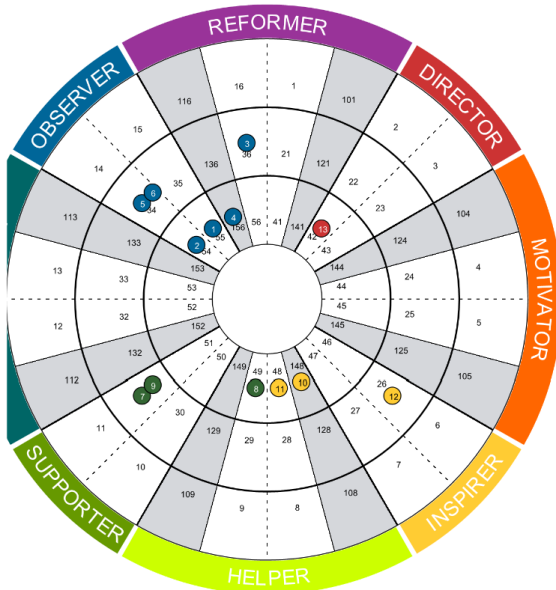
# BUILDING A HIGH PERFORMANCE TEAM



# TEAM TRAINING WITH PERSONALITY PROFILES



# BUILDING A HIGH PERFORMANCE TEAM BY DIVERSITY OF SKILLS



## Speaker

Marco Böhler.

## Visit us



<http://www.etrp.eu>



[contact@etrp.eu](mailto: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.

