GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Big Humanities Data: About the role of Automatic Natural Language Processing Techniques in the Digital Humanities

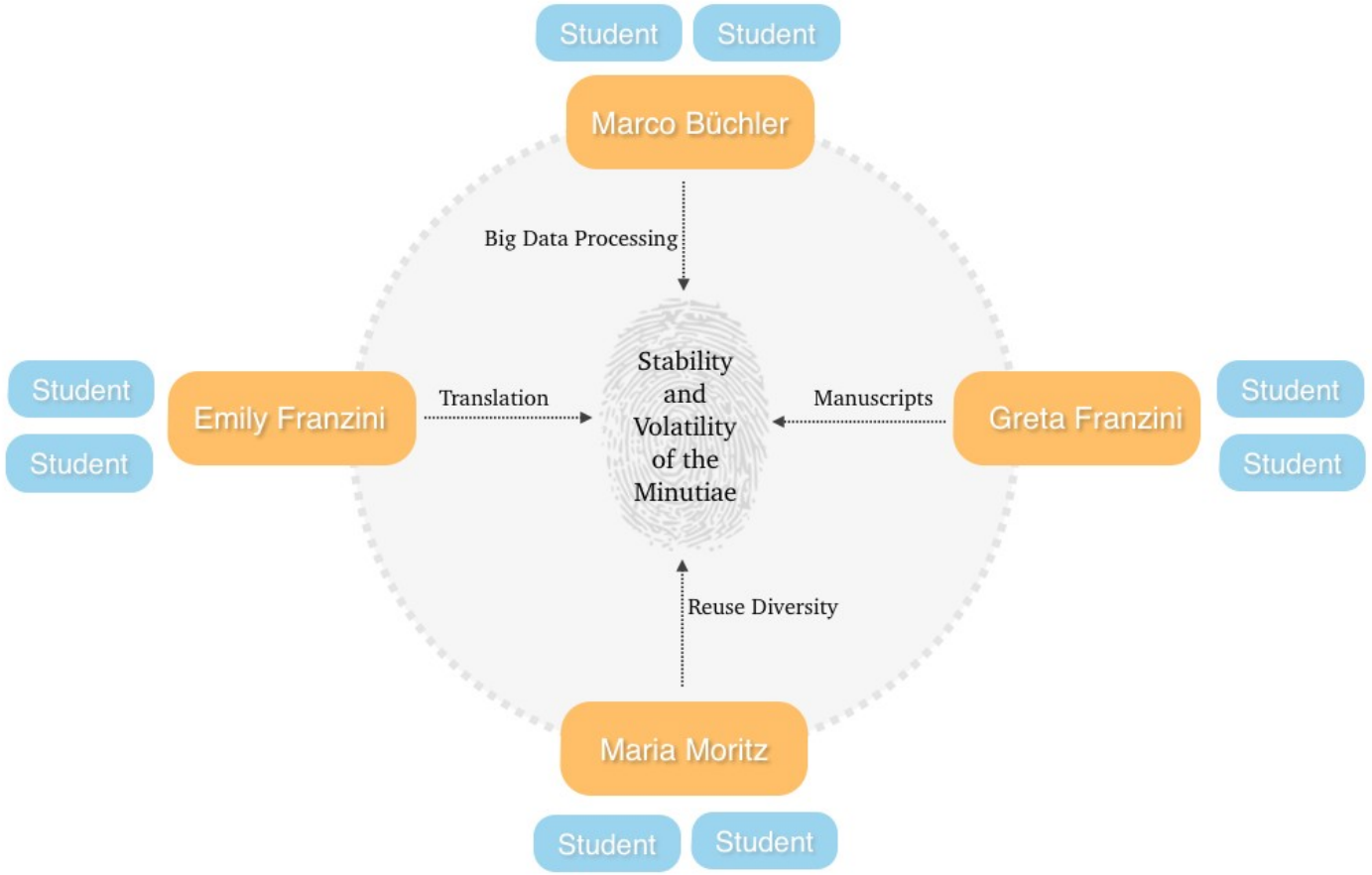
Marco Büchler
eTRAP Research Group
Göttingen Centre for Digital Humanities
Institute of Computer Science
Georg August University Göttingen, Germany

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Who am I?

- 2001/2 Head of Quality Assurance department in a software company
- 2006 Diploma in Computer Science on big scale co-occurrence analysis
- 2007- Consultant for several SME in IT sector
- 2008 Technical project management of eAQUA project
- 2011 PI and project manager eTRACES project
- 2013 PhD in „Digital Humanities“ on Text Reuse
- 2014- Head of Early Career Research Group eTRAP at Göttingen Centre for Digital Humanities

eTRAP – Electronic Text Reuse Acquisition Project



Overview

- Big Humanities Data
- Filling gaps in inscriptions
- Uncovering unexpected relations
- Text Reuse

Big (Humanities) Data

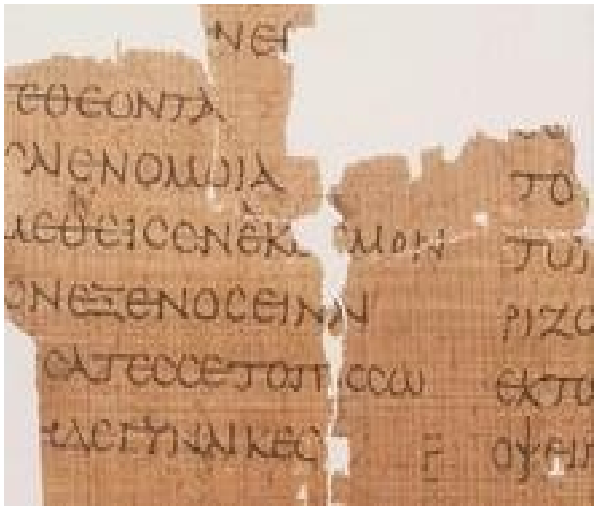
- **3 aspects** (by Ulrike Rieß, Big Data bestimmt die IT-Welt):
  - **Huge amount of data** that can‘t be processed and analyzed manually
  - **Less structured data**; e. g. in comparison to databases and data warehouse systems
  - Linked data between **heterogeneous and distributed resources**

- The fastest growing sources of Big Data are text and images.

- Researchers easily get lost in the **information overload** (Big Data) and in the **information poverty** (Humanities Data).

Filling (Guessing) gaps in inscriptions

The data



Textcorrection

Possible passage in the text:
Platon Timaios, 38c7 bis 38d4 (from TLG-Online):
σώματα δὲ αὐτῶν ἑκάστων ποιήσας ὁ θεὸς ἔθηκεν εἰς τὰς
@1 περιφορὰς ἃς ἡ θατέρου περίοδος ᾖειν, ἑπτὰ οὔσας ὄντα
(d.) ἑπτά, σελήνην μὲν εἰς τὸν περὶ γῆν πρῶτον, ἥλιον δὲ εἰς
τὸν δεύτερον ὑπὲρ γῆς, ἑωσφόρον δὲ καὶ τὸν ἱερὸν Ἑρμοῦ
λεγόμενον εἰς [τὸν] τάχει μὲν ἰσόδρομον ἡλίῳ κύκλον ἰόντας,τὴν δὲ ἐναντίαν εἰληχότας αὐτῷ δύναμιν·

Leiden convention

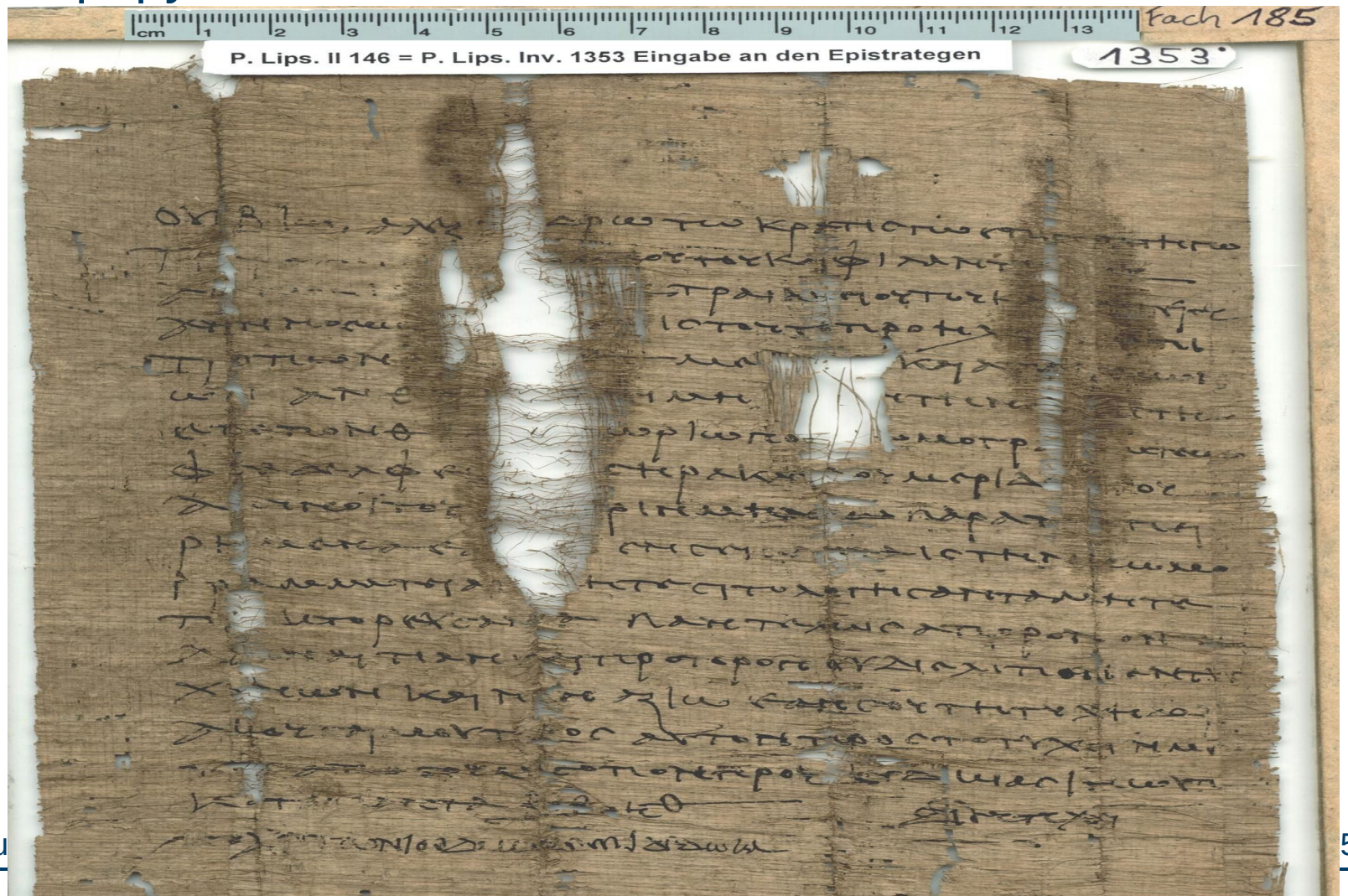
Textcorrection

Possible passage in the text:
Platon Timaios, 38c7 bis 38d4 (from TLG-Online):
σώματα δὲ αὐτῶν ἑκάστων ποιήσας ὁ θεὸς ἔθηκεν εἰς τὰς
@1 περιφορὰς ἃς ἡ θατέρου περίοδος ᾖειν, ἑπτὰ οὔσας ὄντα
(d.) ἑπτά, σελήνην μὲν εἰς τὸν περὶ γῆν πρῶτον, ἥλιον δὲ εἰς
τὸν δεύτερον ὑπὲρ γῆς, ἑωσφόρον δὲ καὶ τὸν ἱερὸν Ἑρμοῦ
λεγόμενον εἰς [τὸν] τάχει μὲν ἰσόδρομον ἡλίῳ κύκλον ἰόντας,τὴν δὲ ἐναντίαν εἰληχότας αὐτῷ δύναμιν·

Detection of words by Leiden Conventions (Source: Wikipedia):

- **[abc]**: letters missing from the original text due to lacuna, but restored by the editor
- **<ab>**: characters erroneously omitted by the ancient scribe, restored by the editor
- **[[abc]]**: deleted letters

The papyrus


Fach 185
P. Lips. II 146 = P. Lips. Inv. 1353 Eingabe an den Epistrategen
1353

Transcribed data

Οὐιβίῳ **Ἀλεξά**[**ν**]**δρῳ** τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὐ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἐα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-
γραμματείαν [μ]ήτε σιτολογήσαντα μήτε
πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α.
δι᾽ ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ-
χάνων καὶ νῦν ἀξιῶ, ἐάν σου τῇ τύχῃ δόξ[ῃ],
ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με
τῆς ἀπὸ σοῦ [μι]σοπονήρου ἐγδ[ι]κίας, ἵν᾽ ὦ ὑπὸ [σ]οῦ
κατὰ πάντα βεβοηθ(ημένος). διευτύχει
Ἀντώνιος Δόμνος ἐπιδέδωκα.

Input form

Text

Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὗ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-

TLG PHI7 Epiduke Send

#	Sentence

Parsed input (parsed for Leiden conventions)

Text

```
Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὗ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-
```

TLG
PHI7
Epiduke
Send

#	Sentence
0	Οὐιβίῳ [_]λεξά[____] τῷ κρατίστῳ ἐπιστρατήγῳ παρὰ Ἀντ[ωνίου] [Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου Ἀντωνίο[υ] [Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ] [Στρα]τείου Ἀντινοέως
1	[οὐκ] [ἂν] εἰς τοῦτο προήχθ[η]ν ἐπι- τρόπων [μέγιστ]ε μέ[τριος] καὶ ἀπράγμων ὢν ἄνθρ[ωπος] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην ἐπεπόνθ[ειν] [ὑπὸ] Ὡρίωνο[ς] [κ]ωμογρα[μ]ματέως Φ[ι]λαδελφεί[ας] [τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ Ἀρσινοίτου
2	[οὗ] [χά]ριν μην[ύ]ω παρὰ τ[ὰ] [ἀ]πει- ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο- γραμματείαν [μ]ήτε σπολογήσαντα μήτε πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α
3	δι ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ- χάνων καὶ νῦν ἀξιῶ ἐάν σου τῇ τύχῃ δόξ[ῃ] ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με τῆς ἀπὸ σοῦ [μι]σοπονήρου ἐγδ[ι]κίας ἵν ὦ ὑπὸ [σ]οῦ κατὰ πάντα βεβοηθ(ημένος)
4	διευτύχει Ἀντώνιος Δόμνος ἐπιδέδωκα

Strategy 1: use only information of the word

Word length + „survived“ pattern

[_]λεξά[___]

Interpreted word : _λεξά____

Length : 9

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
Ἀλεξάνδρᾳ	2	1.0		1.0					
Ἀλεξάνρου	2	1.0		1.0					
Ἀλεξάνδρα	2	1.0		1.0					
Ἀλεξάνδρῳ	2	1.0		1.0					
Ἀλεξάρχου	2	1.0		1.0					
Ἀλεξάνδου	2	1.0		1.0					
ἐνεχάραξα	1	1.0							
ὁμολογει῎	1	1.0							
ὁλοκλήρον	1	1.0							

Strategy 2: use only of context information

Word bigrams + co-occurrences + classification

[_]λεξά[___]

Interpreted word : _λεξά____

Length : 9

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
νομοῦ	3					0.5	0.4	0.000	
Ἀλεξάνδρῳ	3					1.0	0.8	0.003	
ἀπόδος	2					0.5		0.001	
Δόμνου	2						0.8	0.040	
ἀνέτεινα	2						0.2	0.025	
Αὐρηλίου	2						0.4	0.000	
Ἀχιλλεῖ	2					0.5	0.2		
Ἑπτὰ	2						0.2	0.010	
διαδεχομένῳ	2						0.2	0.250	
Σεουηριανῷ	2					0.5	0.2		
Αἰγύπτου	2						0.4	0.001	
ἡγεμόνι	2						0.2	0.001	
Λικιννιανῷ	2						0.2	0.200	

The „real“ Strategy 2: use only of context information and removing any information about the damaged word

Reparsing

Text

Οὐιβίῳ [_] τῷ κρατίστῳ ἐπιστρατήγῳ
παρὰ Ἀντ[ωνίου Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου.
Ἀντωνίο[υ Ῥωμανο]ῦ Τραιανείου τοῦ κα[ὶ Στρα]τείου
Ἀντινοέως. [οὐκ ἂν] εἰς τοῦτο προήχθ[η]ν, ἐπι-
τρόπων [μέγιστ]ε, μέ[τριος] καὶ ἀπράγμων
ὢν ἄνθρ[ωπος,] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην
ἐπεπόνθ[ειν ὑπὸ] Ὡρίωνο[ς κ]ωμογρα[μ]ματέως
Φ[ι]λαδελφεί[ας τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ
Ἀρσινοίτου. [οὗ χά]ριν μην[ύ]ω παρὰ τ[ὰ ἀ]πει-
ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο-

TLG
PHI7
Epiduke
Send

#	Sentence
0	Οὐιβίῳ [_] τῷ κρατίστῳ ἐπιστρατήγῳ παρὰ Ἀντ[ωνίου] [Δ]όμνου τοῦ καὶ Φιλαντι[νό]ου Ἀντωνίο[υ] [Ῥ]ωμανο]ῦ Τραιανείου τοῦ κα[ὶ] [Στρα]τείου Ἀντινοέως
1	[οὐκ] [ἂν] εἰς τοῦτο προήχθ[η]ν ἐπι- τρόπων [μέγιστ]ε μέ[τριος] καὶ ἀπράγμων ὢν ἄνθρ[ωπος] εἰ μὴ [ὕβρι]ν τὴν μ[εγ]ίστην ἐπεπόνθ[ειν] [ὑπὸ] Ὡρίωνο[ς] [κ]ωμογρα[μ]ματέως Φ[ι]λαδελφεί[ας] [τῆ]ς Ἡρακλείδου μερίδο[ς] τοῦ Ἀρσινοίτου
2	[οὗ] [χά]ριν μην[ύ]ω παρὰ τ[ὰ] [ἀ]πει- ρημένα ἑα[υτὸ]ν ἐνσείσαντα εἰς τὴν κωμο- γραμματείαν [μ]ήτε σπολογήσαντα μήτε πρ[α]κτορεύσαντα παντελῶς ἄπορον ὄν[τ]α
3	δι ἣν αἰτίαν καὶ πρότερον οὐ διέλιπον ἐντυγ- χάνων καὶ νῦν ἀξιῶ ἐάν σου τῇ τύχῃ δόξ[ῃ] ἀκοῦσαί μου π[ρ]ὸς αὐτὸν πρὸς τὸ τυχεῖν με τῆς ἀπὸ σοῦ [μ]σοπονήρου ἐγδ[ι]κίας ἵν ὦ ὑπὸ [σ]οῦ κατὰ πάντα βεβοηθ(ημένος)
4	διευτύχει Ἀντώνιος Δόμνος ἐπιδέδωκα

Word bigrams + co-occurrences + classification

[_]

Interpreted word : _

Length : 1

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
νομοῦ	3					0.5	0.4	0.000	
Ἀλεξάνδρῳ	3					1.0	0.8	0.003	
ἀπόδος	2					0.5		0.001	
Δόμνου	2						0.8	0.040	
ἀνέτενα	2						0.2	0.025	
Αὐρηλίου	2						0.4	0.000	
Ἀχιλλεῖ	2					0.5	0.2		
Ἑπτὰ	2						0.2	0.010	
διαδεχομένῳ	2						0.2	0.250	
Σεουηριανῷ	2					0.5	0.2		
Αἰγύπτου	2						0.4	0.001	
ἡγεμόνι	2						0.2	0.001	
Λικιννιανῷ	2						0.2	0.200	

Strategy 3: Complete strategy

Multiple options

[ ]λεξά[ ]

Interpreted word : _λεξά____

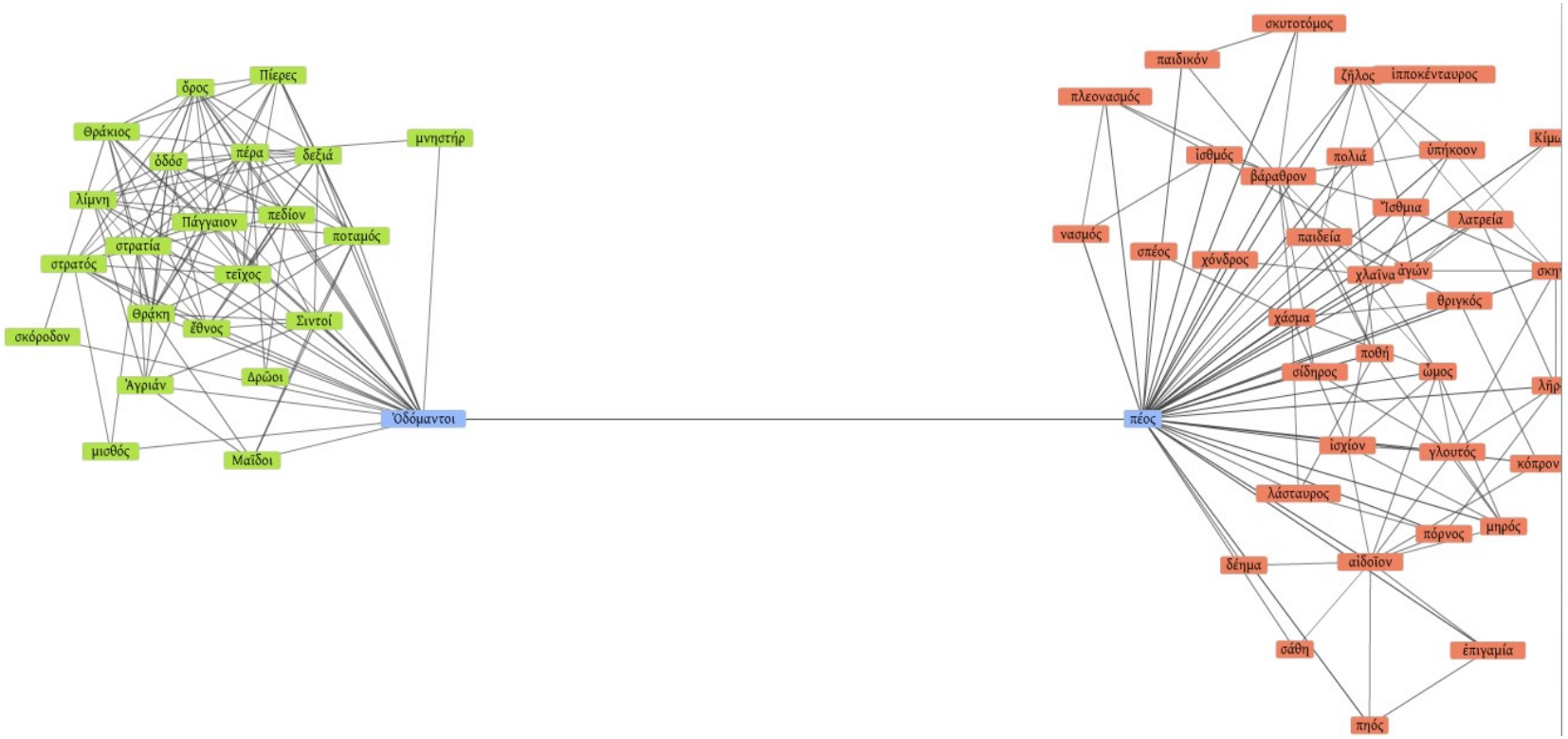
Length : 9

Candidate	Score	☐ Word length	☐ Neighboured letter bigrams	☐ Word similarity (letters)	☐ Named Entity	☐ Word bigram	☐ Semantic context	☐ Classification	Show
Ἀλεξάνδρῳ	5	1.0		1.0		1.0	0.8	0.003	
γενομένην	3	1.0				0.5	0.2		
διοίκησιν	3	1.0					0.2	0.008	
νομοῦ	3					0.5	0.4	0.000	
Ἀντινοέως	3	1.0					0.8	0.011	
βιβλιδίου	3	1.0					0.4	0.003	
ἐπιπρόπων	3	1.0					0.2	0.005	
Ἀντινοέων	3	1.0					0.4	0.002	
Δημητρίωι	3	1.0					0.2	0.002	
βιβλιδίων	3	1.0					0.2	0.002	
Στρατείου	3	1.0					0.6	0.214	
στρατηγὸς	3	1.0					0.2	0.001	
Ἀλεξάρχου	2	1.0		1.0					

Contrastive Semantics

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Visualisation of Contrastive semantics



Main properties

- **Contrast:**

$$contrast(w_i, w_j) = \begin{cases} 1 - sim_{dice}(w_i, w_j) & \text{if } sim_{dice}(w_i, w_j) \leq eps \\ 0 & \text{if } sim_{dice}(w_i, w_j) > eps \end{cases}$$

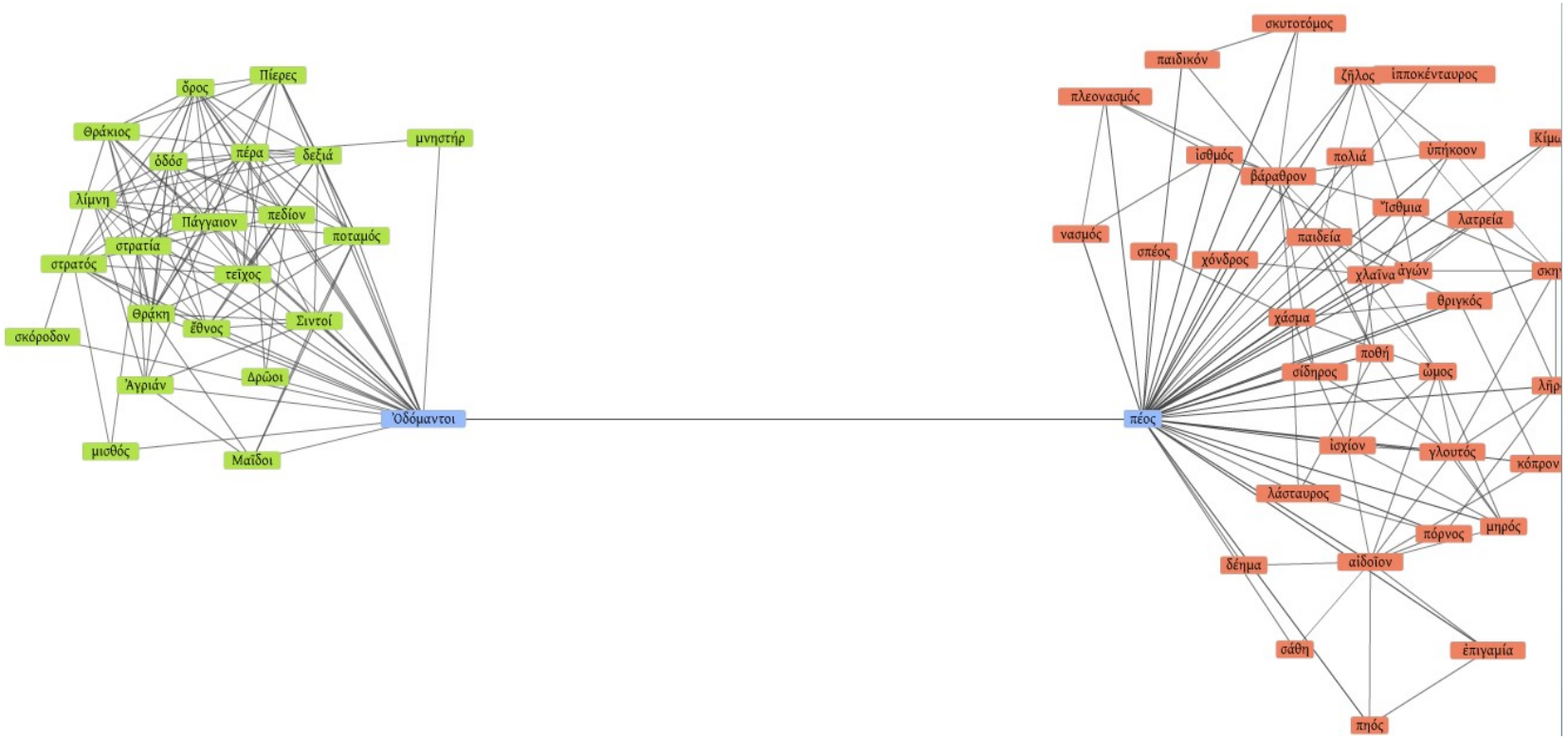
$$sim_{dice}(w_i, w_j) = 2 * \frac{|K_{w_i} \cap K_{w_j}|}{|K_{w_i}| + |K_{w_j}|}$$

- **Locality**:

$$dist(w_i, w_j) \leq eps_{dist} \text{ aus } (w_i, w_j) \in C$$

- **Frequency range of contrastive semantic relations:**
  - Generally less than 10 times of common occurrences

Connectivity?



Some observations

- **Identified clusters:**
  - As shown in examples comedy
  - Sarcasm
  - Cynicism
  - Artificial ambiguity like „Michael Schumacher the red king“ (translated from a German corpus)
  - Scope to gnomology

- **Is there a relation between contrastive semantics and textual reuse?**
  - Clearly, yes.
  - „Evaluation results“: More than 90% of the contrastive semantics have a relation to text reuse

Historical Text Reuse

An overview of text reuse

- **General**: Text reuse describes the spoken and written repetition of content.
- **Example**: quotations, paraphrases but also translations
- **Historical changes**: language evolution, different dialects, “spelling errors” but also copy errors (by scribes in the middle ages)

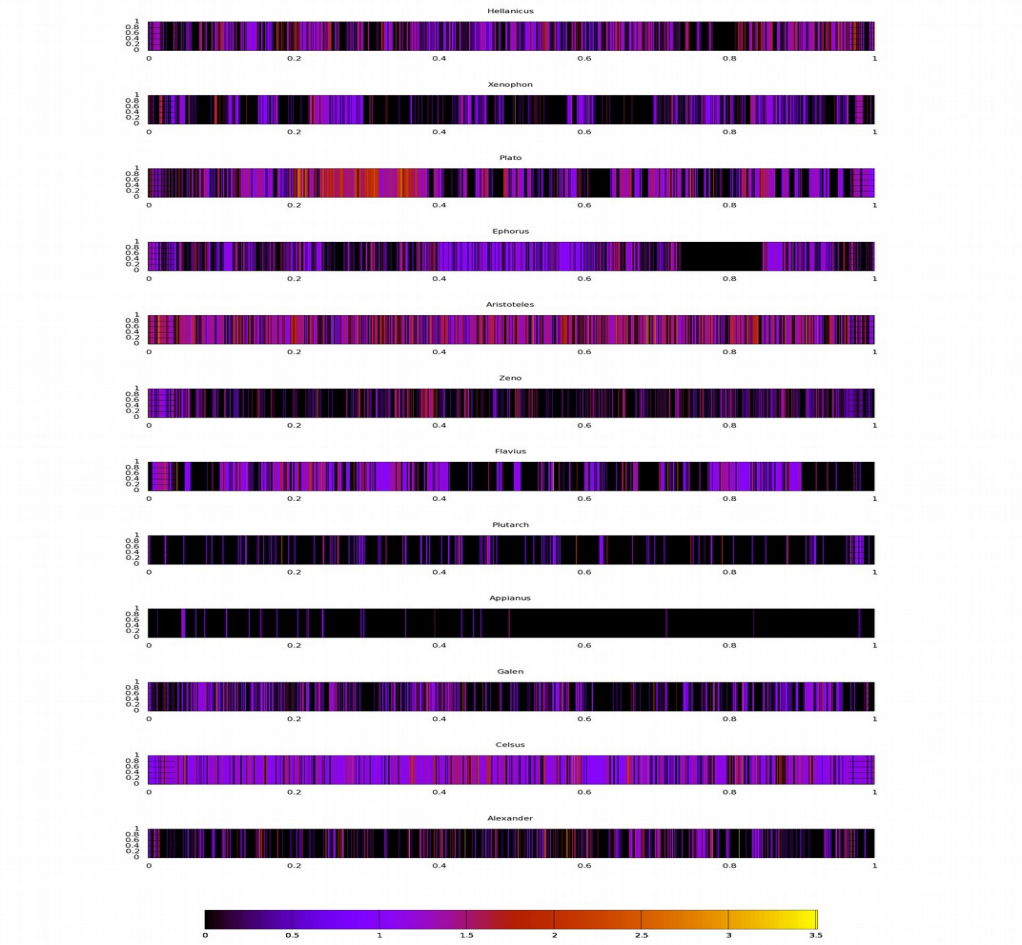
Text Reuse for Humanities and Computer Science

- **Question**: Why is Text Reuse so relevant for Humanities and Computer Science?
- **Premise**: The amount of digitally available data is growing exponentially (Big Data)

- Humanities:
  - Lines of transmission and textual criticism
  - Transmissions of ideas/thoughts under different circumstances and conditions

- Computer Science:
  - Text Decontamination for stylometry and authorship attribution, dating of texts
  - gen. Text Mining, Corpus Linguistics

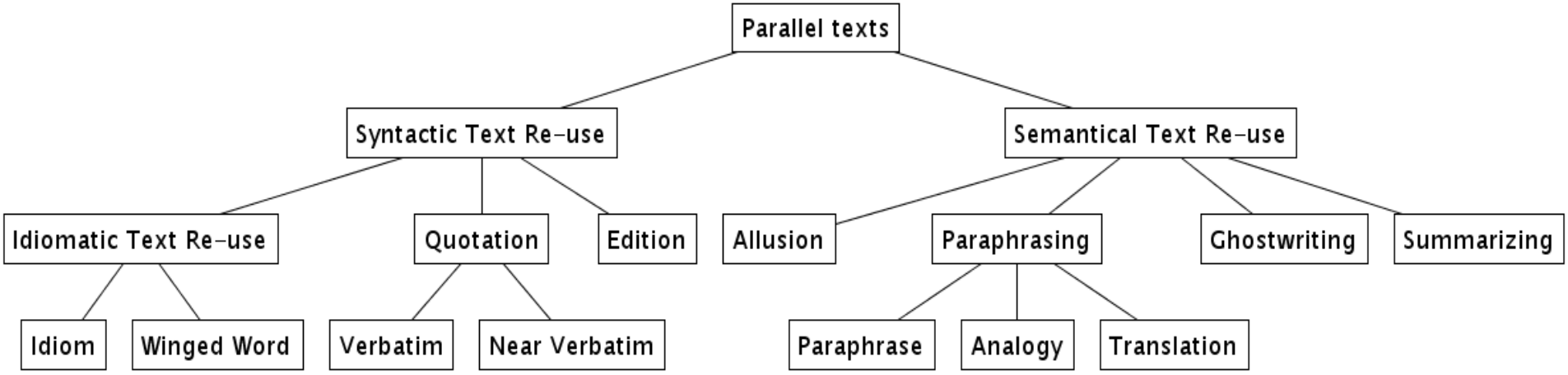
Temperature Map



ACID for the Digital Humanities – Diversity (Reuse Types)

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

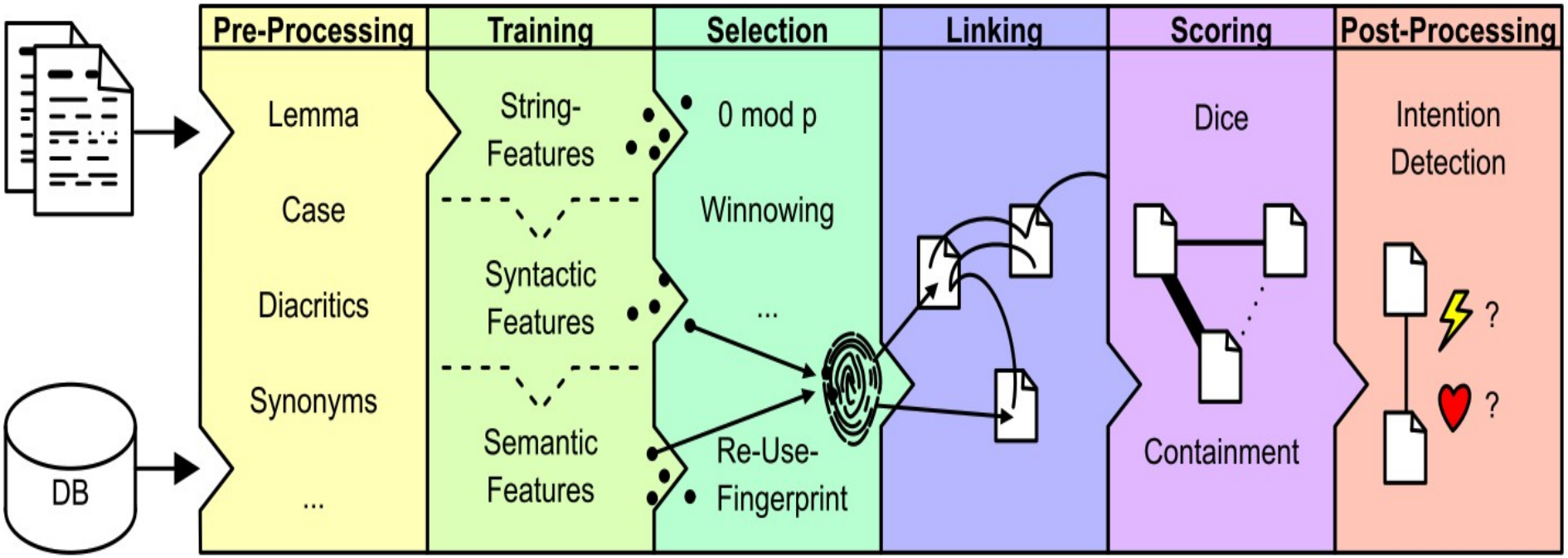
ACID for the Digital Humanities – Diversity (Reuse Styles)



Key problem

Basic question: *Distribution of Reuse Types und Reuse Styles are often unknown: Which model(s) should be chosen?*

Current approach



Text Reuse on English Bible versions

- Why the use of the Bible makes sense?:
  - The Bible is easy to evaluate.
  - There are different editions written for different purposes.

Text Reuse on English Bible versions with different intentions

- **American Standard Version** (ASV): 20th century, focus is USA
- **Bible in Basic English** (BBE): Verses are written in a simplified language
- **Darby** Version (DBY): Created in the 19th century from Hebrew and Greek texts, multiple authors through death of Darby
- **King James Version** (KJV): One of the oldest English Bible versions (16th Cent.)
- **Webster's Revision** (WBS): Revision of KJV in 19th century
- **World English Bible** (WEB): 21st century, global focus
- **Young Literal Translation** (YLT): Verses in Hebrew syntax

Text Reuse on English Bible versions Evaluation

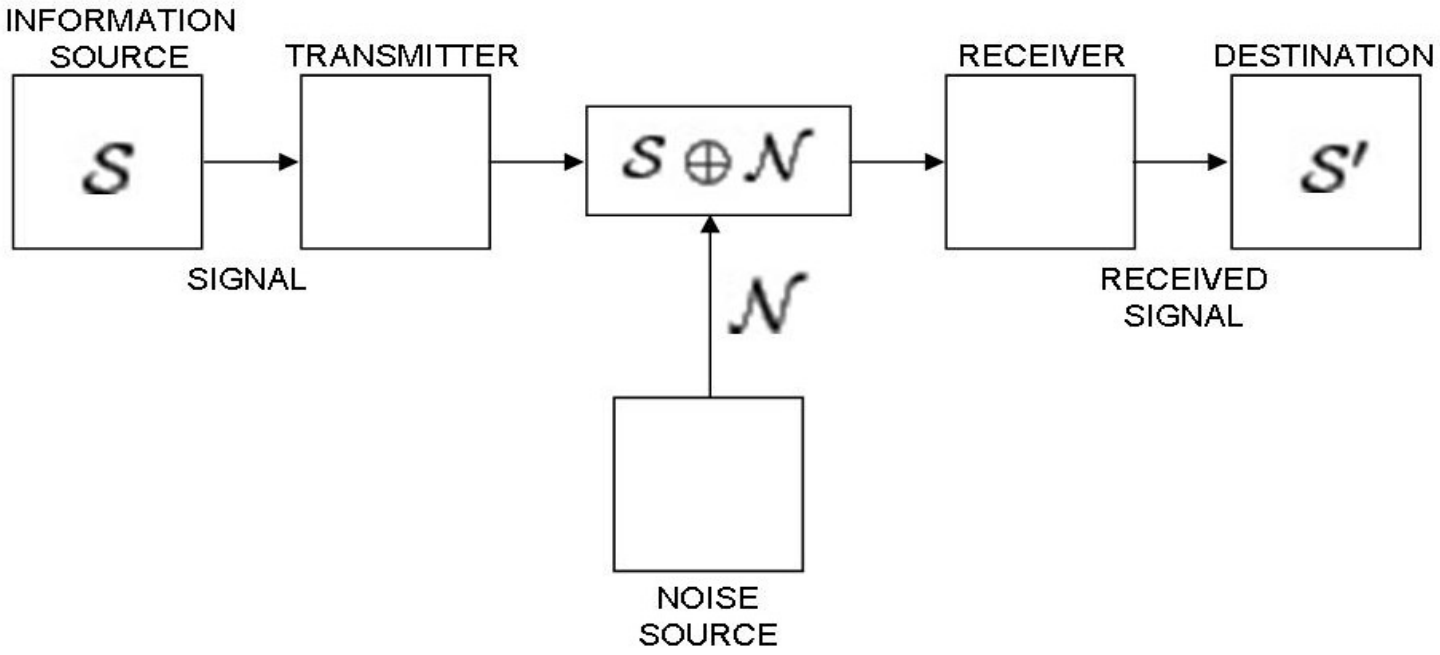
- **Example: book Genesis, chapter 1, verse 1**

ASV	In the beginning God created the heavens and the earth.
BBE	At the first God made the heaven and the earth.
DBY	In the beginning God created the heavens and the earth.
KJV	In the beginning God created the heaven and the earth.
Webster	In the beginning God created the heaven and the earth.
WEB	In the beginning God created the heavens and the earth.
YLT	In the beginning of God's preparing the heavens and the earth.

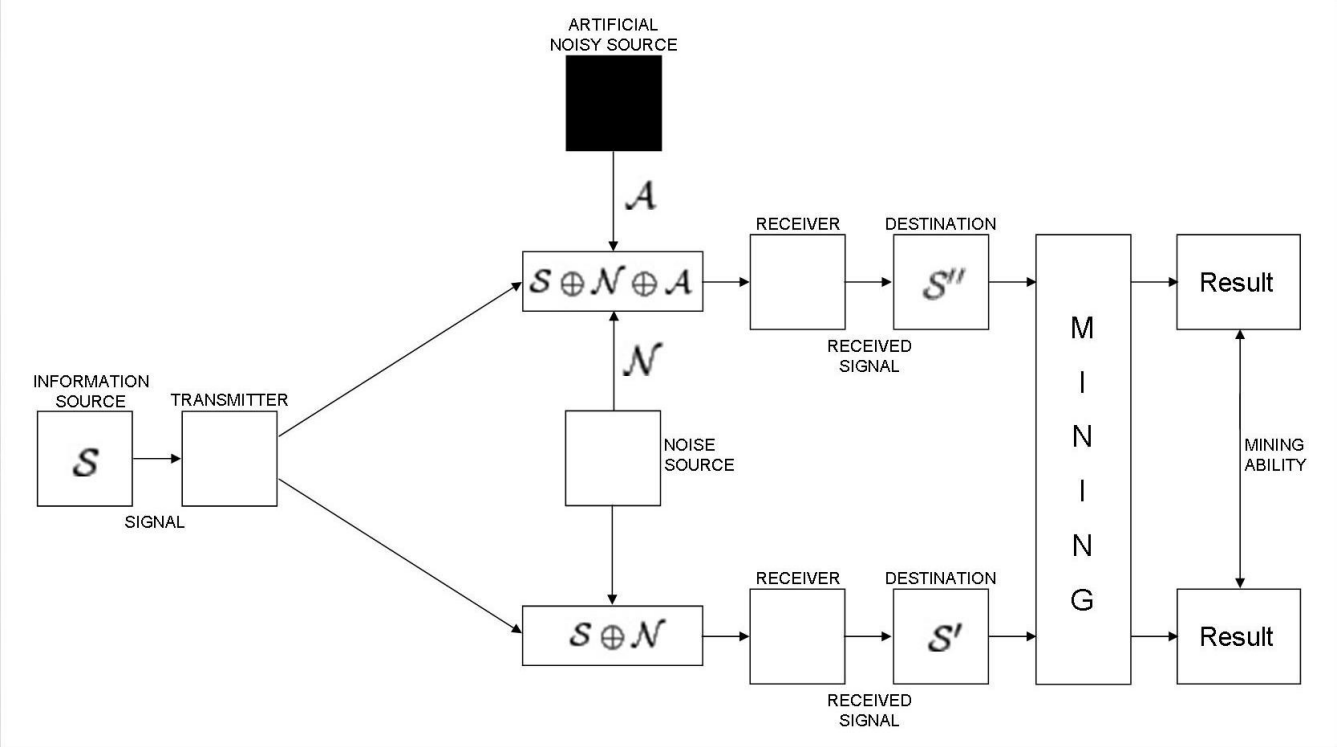
Reduced Bibels: all seven reduced Bible versions contain "only" the 28632 verses contained in all seven editions.

Framework

- **Basic idea**: Embed Historical Text Reuse in Shannon's Noisy Channel Theorem



Noisy Channel Evaluation



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

Text Reuse on English Bible versions Setup

- **Segmentation:** disjoint and versewise 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
- **Postprocessing:** not used

Text Reuse on 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 on 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
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

Recall vs. Text Reuse Compression

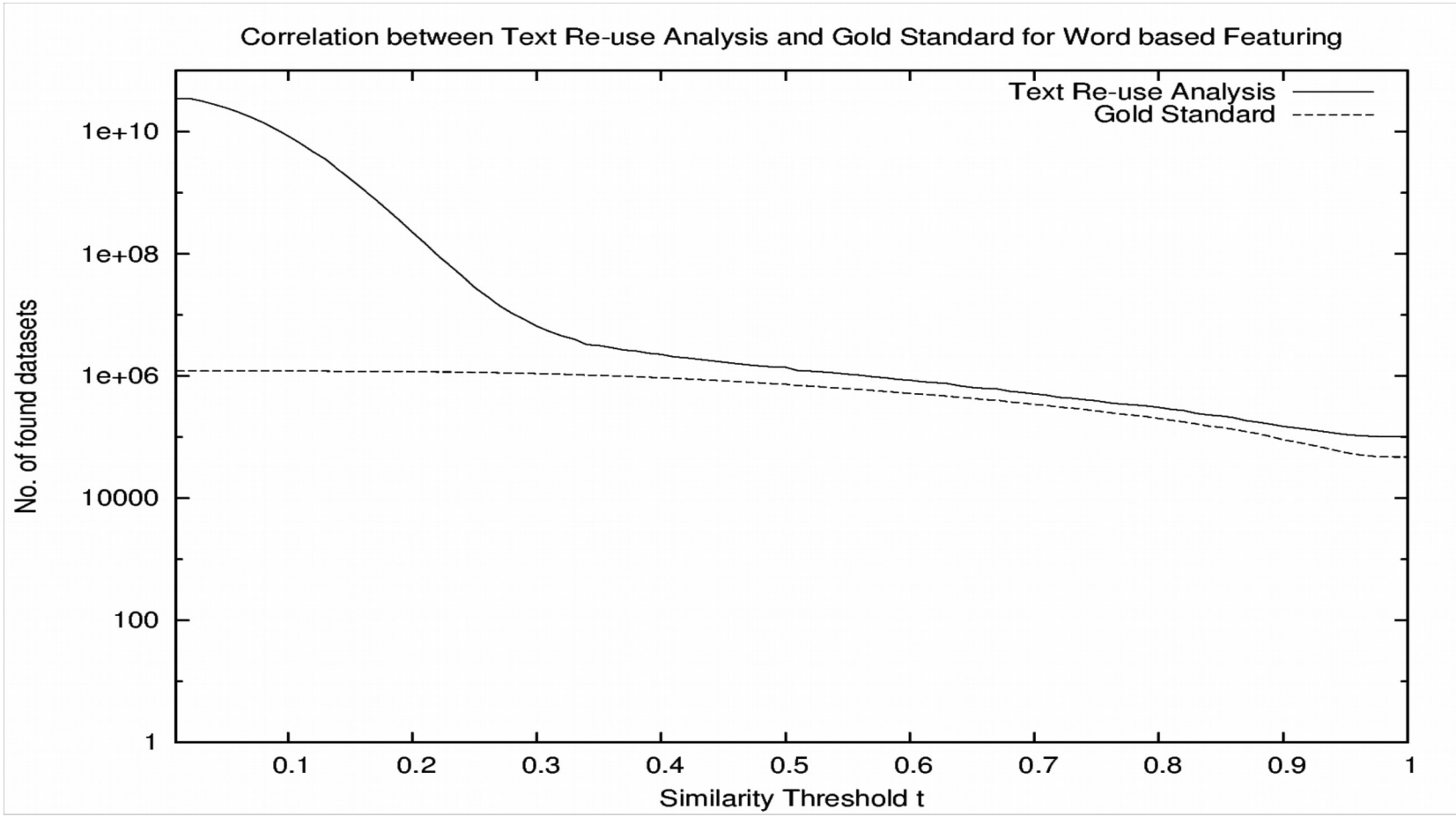
With

	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	0.17	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	0.88	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

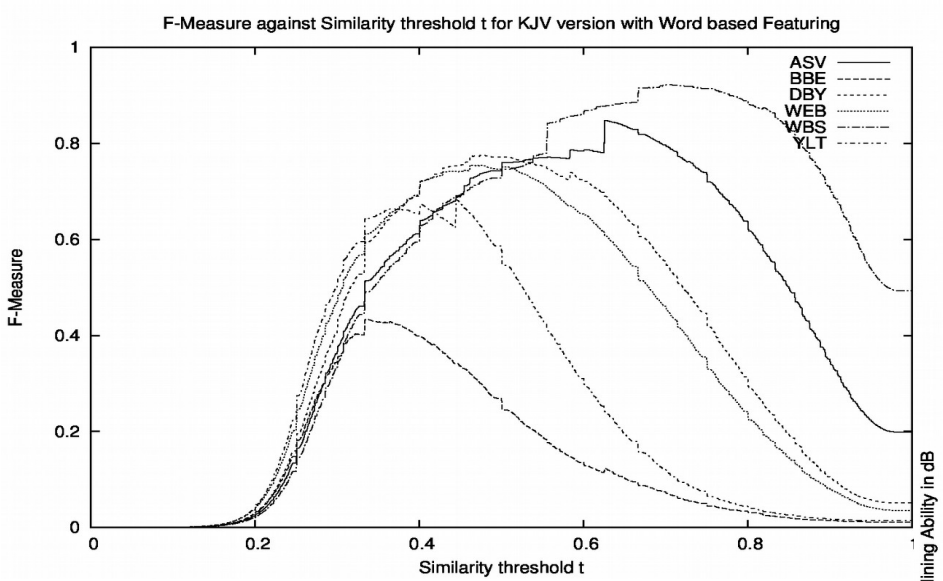
Without

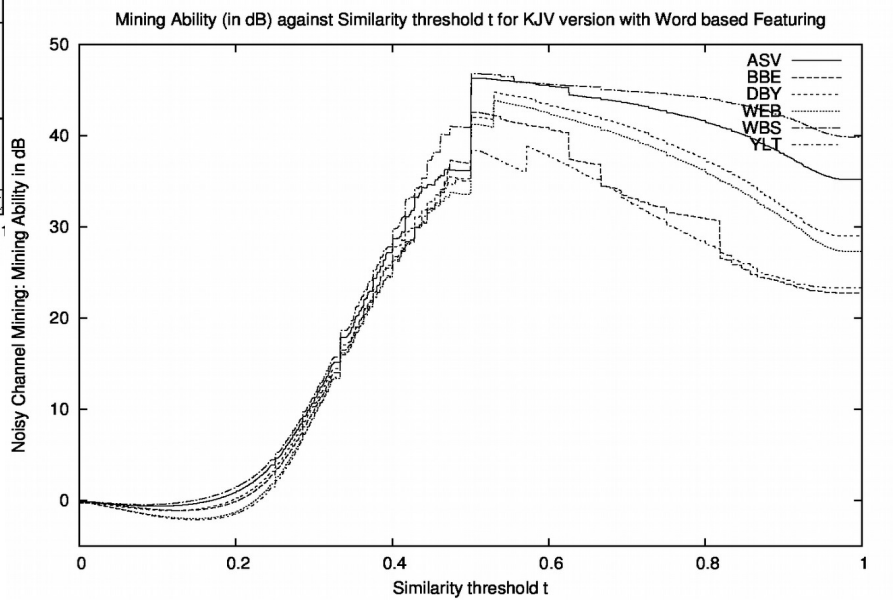
	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	6.16	6.15	6.16	6.18	6.02	6.01	6.01	5.99	5.42	5.39	5.37	5.33
ASV vs. DBY	5.22	5.19	5.20	5.19	4.98	4.96	4.97	4.95	4.60	4.58	4.58	4.57
ASV vs. 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	4.6	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	4.6	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	5.2	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	5.99	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

Dependency of recall and TR compression



F-Measure and Noisy Channel Evaluation





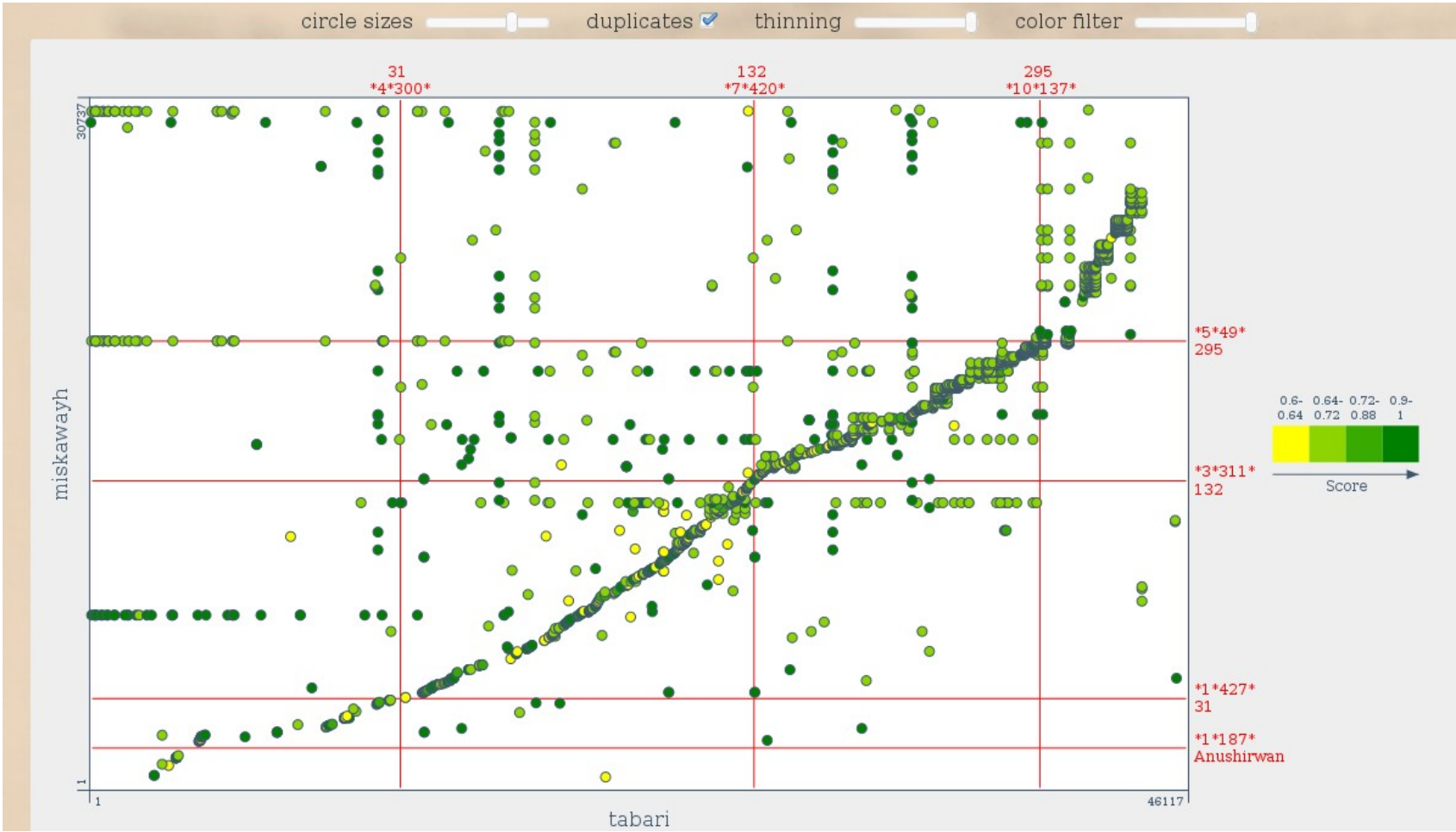
F-Measure: *WBS, ASV, DBY, WEB, YLT, BBE*
NCE: *WBS, ASV, DBY, WEB, BBE, YLT*

Relation of preprocessing, featuring, and reuse style

	ASV	BBE	DBY	WEB	WBS	YLT
ξ_1	(0.38, 0.77)	(0.08, 0.40)	(0.15, 0.74)	(0.15, 0.74)	(0.37, 0.91)	(0.08, 0.52)
ξ_2	(0.38, 0.79)	(0.08, 0.40)	(0.16, 0.74)	(0.15, 0.74)	(0.40, 0.92)	(0.08, 0.53)
ξ_3	(0.38, 0.78)	(0.08, 0.41)	(0.15, 0.74)	(0.15, 0.74)	(0.39, 0.91)	(0.08, 0.52)
ξ_4	(0.38, 0.78)	(0.08, 0.42)	(0.16, 0.74)	(0.15, 0.74)	(0.39, 0.91)	(0.09, 0.53)

	ASV	BBE	DBY	WEB	WBS	YLT
ξ_1	(0.63, 0.84)	(0.34, 0.44)	(0.46, 0.77)	(0.46, 0.75)	(0.70, 0.92)	(0.34, 0.64)
ξ_2	(0.63, 0.85)	(0.34, 0.45)	(0.48, 0.78)	(0.46, 0.76)	(0.70, 0.92)	(0.36, 0.65)
ξ_3	(0.63, 0.85)	(0.34, 0.43)	(0.48, 0.78)	(0.46, 0.76)	(0.70, 0.92)	(0.44, 0.68)
ξ_4	(0.63, 0.85)	(0.36, 0.44)	(0.48, 0.78)	(0.47, 0.76)	(0.70, 0.92)	(0.44, 0.70)

Dotplot view



Special issue on text reuse

Special Issue Call for Contribution

Computer-Aided Processing of Intertextuality
in Ancient Languages

Journal of Data Mining and Digital Humanities

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

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

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Thank you!

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

SPONSORED BY THE
Federal Ministry
of Education
and Research

Visit us at http://etrap.gcdh.de