eTRACES/eTRAP: Computational Aspects of *Historical Text Re-use*

Marco Büchler

Göttingen Centre for Digital Humanities
Georg-August University Göttingen, Germany

Feb. 14Th 2014

A fundamental question

How can the computer really support to identify lines of transmissions (text re-use) on big data?

A fundamental answer

Naive method:

- Compare every text chunk (like sentence) with each other.

- TLG: 5,500,000*5,500,000 = **3.025e13** comparisons

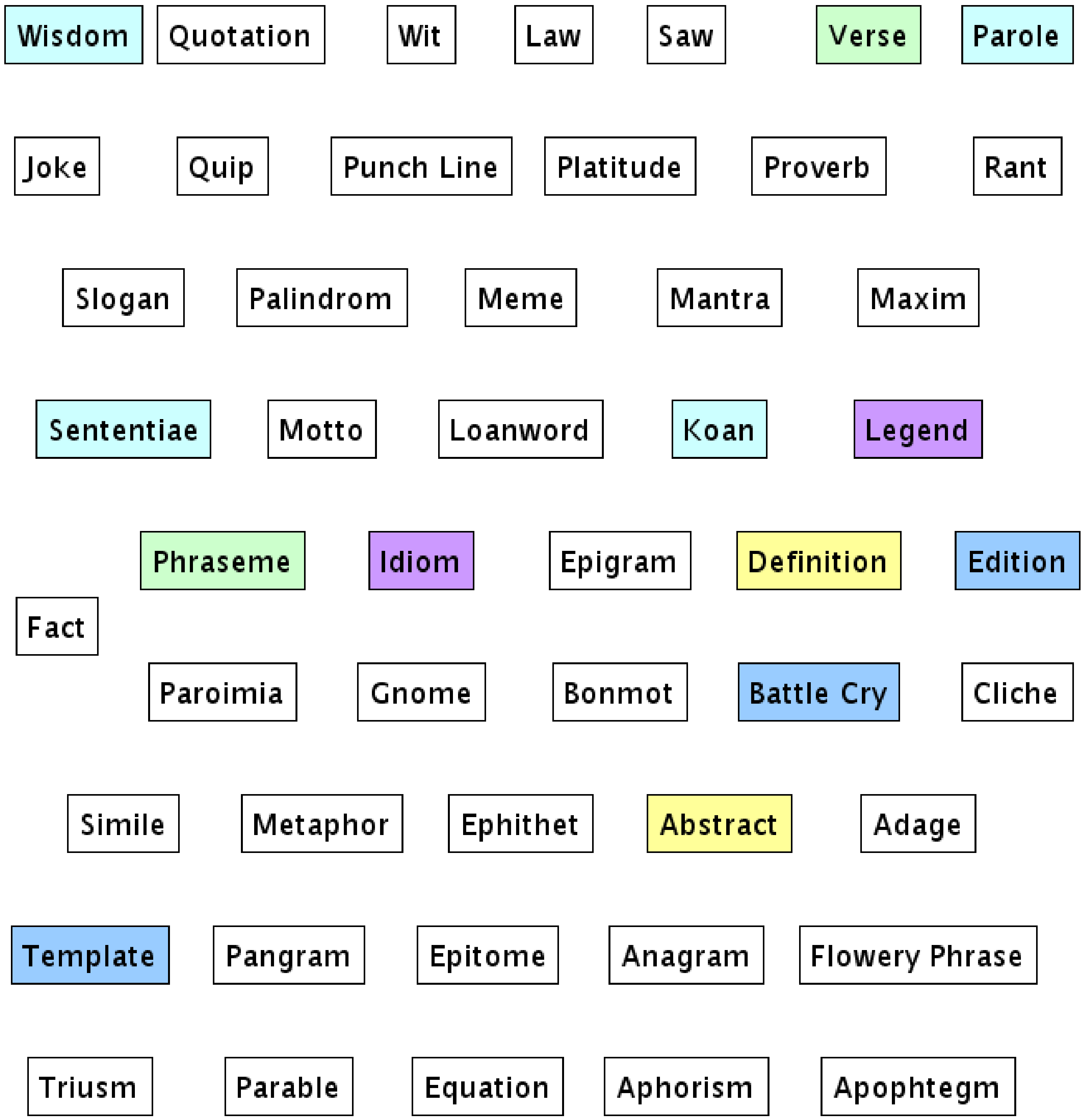
- Assumption: Comparison rate of **1000 sentences/sec**.

- This process would run about **3.025e10** seconds or more than **959 years.**

A fundamental question

How can the computer really support to identify lines of transmissions (text re-use)?

Complete view: **Re-use Types**



Wisdom Quotation Wit Law Saw Verse Parole

Joke Quip Punch Line Platitude Proverb Rant

Slogan Palindrom Meme Mantra Maxim

Sententiae Motto Loanword Koan Legend

Phraseme Idiom Epigram Definition Edition

Fact

Paroimia Gnome Bonmot Battle Cry Cliche

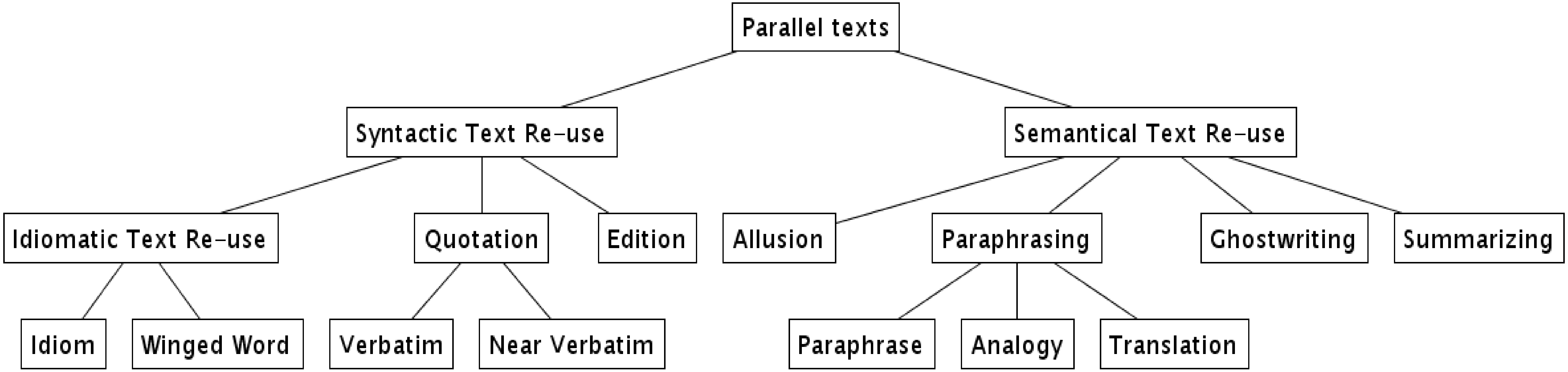
Simile Metaphor Ephithet Abstract Adage

Template Pangram Epitome Anagram Flowery Phrase

Triusm Parable Equation Aphorism Apophtegm

- **Stability (yellow)**: syntactic vs. semantic
- **Purpose (green)**
- **Size of Text Re-use (blue)**
- **Literary classification (light blue)**
- **Degree of distribution (lila)**
- **Written and oral transmission**

Complete view: **Re-use Styles**



Basic Question

Basic question: Distribution of *Re-use Types* und *Re-use Styles* are most often unknown. Question: Which model(s) should be chosen?

Approach

7-Level-Architecture to deal with *Data Diversity*:

a. **Segmentation**: Defining *Re-use Units*

b. **Preprocessing**: Cleaning of *Re-use Units*

c. **Featuring**: Creating of *Fingerprints*

d. **Selection**: Selection of *Features* from *Fingerprint*

e. **Linking**: Linking of *Re-use Units* given common Features

f. **Scoring**: Scoring of Re-use Overlaps of pairwise linked Re-use Units

g. **Postprocessing**: optional Post-Processing of the Re-use Graph

Implemented in TRACER software: more than a million permutations of implementations of different levels are recently possible

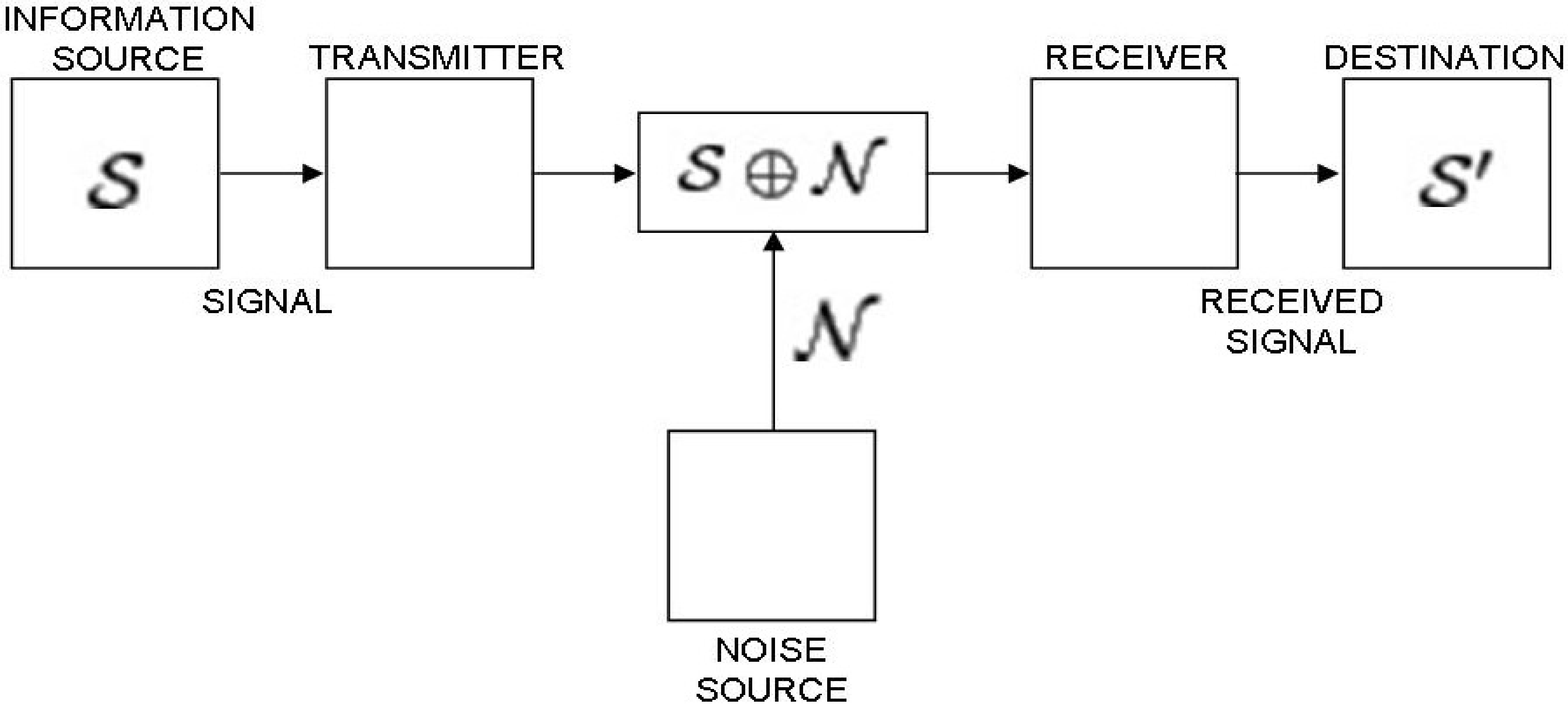
A (second) fundamental question

Using the computer for big data with the data diversity,

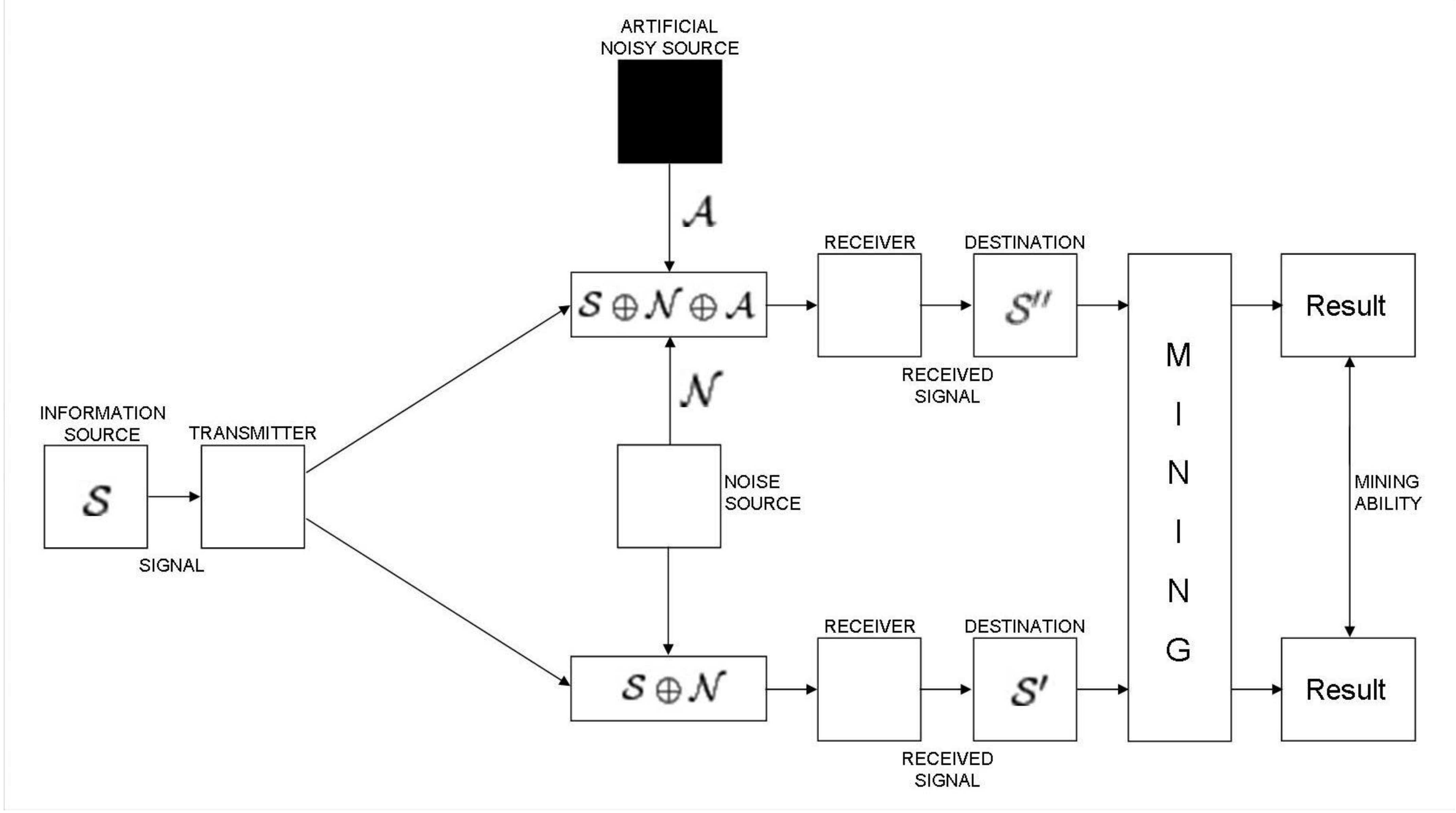
how do you evaluate with such a data diversity?

Methodology

Basic idea: Embed *Historical Text Re-use* in Shannon’s *Noisy Channel Theorem*



Evaluation methodology: *Noisy Channel Evaluation* I



Hint: There is always the entire result of a Digital Library compared to a *Randomised Digital Library*.

Evaluation methodology: *Noisy Channel Evaluation* II

Signal-Noise-Ratio from signal and satellite technique:

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{4.20}$$

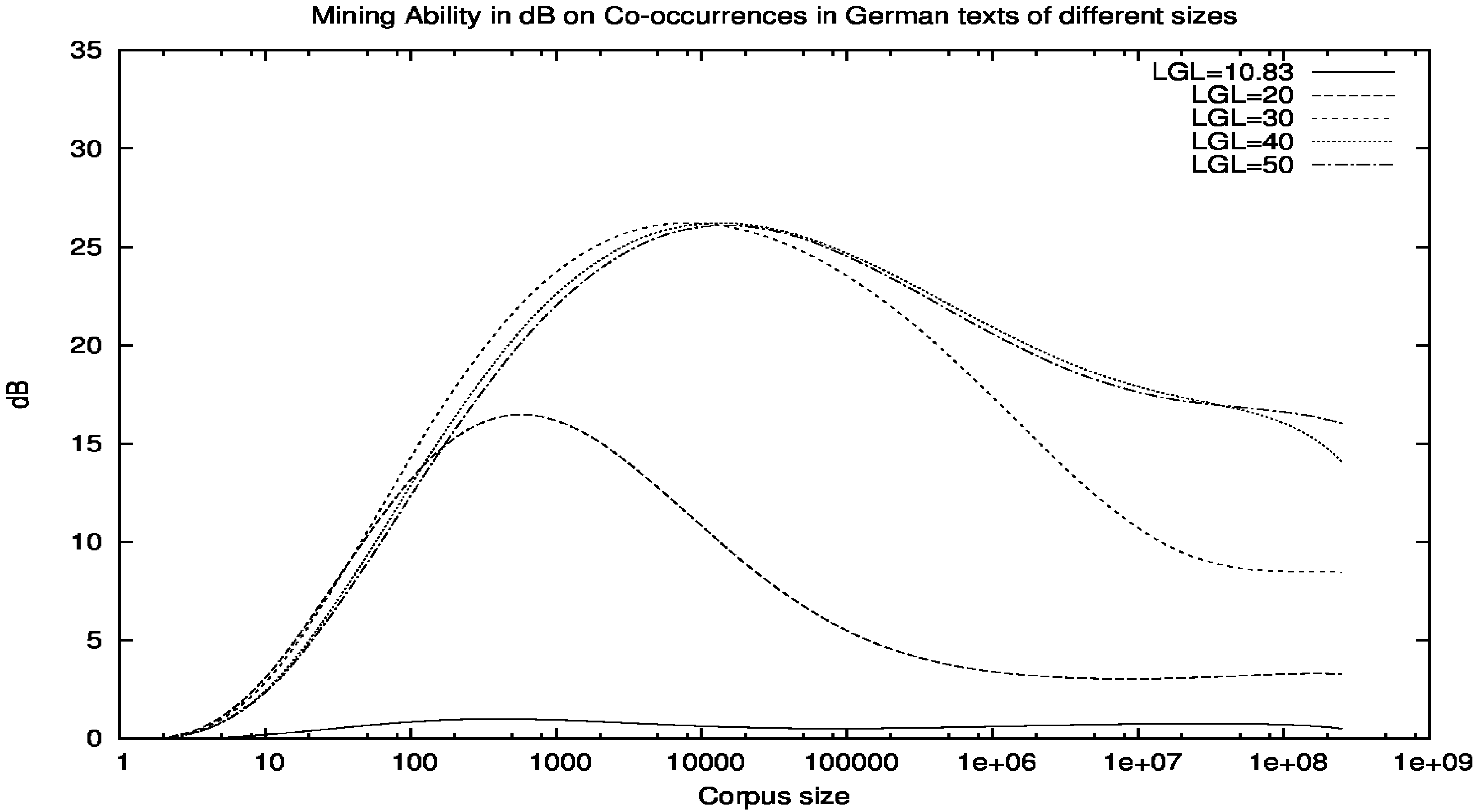
Scaled *Signal-Noise-Ratio,* in unit Dezibel (dB):

$$SNR_{db} = 10 \cdot \log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{4.21}$$

Mining Ability (in dB): With *Mining Ability* it is measured how good a technique (both manually and automatically) can distinguish a pattern from noise.

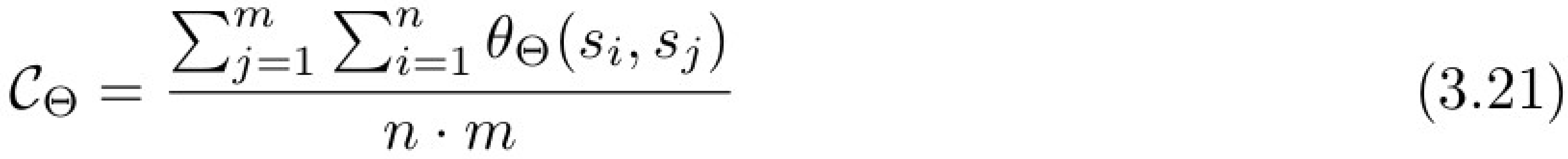
$$\mathcal{L}_{Quant}(\Theta) = 10 \cdot \log_{10} \frac{|E_{D_S,\phi_\Theta}|}{\max(1, |E_{D_S^m,\phi_\Theta}|)} dB \tag{4.22}$$

Evaluation methodology: *Noisy Channel Evaluation* III



Be careful with Big Data Analysis in general.

Untersuchungsmethodik:
Text Re-use Compression



$$\mathcal{C}_\Theta = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} \theta_\Theta(s_i, s_j)}{n \cdot m} \tag{3.21}$$

Text Re-use on English Bibles
Different intentions

a. **American Standard Version** (ASV): 20. Jh., Fokus auf USA

b. **Bible in Basic English** (BBE): Verse in einfacher Sprache (gleicher Inhalt)

c. **Darby Version** (DBY): im 19. Jh. aus hebräischen und griechischen Texten erstellt, mehrere Editoren durch den Tod von Darby

d. **King James Version** (KJV): englischen Bibelversionen aus dem 16. Jh.

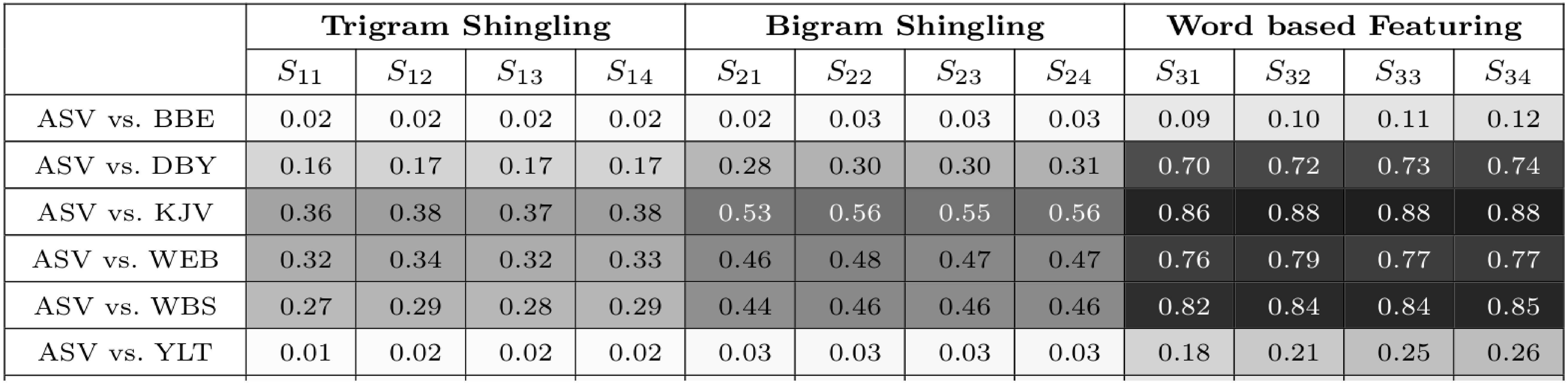
e. **Webster's Revision** (WBS): Revision von KJV im 19. Jh.

f. **World English Bible** (WEB): 21. Jh., Fokus auf weltweite Gültigkeit

g. **Young Literal Translation** (YLT): Verse an hebräischer Syntax orientiert (Syntax)

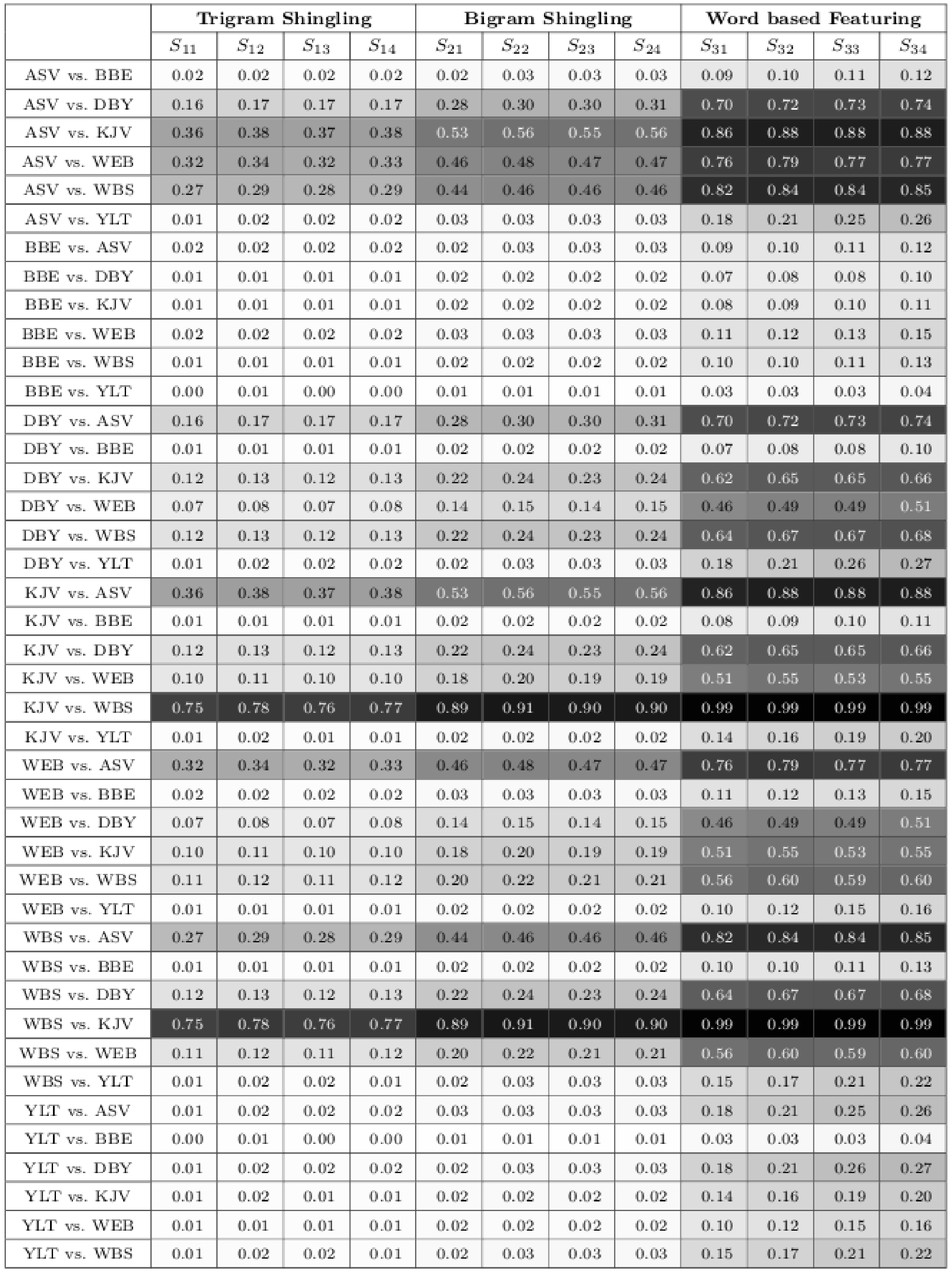
Text Re-use on English Bibles
Results – Recall I

Re-use Style: Bis zu welchem Grad der Veränderung kann ein *Text Re-use* noch automatisch erkannt werden?



	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 Re-use on English Bibles
Results – Recall II



	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

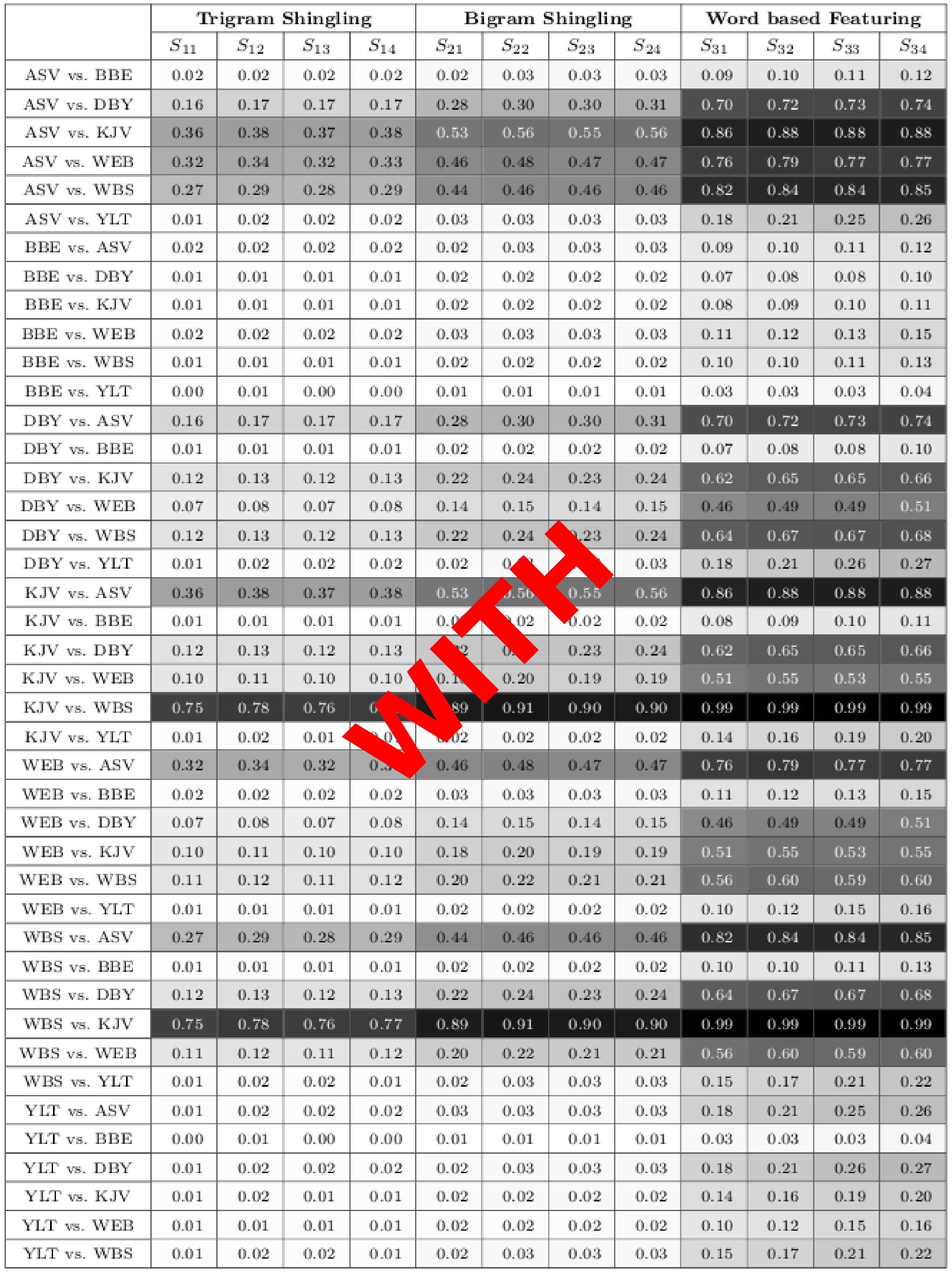
Re-use Style: Bis zu welchem Grad der Veränderung kann ein *Text Re-use* noch automatisch erkannt werden?

- *Word based featuring* eignet sich besser solange es sich nicht um Duplikate handelt

- Analyse von *KJV vs. WBS* zeigt jedoch auch, dass Shingling-Techniken anwendbar sind

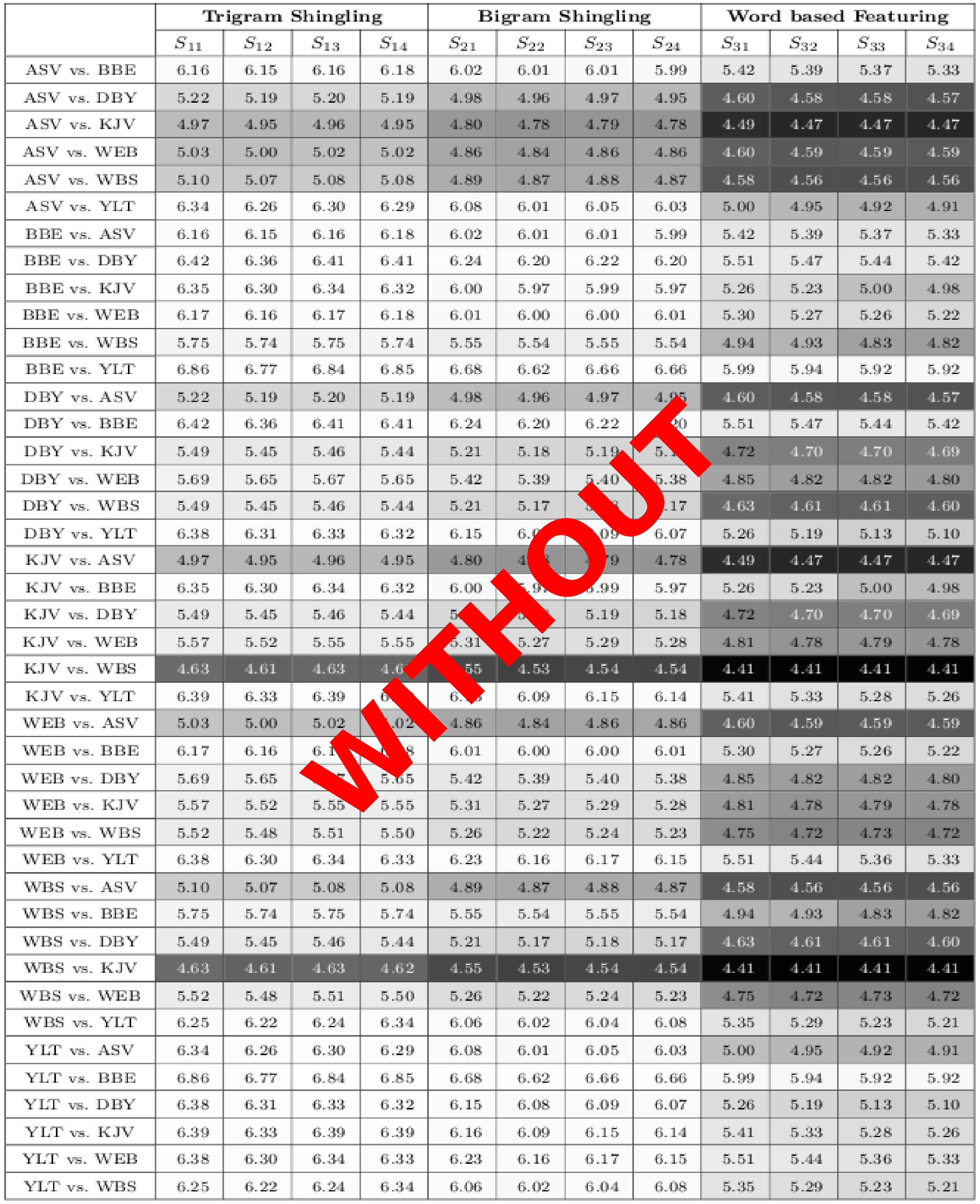
- *BBE* (semantische Veränderung) und *YLT* (syntaktische Veränderung) zeigen, dass bei größeren Abweichungen bereits eine deutlich schlechtere Performance gemessen werden kann

Text Re-use on English Bibles
Recall vs. Text Re-use Compression



WITH

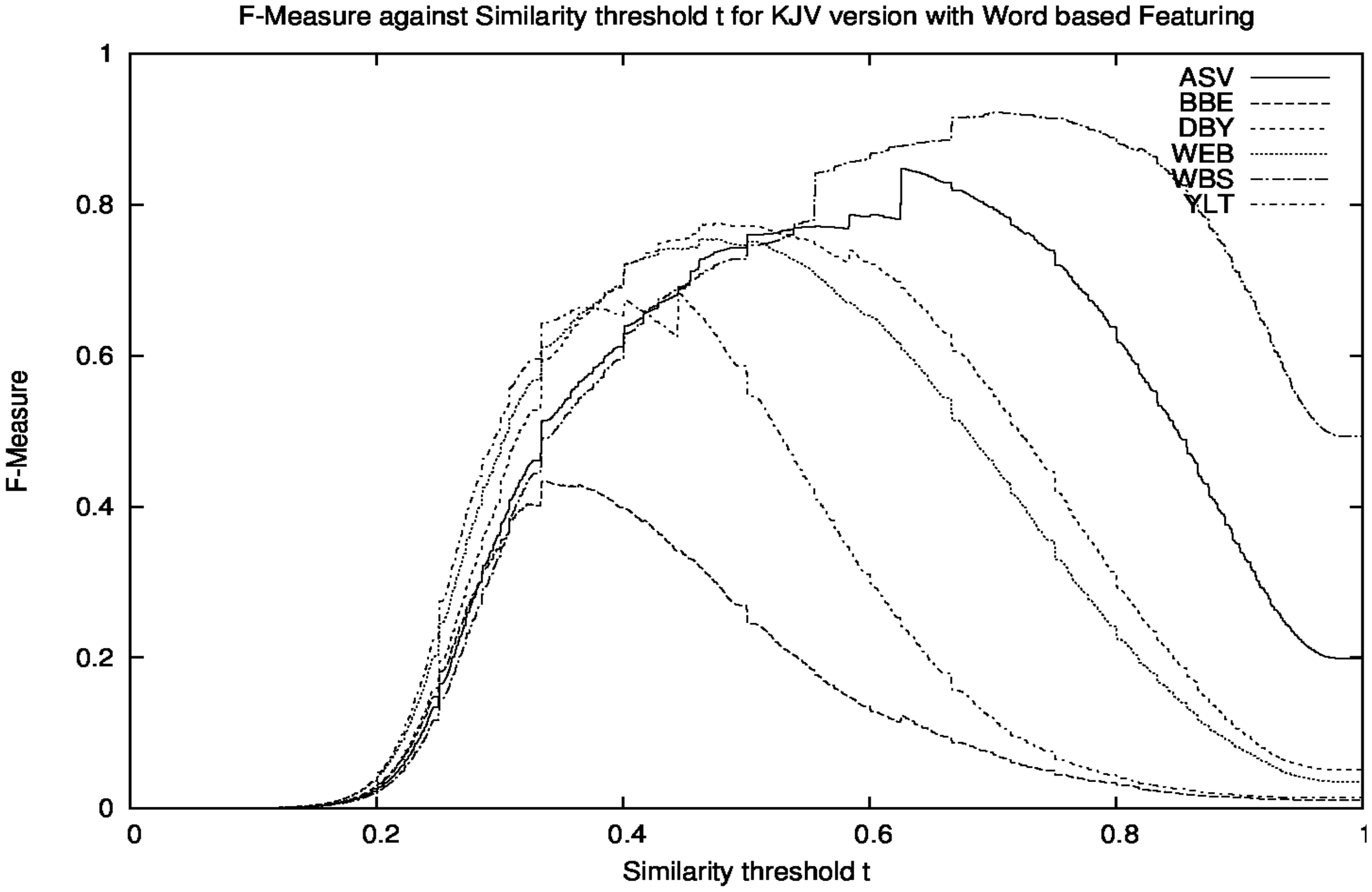
	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.09	0.10	0.11	0.12
ASV vs. DBY	0.16	0.17	0.17	0.17	0.28	0.30	0.30	0.31	0.70	0.72	0.73	0.74
ASV vs. 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	[illegible]	[illegible]	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	[illegible]	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	[illegible]	[illegible]	0.23	0.24	0.62	0.65	0.65	0.66
KJV vs. WEB	0.10	0.11	0.10	0.10	[illegible]	0.20	0.19	0.19	0.51	0.55	0.53	0.55
KJV vs. WBS	0.75	0.78	0.76	[illegible]	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	[illegible]	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	[illegible]	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

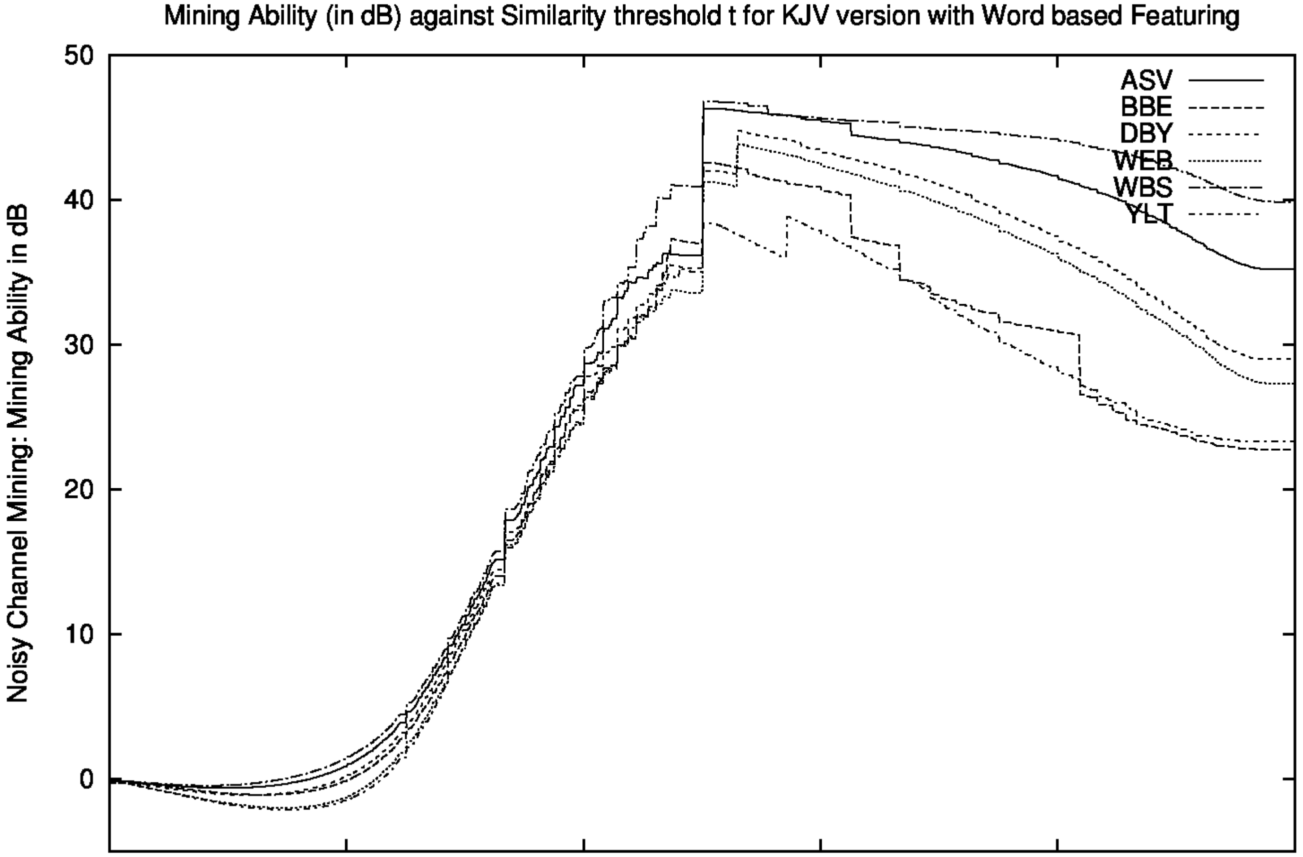


WITHOUT

	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	[illegible]	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	[illegible]	5.17	4.63	4.61	4.61	4.60
DBY vs. YLT	6.38	6.31	6.33	6.32	6.15	[illegible]	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	[illegible]	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	[illegible]	5.99	5.97	5.26	5.23	5.00	4.98
KJV vs. DBY	5.49	5.45	5.46	5.44	[illegible]	[illegible]	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	[illegible]	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	[illegible]	[illegible]	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	[illegible]	[illegible]	6.01	6.00	6.00	6.01	5.30	5.27	5.26	5.22
WEB vs. DBY	5.69	5.65	[illegible]	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 Re-use on English Bibles
F-Measure vs. Noisy Channel Eval.



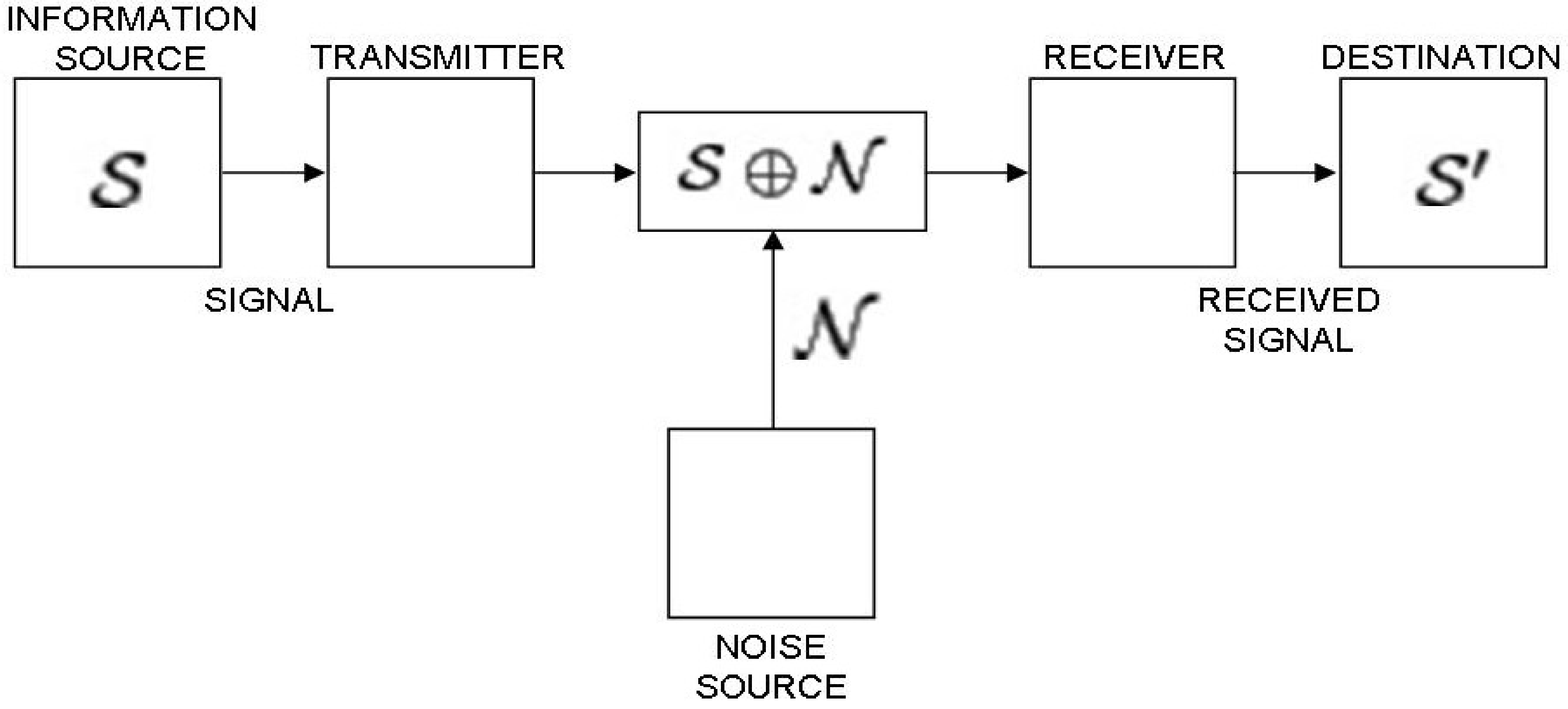


F-Measure: $WBS, ASV, DBY, WEB, YLT, BBE$

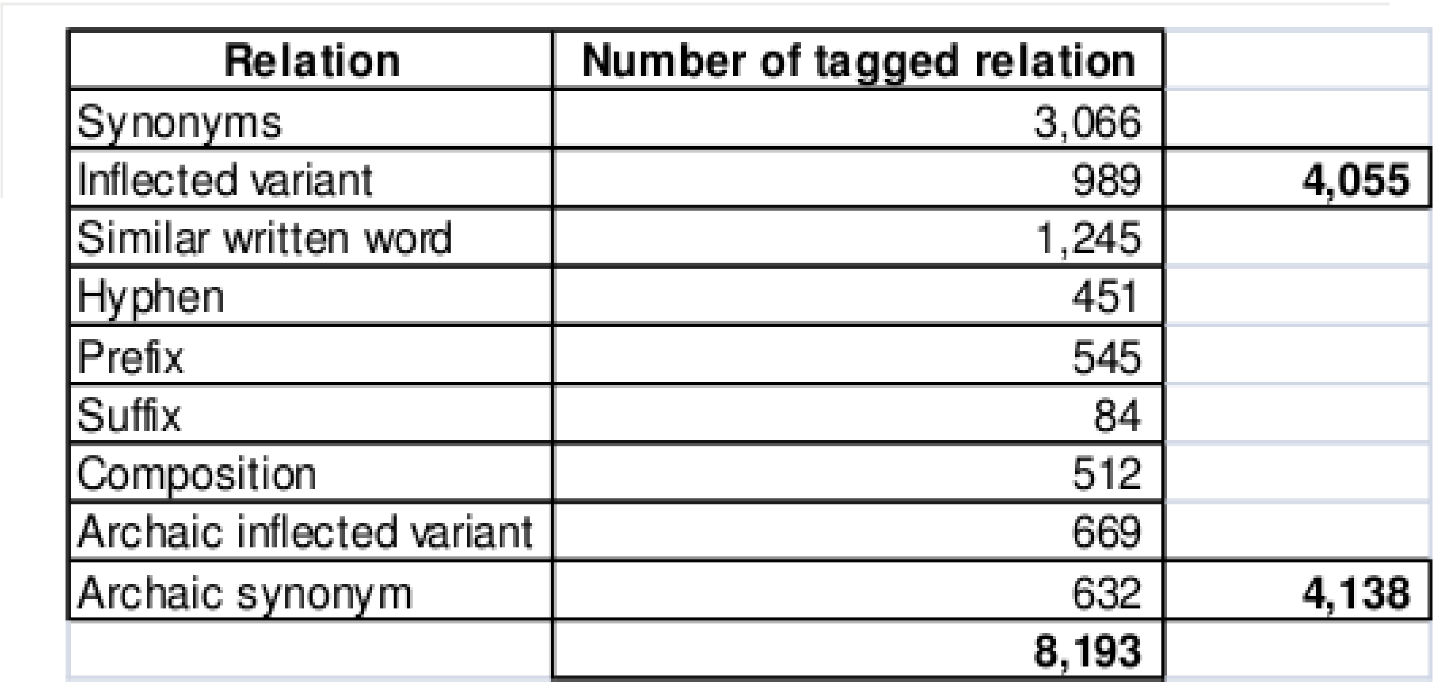
NCE: $WBS, ASV, DBY, WEB, BBE, YLT$

Reminder: Methodology

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



Noisy Channel Mining I



Relation	Number of tagged relation	
Synonyms	3,066	
Inflected variant	989	**4,055**
Similar written word	1,245	
Hyphen	451	
Prefix	545	
Suffix	84	
Composition	512	
Archaic inflected variant	669	
Archaic synonym	632	**4,138**
	8,193	

Table 5: Identified systematic classes from extracting paradigmatic relations.

- **Hyphen**:

  birth-day vs. birthday,

  back-bone vs. backbone

  zareth-shahar vs. zarethshahar

- **Prefix**:

  ambush vs. ambushment

  shimite vs. shimites

- **Suffix:**

  bearing vs. childbearing

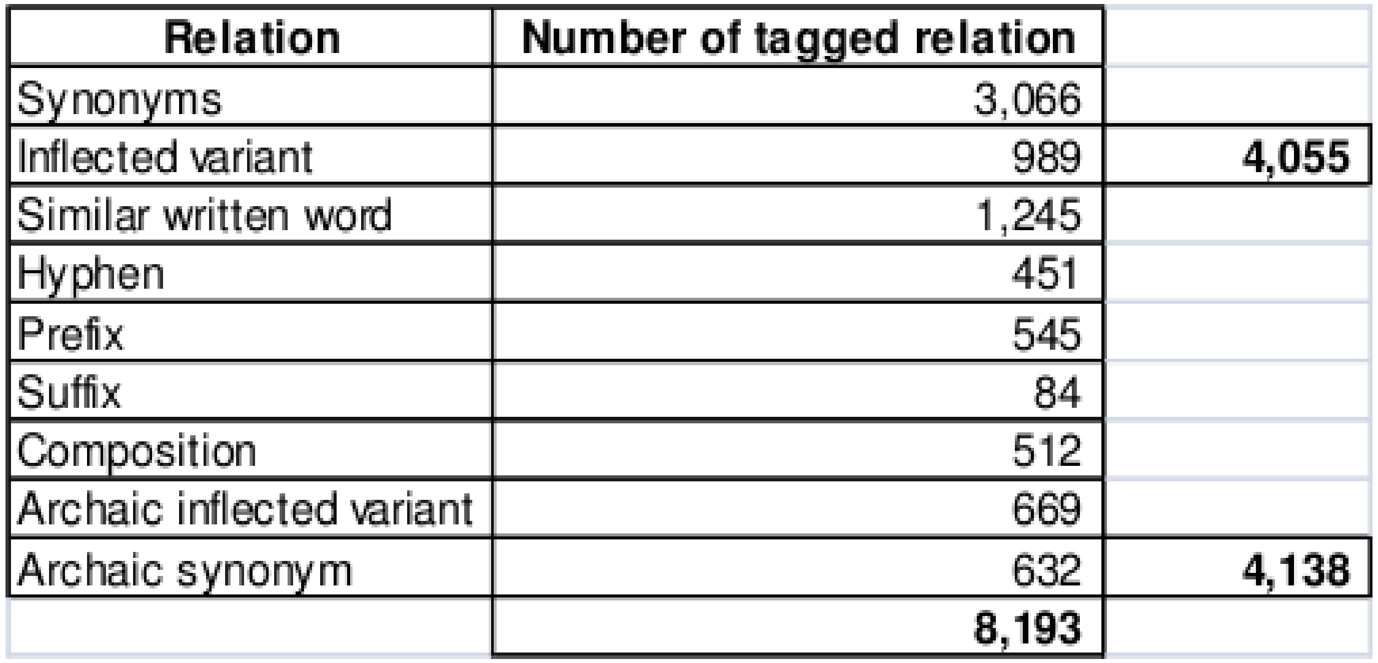
- **Composition**:

  sea-beast vs. sea-monster (synonym)

  sea-gull vs. sea-mew vs. sea-hawk (cohyponym)

  apple-tree vs. citron-tree (cohyponym)

Noisy Channel Mining II



Relation	Number of tagged relation	
Synonyms	3,066	
Inflected variant	989	**4,055**
Similar written word	1,245	
Hyphen	451	
Prefix	545	
Suffix	84	
Composition	512	
Archaic inflected variant	669	
Archaic synonym	632	**4,138**
	8,193	

Table 5: Identified systematic classes from extracting paradigmatic relations.

- **Similar written words**:

    anathothite vs. anethothite vs. anetothite vs. annethothite vs. antothite

- **Further around *4,000* word pairs are extracted but not classified that contains** noise
But also: punishment vs. torment

- **Ignored: any kind of negation (e. g. book Genesis, chapter 34, verse 19):**

    not defer (ASV, KJV, Webster), *without loss of time* (Basic), *not delay* (Darby, YLT), and *not wait* (WEB)

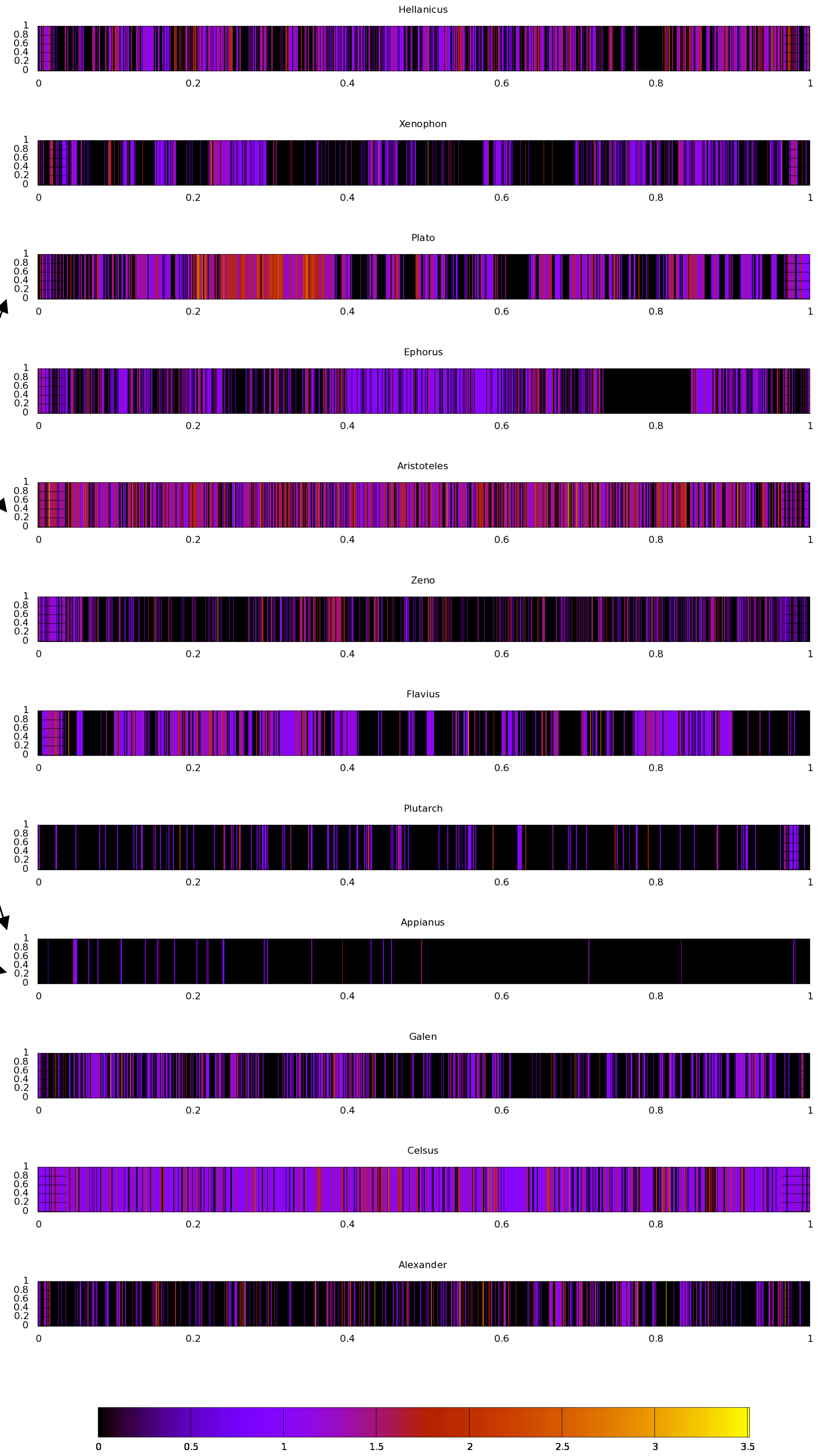
Big Data

Why do Big Data make sense?

Level 6: Post processing

A text re-use from a document with a high text re-use coverage is more trustworthy than from a less frequently re-used text.

A text re-use from a section of a document with a high text re-use temperature is more trustworthy than from a less frequently re-used part of a document.



Visualisation

How do you visualize Big Data from Text Re-use?

Microview I

Plato: Timaeus 91b7 ff.

αἱ δ' ἐν ταῖς γυναιξὶν αὖ μῆτραί τε καὶ ὑστέραι λεγόμεναι διὰ τὰ αὐτὰ ταῦτα ζῷον ἐπιθυμητικὸν ἐνὸν τῆς παιδοποιίας ὅταν ἄκαρπον παρὰ τὴν ὥραν χρόνον πολὺν γίγνηται χαλεπῶς ἀγανακτοῦν φέρει καὶ πλανώμενον πάντῃ κατὰ τὸ σῶμα τὰς τοῦ πνεύματος διεξόδους ἀποφράττον ἀναπνεῖν οὐκ ἐῶν εἰς ἀπορίας τὰς ἐσχάτας ἐμβάλλει καὶ νόσους παντοδαπὰς ἄλλας παρέχει μέχριπερ ἂν ἑκατέρων ἡ ἐπιθυμία καὶ ὁ ἔρως συναγαγόντες οἷον ἀπὸ δένδρων καρπὸν καταδρέψαντες ὡς εἰς ἄρουραν τὴν μήτραν ἀόρατα ὑπὸ σμικρότητος καὶ ἀδιάπλαστα ζῷα κατασπείραντες καὶ πάλιν διακρίναντες μεγάλα ἐντὸς ἐκθρέψωνται καὶ μετὰ τοῦτο εἰς φῶς ἀγαγόντες ζῴων ἀποτελέσωσι γένεσιν

αἱ δ' ἐν ταῖς γυναιξὶν αὖ μῆτραί τε καὶ ὑστέραι λεγόμεναι διὰ τὰ αὐτὰ ταῦτα ζῷον ἐπιθυμητικὸν ἐνὸν τῆς παιδοποιίας ὅταν ἄκαρπον περὶ τὴν ὥραν χρόνον πολὺν γίγνηται χαλεπῶς ἀγανακτοῦν φέρει καὶ πλανώμενον πάντῃ κατὰ τὸ σῶμα τὰς τοῦ πνεύματος διεξόδους ἀποφράττον ἀναπνεῖν οὐκ ἐῶν εἰς ἀπορίας τὰς ἐσχάτας ἐμβάλλει καὶ νόσους παντοδαπὰς ἄλλας παρέχει μέχριπερ ἂν ἑκατέρων ἡ ἐπιθυμία καὶ ὁ ἔρως ξυναγαγόντες οἷον ἀπὸ δένδρων καρπὸν καταδρέψαντες ὡς εἰς ἄρουραν τὴν μήτραν ἀόρατα ὑπὸ σμικρότητος καὶ ἀδιάπλαστα ζῷα κατασπείραντες καὶ πάλιν διακρίναντες μεγάλα ἐντὸς ἐκθρέψωνται καὶ μετὰ τοῦτο εἰς φῶς ἀγαγόντες ζῴων ἀποτελέσωσι γένεσιν

περὶ δὲ τῆς μήτρας ὅτι τε ζῷόν ἐστι καὶ αὕτη καὶ τὰ ἀπὸ τοῦ πατρὸς ἐξερχόμενα μόρια ταῦτα πάλιν λέγει Πλάτων αἱ δ' ἐν ταῖς γυναιξὶν αὖ μῆτραί τε καὶ ὑστέραι λεγόμεναι διὰ τὰ αὐτὰ ταῦτα ζῷον ἐπιθυμητικὸν ἐνὸν τῆς παιδοποιίας ὅταν ἄκαρπον παρὰ τὴν ὥραν χρόνον πολὺν γίνηται χαλεπῶς ἀγανακτοῦν φέρει καὶ πλανώμενον πάντῃ κατὰ τὸ σῶμα τὰς τοῦ πνεύματος διεξόδους ἀποφράττον καὶ ἀναπνεῖν οὐκ ἐῶν εἰς ἀπορίας τὰς ἐσχάτας ἐμβάλλει καὶ νόσους παντοδαπὰς ἄλλας παρέχει μέχριπερ ἂν ἑκατέρων ἡ ἐπιθυμία καὶ ὁ ἔρως ξυναγαγόντες οἷον ἀπὸ δένδρων καρπὸν καταδρέψαντες ὡς εἰς ἄρουραν τὴν μήτραν ἀόρατα ὑπὸ σμικρότητος καὶ ἀδιάπλαστα ζῷα κατασπείραντες καὶ πάλιν διακρίναντες μεγάλα ἐντὸς ἐκθρέψωνται καὶ μετὰ τοῦτο εἰς φῶς ἀγαγόντες ζῴων ἀποτελέσωσι γένεσιν

Microview II

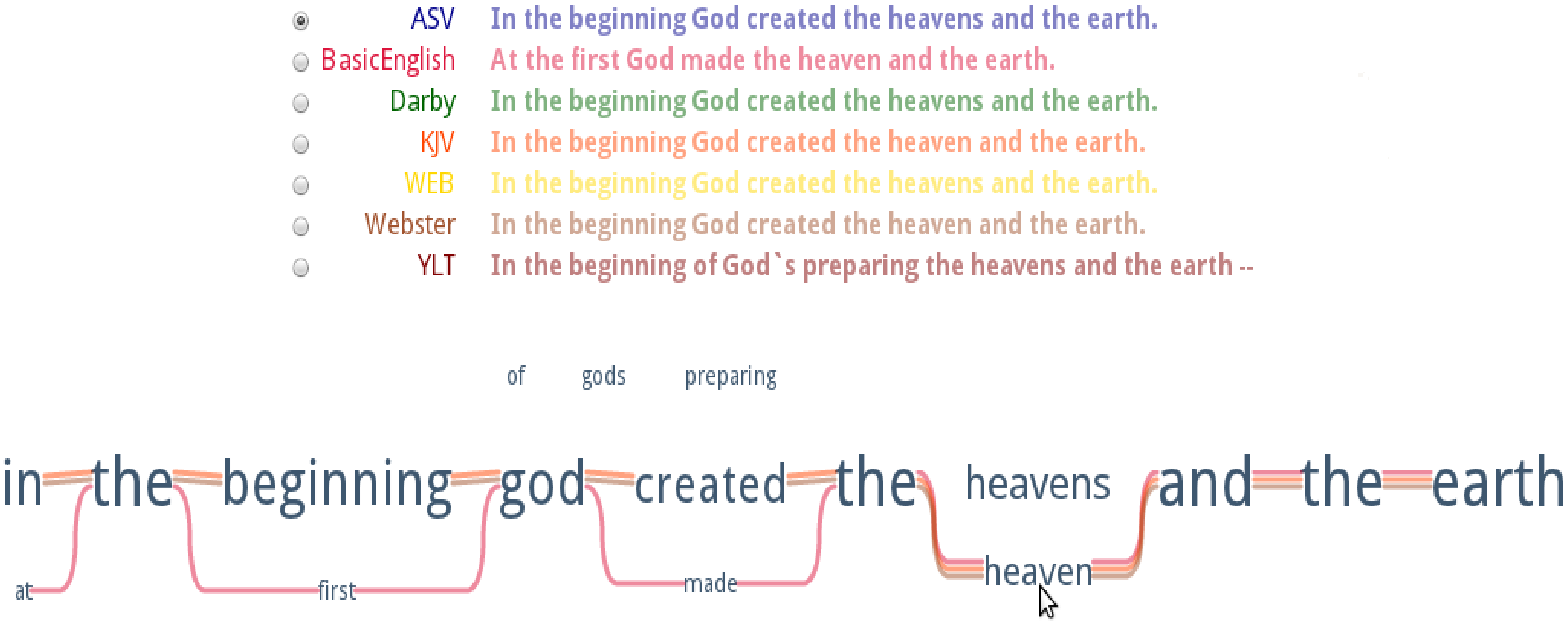
Plato: Timaeus 91b7 ff.

αἱ δ' ἐν ταῖς γυναιξὶν αὖ μῆτραί τε καὶ ὑστέραι λεγόμεναι διὰ τὰ αὐτὰ ταῦτα ζῷον ἐπιθυμητικὸν ἐνὸν τῆς παιδοποιίας ὅταν ἄκαρπον παρὰ τὴν ὥραν χρόνον πολὺν γίγνηται χαλεπῶς ἀγανακτοῦν φέρει καὶ πλανώμενον πάντῃ κατὰ τὸ σῶμα τὰς τοῦ πνεύματος διεξόδους ἀποφράττον ἀναπνεῖν οὐκ ἐῶν εἰς ἀπορίας τὰς ἐσχάτας ἐμβάλλει καὶ νόσους παντοδαπὰς ἄλλας παρέχει μέχριπερ ἂν ἑκατέρων ἡ ἐπιθυμία καὶ ὁ ἔρως συναγαγόντες οἷον ἀπὸ δένδρων καρπὸν καταδρέψαντες ὡς εἰς ἄρουραν τὴν μήτραν ἀόρατα ὑπὸ σμικρότητος καὶ ἀδιάπλαστα ζῷα κατασπείραντες καὶ πάλιν διακρίναντες μεγάλα ἐντὸς ἐκθρέψωνται καὶ μετὰ τοῦτο εἰς φῶς ἀγαγόντες ζῴων ἀποτελέσωσι γένεσιν

αἱ δ' ἐν ταῖς γυναιξὶν αὖ μῆτραί τε καὶ ὑστέραι λεγόμεναι διὰ τὰ αὐτὰ ταῦτα ζῷον ἐπιθυμητικὸν ἐνὸν τῆς παιδοποιίας ὅταν ἄκαρπον περὶ τὴν ὥραν χρόνον πολὺν γίγνηται χαλεπῶς ἀγανακτοῦν φέρει καὶ πλανώμενον πάντῃ κατὰ τὸ σῶμα τὰς τοῦ πνεύματος διεξόδους ἀποφράττον ἀναπνεῖν οὐκ ἐῶν εἰς ἀπορίας τὰς ἐσχάτας ἐμβάλλει καὶ νόσους παντοδαπὰς ἄλλας παρέχει μέχριπερ ἂν ἑκατέρων ἡ ἐπιθυμία καὶ ὁ ἔρως ξυναγαγόντες οἷον ἀπὸ δένδρων καρπὸν καταδρέψαντες ὡς εἰς ἄρουραν τὴν μήτραν ἀόρατα ὑπὸ σμικρότητος καὶ ἀδιάπλαστα ζῷα κατασπείραντες καὶ πάλιν διακρίναντες μεγάλα ἐντὸς ἐκθρέψωνται καὶ μετὰ τοῦτο εἰς φῶς ἀγαγόντες ζῴων ἀποτελέσωσι γένεσιν

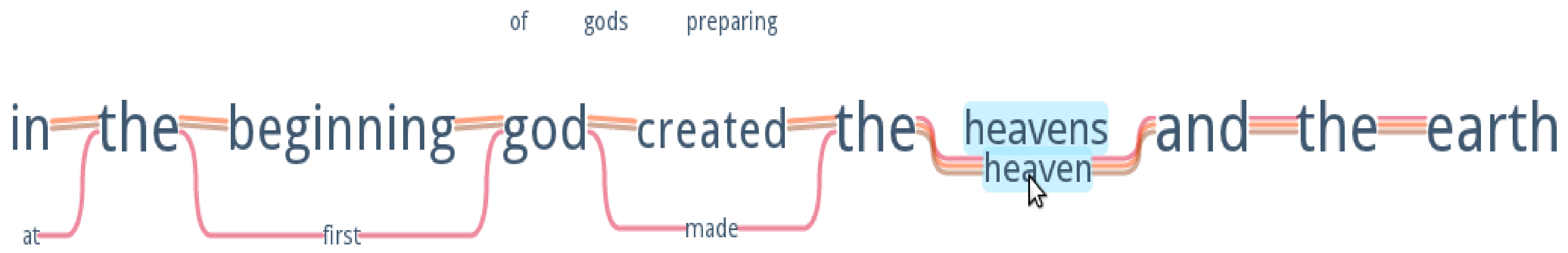
περὶ δὲ τῆς μήτρας ὅτι τε ζῷόν ἐστι καὶ αὕτη καὶ τὰ ἀπὸ τοῦ πατρὸς ἐξερχόμενα μόρια ταῦτα πάλιν λέγει Πλάτων αἱ δ' ἐν ταῖς γυναιξὶν αὖ μῆτραί τε καὶ ὑστέραι λεγόμεναι διὰ τὰ αὐτὰ ταῦτα ζῷον ἐπιθυμητικὸν ἐνὸν τῆς παιδοποιίας ὅταν ἄκαρπον παρὰ τὴν ὥραν χρόνον πολὺν γίνηται χαλεπῶς ἀγανακτοῦν φέρει καὶ πλανώμενον πάντῃ κατὰ τὸ σῶμα τὰς τοῦ πνεύματος διεξόδους ἀποφράττον καὶ ἀναπνεῖν οὐκ ἐῶν εἰς ἀπορίας τὰς ἐσχάτας ἐμβάλλει καὶ νόσους παντοδαπὰς ἄλλας παρέχει μέχριπερ ἂν ἑκατέρων ἡ ἐπιθυμία καὶ ὁ ἔρως ξυναγαγόντες οἷον ἀπὸ δένδρων καρπὸν καταδρέψαντες ὡς εἰς ἄρουραν τὴν μήτραν ἀόρατα ὑπὸ σμικρότητος καὶ ἀδιάπλαστα ζῷα κατασπείραντες καὶ πάλιν διακρίναντες μεγάλα ἐντὸς ἐκθρέψωνται καὶ μετὰ τοῦτο εἰς φῶς ἀγαγόντες ζῴων ἀποτελέσωσι γένεσιν

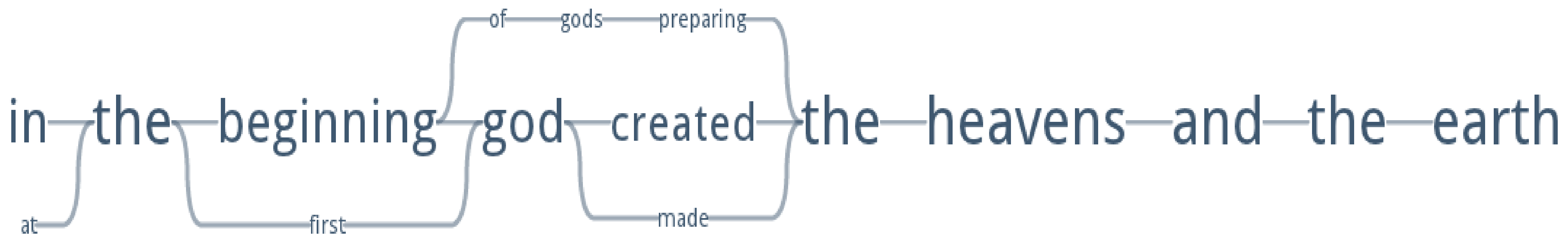
Microview III



●	ASV	In the beginning God created the heavens and the earth.
○	BasicEnglish	At the first God made the heaven and the earth.
○	Darby	In the beginning God created the heavens and the earth.
○	KJV	In the beginning God created the heaven and the earth.
○	WEB	In the beginning God created the heavens and the earth.
○	Webster	In the beginning God created the heaven and the earth.
○	YLT	In the beginning of God`s preparing the heavens and the earth --

Microview IV





Contacts

For more details:

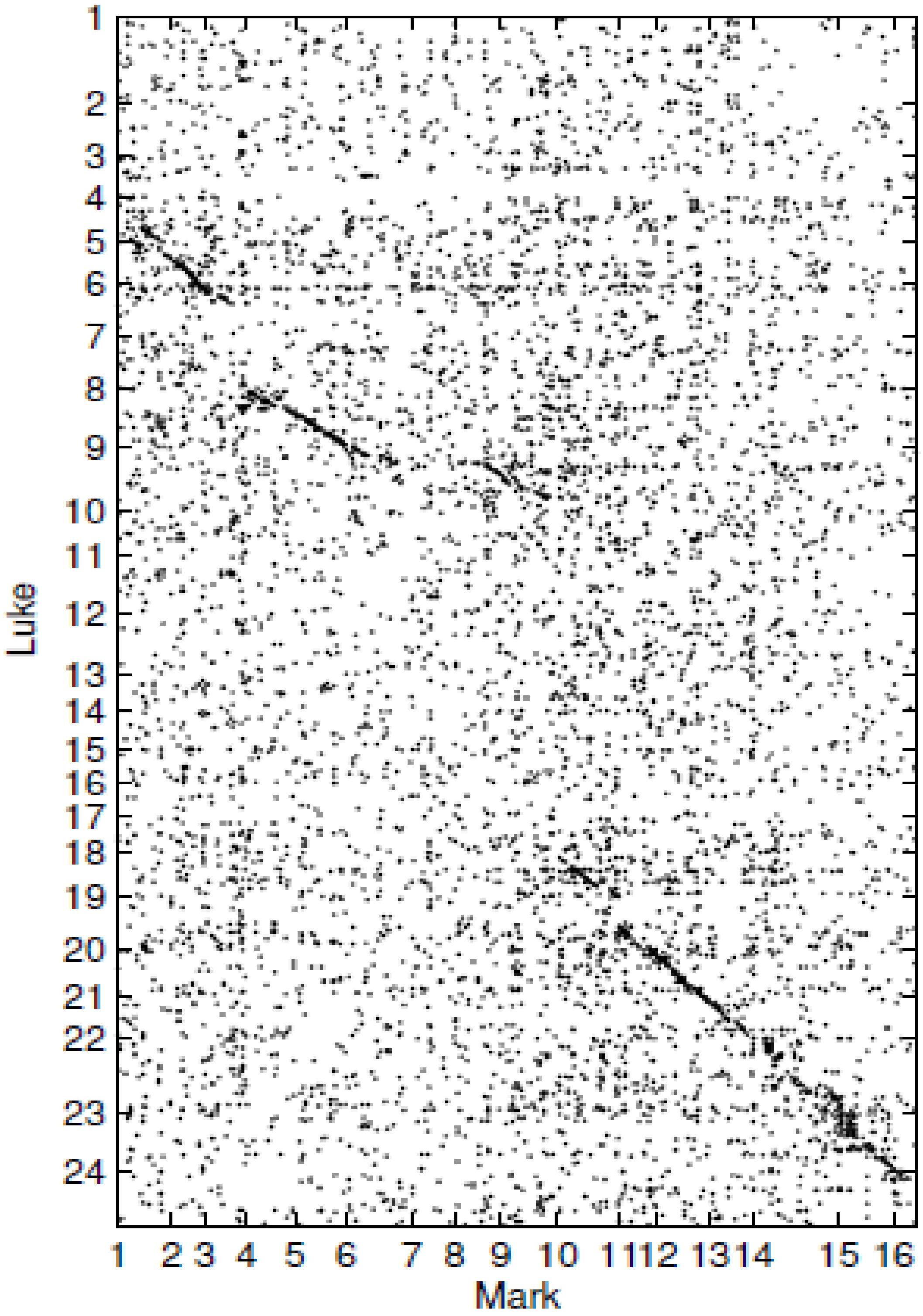
http://www.etraces.e-humanties.net

Google group for Historical Text Re-use:

http://groups.google.com/group/historical-text-re-use

Marco Büchler
Göttingen Centre for Digital Humanities
Georg August University Göttingen, Germany
mbuechler@e-humanities.net

Dotplot view



ACID for the eHumanities

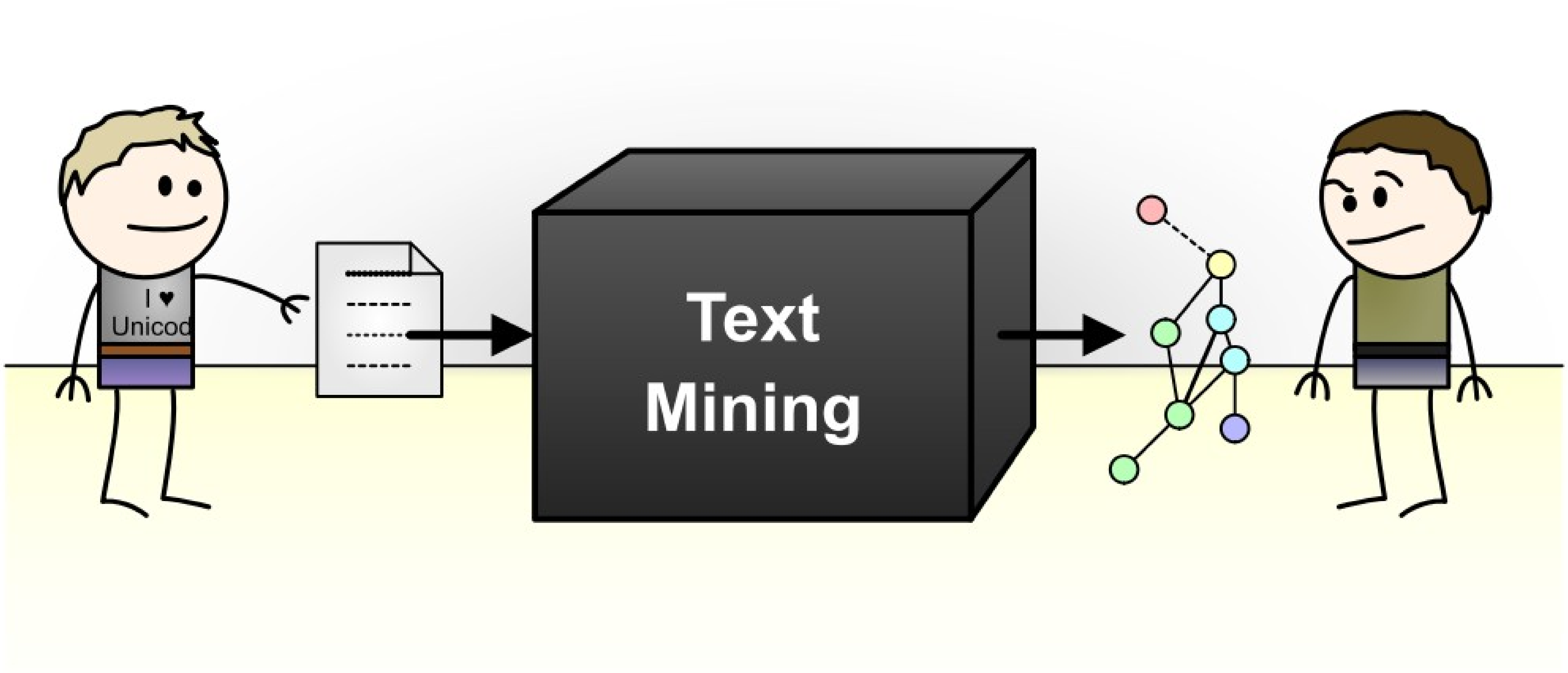
Acceptance

Complexity

Interoperability

Diversity

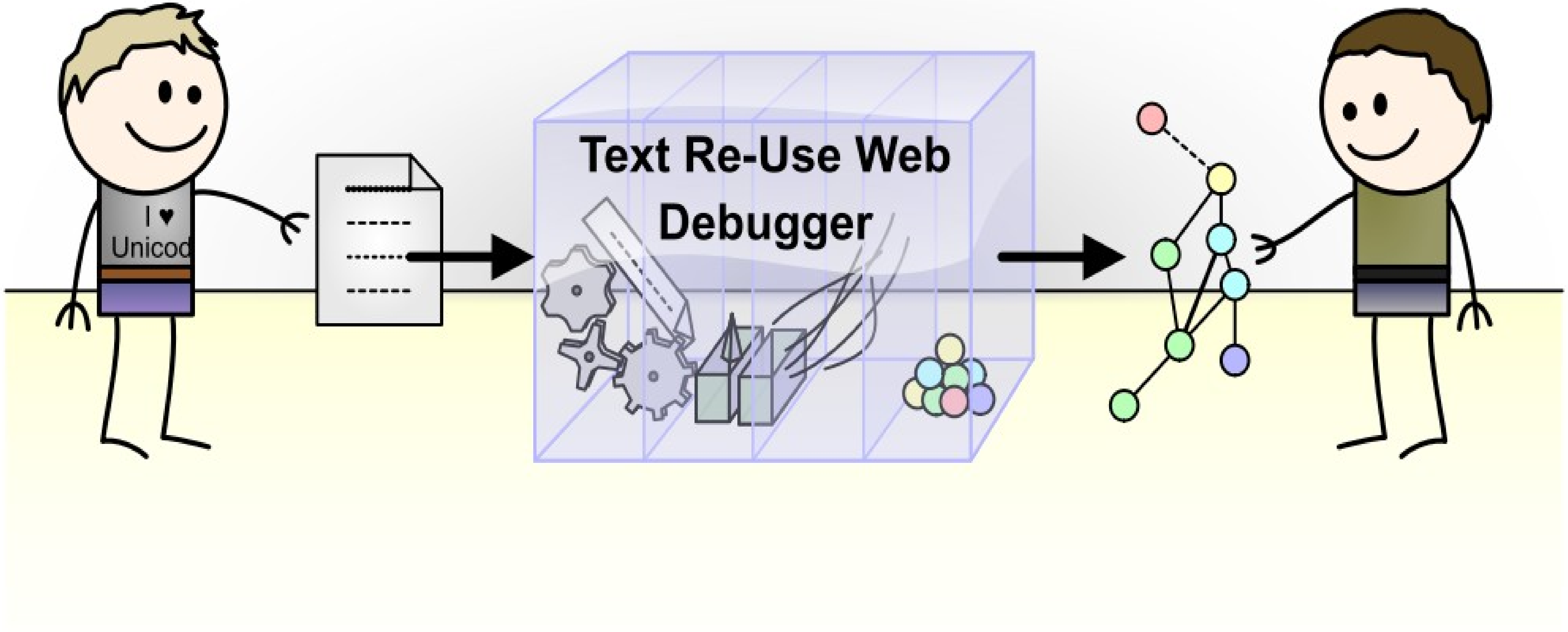
Interdisciplinary collaborations: The problem!



Interdisciplinary collaborations: The problem!

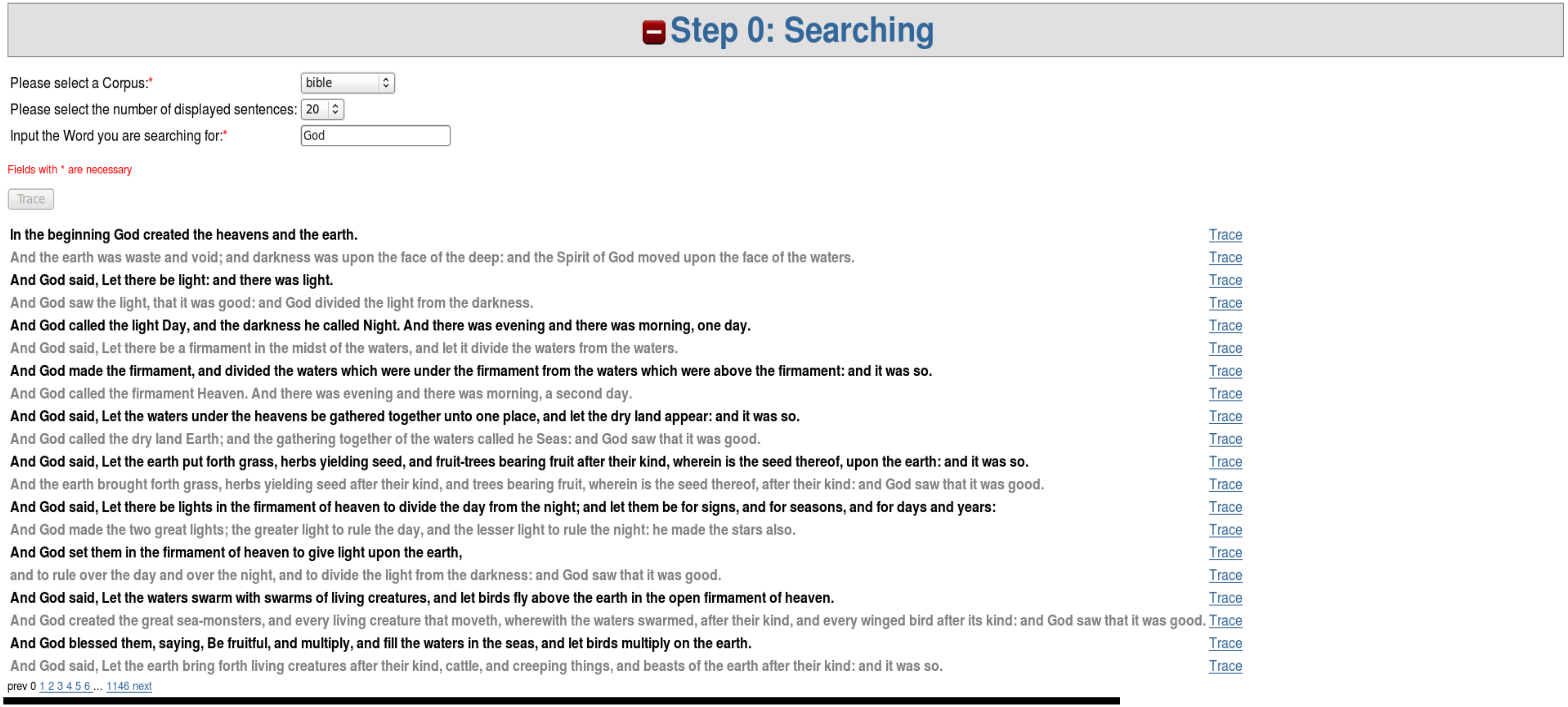
How to get acceptance of humanists if text mining is a black box, that can't be looked in?

Interdisciplinary collaborations: The problem!



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

Jumping into the mining process: Level 0 – Initial request



Step 0: Searching

Please select a Corpus:* bible

Please select the number of displayed sentences: 20

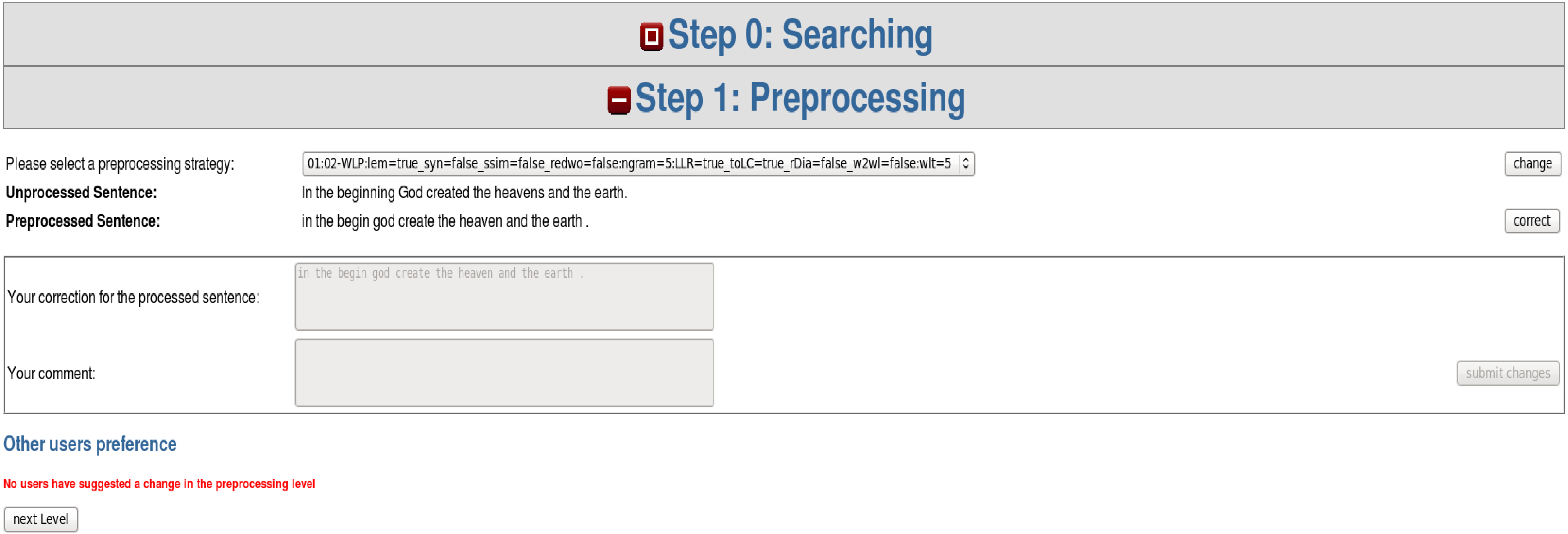
Input the Word you are searching for:* God

Fields with * are necessary

In the beginning God created the heavens and the earth.	Trace
And the earth was waste and void; and darkness was upon the face of the deep: and the Spirit of God moved upon the face of the waters.	Trace
And God said, Let there be light: and there was light.	Trace
And God saw the light, that it was good: and God divided the light from the darkness.	Trace
And God called the light Day, and the darkness he called Night. And there was evening and there was morning, one day.	Trace
And God said, Let there be a firmament in the midst of the waters, and let it divide the waters from the waters.	Trace
And God made the firmament, and divided the waters which were under the firmament from the waters which were above the firmament: and it was so.	Trace
And God called the firmament Heaven. And there was evening and there was morning, a second day.	Trace
And God said, Let the waters under the heavens be gathered together unto one place, and let the dry land appear: and it was so.	Trace
And God called the dry land Earth; and the gathering together of the waters called he Seas: and God saw that it was good.	Trace
And God said, Let the earth put forth grass, herbs yielding seed, and fruit-trees bearing fruit after their kind, wherein is the seed thereof, upon the earth: and it was so.	Trace
And the earth brought forth grass, herbs yielding seed after their kind, and trees bearing fruit, wherein is the seed thereof, after their kind: and God saw that it was good.	Trace
And God said, Let there be lights in the firmament of heaven to divide the day from the night; and let them be for signs, and for seasons, and for days and years:	Trace
And God made the two great lights; the greater light to rule the day, and the lesser light to rule the night: he made the stars also.	Trace
And God set them in the firmament of heaven to give light upon the earth,	Trace
and to rule over the day and over the night, and to divide the light from the darkness: and God saw that it was good.	Trace
And God said, Let the waters swarm with swarms of living creatures, and let birds fly above the earth in the open firmament of heaven.	Trace
And God created the great sea-monsters, and every living creature that moveth, wherewith the waters swarmed, after their kind, and every winged bird after its kind: and God saw that it was good.	Trace
And God blessed them, saying, Be fruitful, and multiply, and fill the waters in the seas, and let birds multiply on the earth.	Trace
And God said, Let the earth bring forth living creatures after their kind, cattle, and creeping things, and beasts of the earth after their kind: and it was so.	Trace

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

Jumping into the mining process: Level 1 – Preprocessing



Step 0: Searching

Step 1: Preprocessing

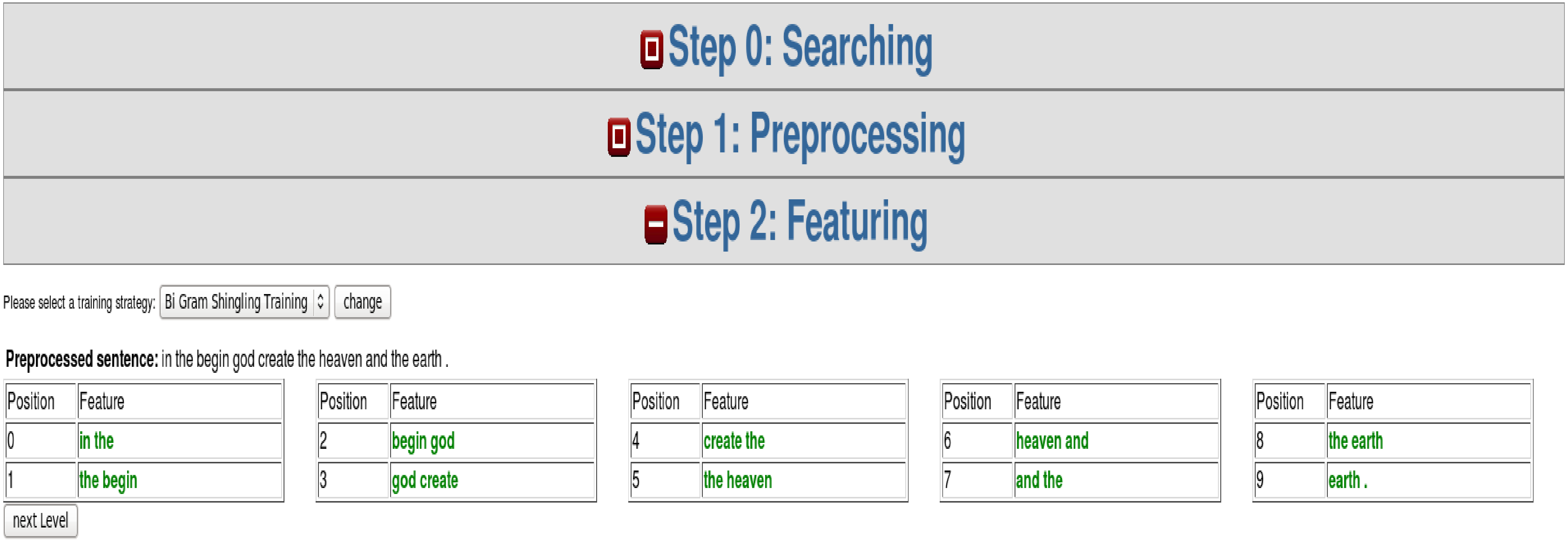
Please select a preprocessing strategy:	01:02-WLP:lem=true_syn=false_ssim=false_redwo=false:ngram=5:LLR=true_toLC=true_rDia=false_w2wl=false:wlt=5	change
Unprocessed Sentence:	In the beginning God created the heavens and the earth.	
Preprocessed Sentence:	in the begin god create the heaven and the earth .	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

Jumping into the mining process: Level 2 – Featuring



Please select a training strategy: Bi Gram Shingling Training

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

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

TRACER

Implemented in TRACER (http://etraces.e-humanities.net/TRACER):

- Tool available in in Q3/4 2013

- Teaching courses (full week) are planned for Q3/4 2013

- More than one million permutations of implementations of the 7 levels possible (06/2013)

- Starting in Q3/2013 to integrate a CTS client for getting data directly from at least one repository

Contacts

For more details:
http://www.etraces.e-humanties.net

Google group for Historical Text Re-use:
http://groups.google.com/group/historical-text-re-use

Marco Büchler
Natural Language Processing Group
Department of Computer Science
University of Leipzig
mbuechler@e-humanities.net