GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

# Probleme und Herausforderungen von Text-Reuse-Erkennung auf großen Datenmengen

***Problems and Challenges of Text Reuse Detection on Massive Data***

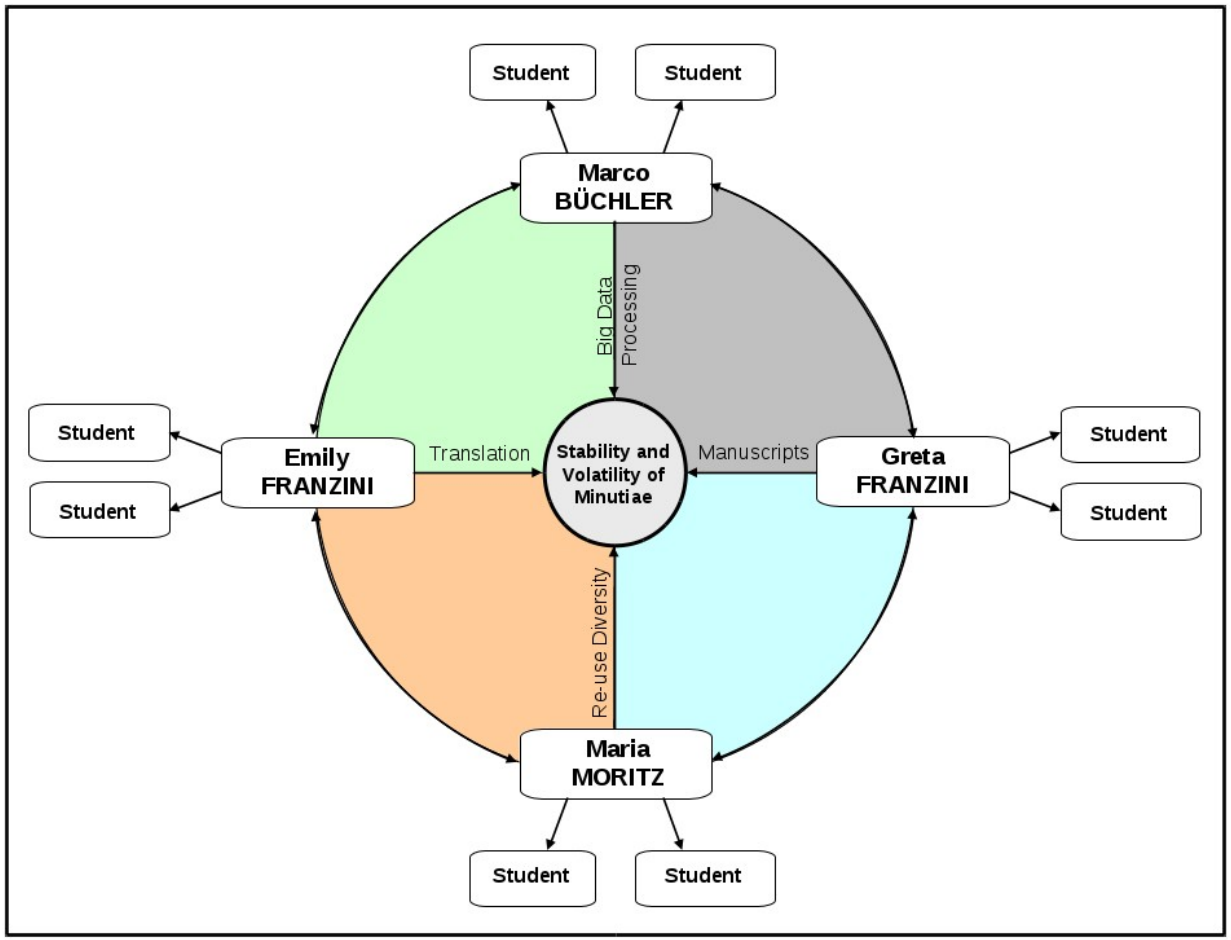
*Marco Büchler, Greta Franzini, Maria Moritz, Emily Franzini*
*eTRAP Research Group*
*Göttingen Centre for Digital Humanities*
*Institute for Computer Science*
*Georg August University Göttingen, Germany*

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



# Question?

What do you associate with text reuse/intertextuality?

# An overview to text reuse

- **General**: Text Re-use 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 monks in the Mid-ages)

# Typical computer scientists' expectation: oversimplification

# Humanists' expectation: oversimplification

# Text Reuse for Humanities and Computer Science

- **Question**: Why is Text Re-use so fundamental for Humanities and Computer Science?
- **Premise**: the amount of digitally available data grows exponentially (Big Data)

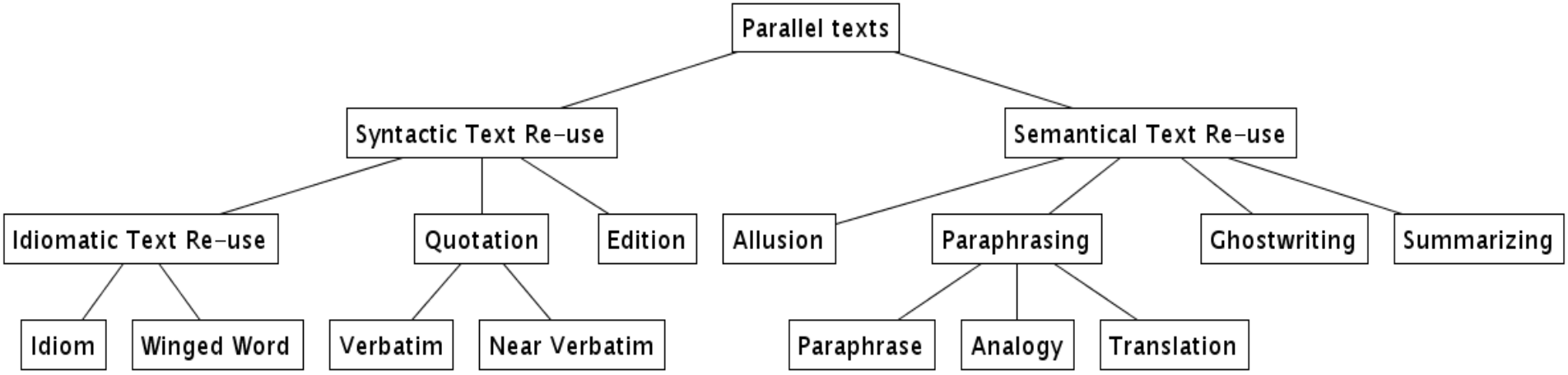
- Humanities:
  - Lines of transmissions and textual criticism
  - Transmissions of ideas/thoughts under different circumstances and conditions

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

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

# Motivation: Naive (computational) 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**.

# Motivation: Naive vs. advanced techniques

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

- **Advanced techniques**:
  - Can break this down to  1-2 processor month for TLG.
  - Simulations on Big Data have shown (given squared complexity) that ...

  - ... by increasing the amount of data to billions words, we need several 100 processor centuries of computational power.

# Scaling Historical Text Reuse

2014 IEEE International Conference on Big Data

## Scaling Historical Text Re-use

Marco Büchler
*Göttingen Centre for Digital Humanities*
*Georg-August-University Göttingen*
*Göttingen, Germany*
*Email: mbuechler@gcdh.de*

Greta Franzini, Emily Franzini, Maria Moritz
*Computer Science Department*
*University of Leipzig*
*Leipzig, Germany*
*Email: [franzini|efranzini|moritz]@informatik.uni-leipzig.de*

***Abstract*—*Text re-use* describes the spoken and written repetition of information. *Historical text re-use*, with its longer time span, embraces a larger set of morphological, linguistic, syntactic, semantic and copying variations, thus adding complication to *text-reuse* detection. Furthermore, it increases the chances of redundancy in a digital library. In *Natural Language Processing* it is crucial to remove these redundancies before we can apply any kind of machine learning techniques to the text. In Humanities, these redundancies foreground textual criticism and allow scholars to identify lines of transmission. Identification can be accomplished by way of automatic or semi-automatic methods. Text re-use algorithms, however, are of squared complexity and call for higher computational power. The present paper addresses this issue of complexity, with a particular focus on its algorithmic implications and solutions.**

***Keywords*-text re-use, performance, scalability**

### I. Introduction

Recent research in Computer Science has witnessed an increased interest in Big Data. Ulrike Rieß, editor of *Spe-*

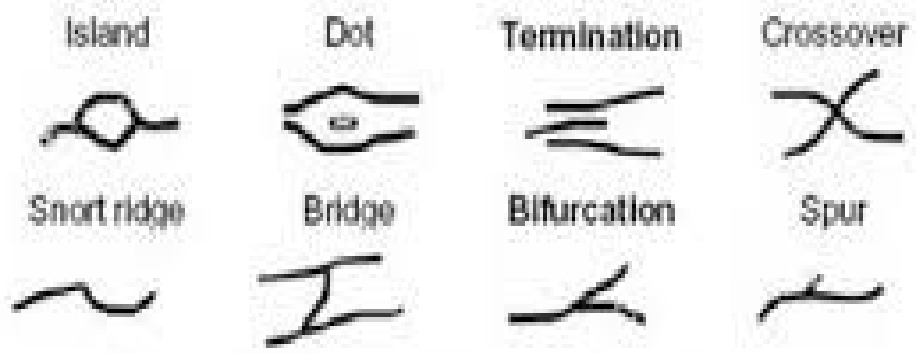
By gradually providing and presenting scholarly Big Data digitally, we are now able to investigate historical texts in broader and more comprehensive ways than ever before. At the same time, the bigger the data, the more difficult it becomes to search and browse large collections. The Digital Humanities contribution to this growth is its support towards the creation of those tools, visualisations and user interfaces, which have now become instrumental in the exploration of mass data, as well as an integral part of digital ecosystems.

In recent years research methodologies in the Humanities have been gradually changing. As recently as thirty years ago, access restrictions to libraries posed numerous challenges to humanists working with printed books and manuscripts. Today, the efforts of mass digitisation provide broader access to these items in digital form. The increasing availability of digitally encoded texts expedites and facilitates the exploration of text patterns. Google's mass digitisation effort, for example, is driving the improvement

# Algorithmic problem: What are the minutiae?

- Question: What are the common primitives in the re-use diversity?

- From biometry (Minutiae):



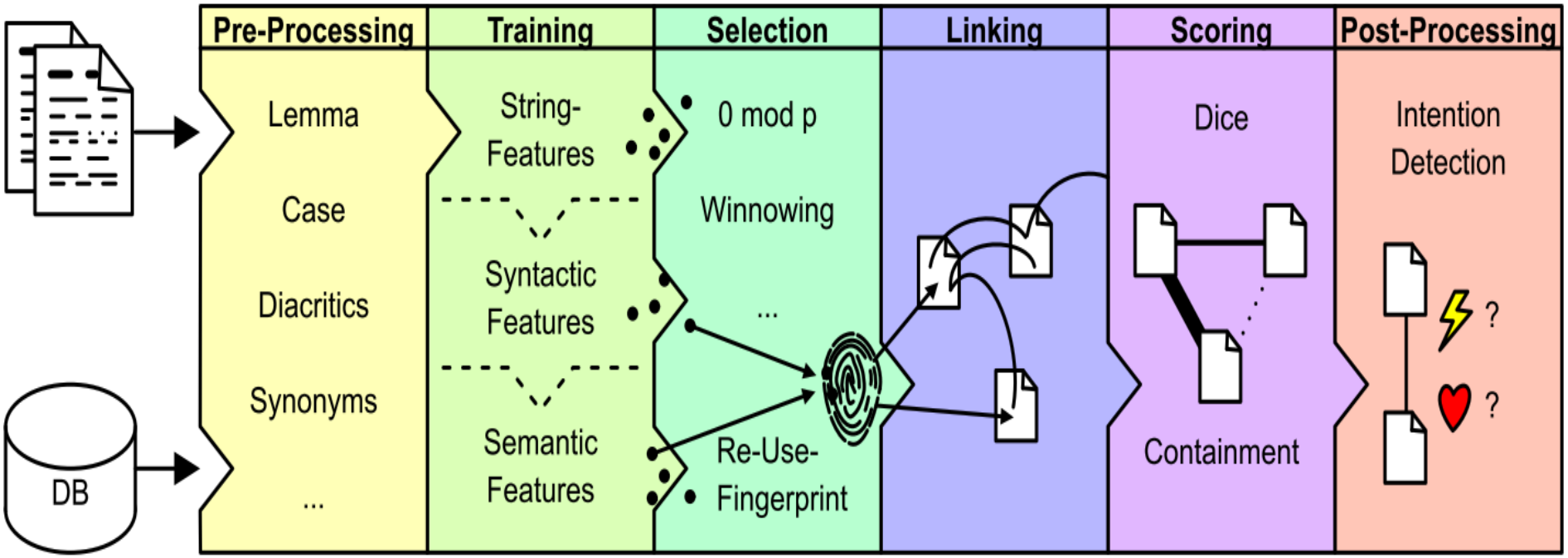

# Algorithmic problem: What are the minutiae?

- An aspect: Easy, we just select content words.

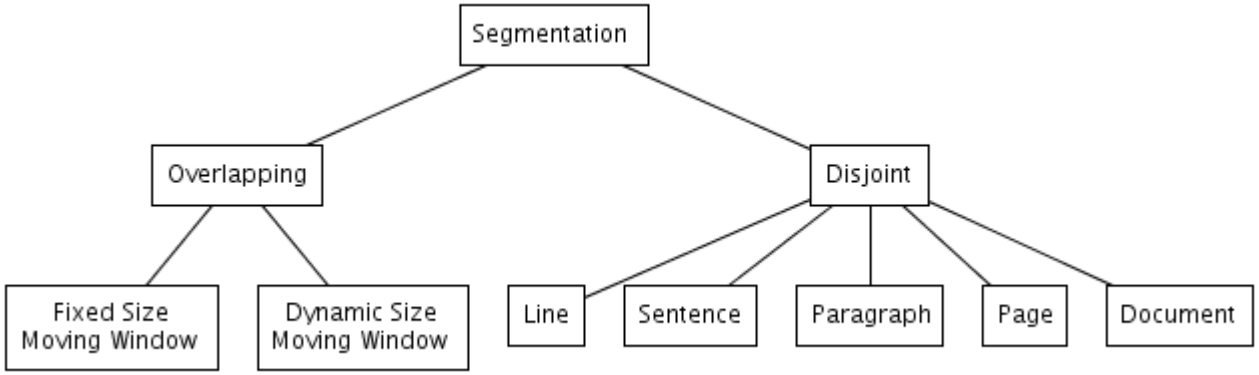
„To be, or not to be, that is the question.“

- Really easy?

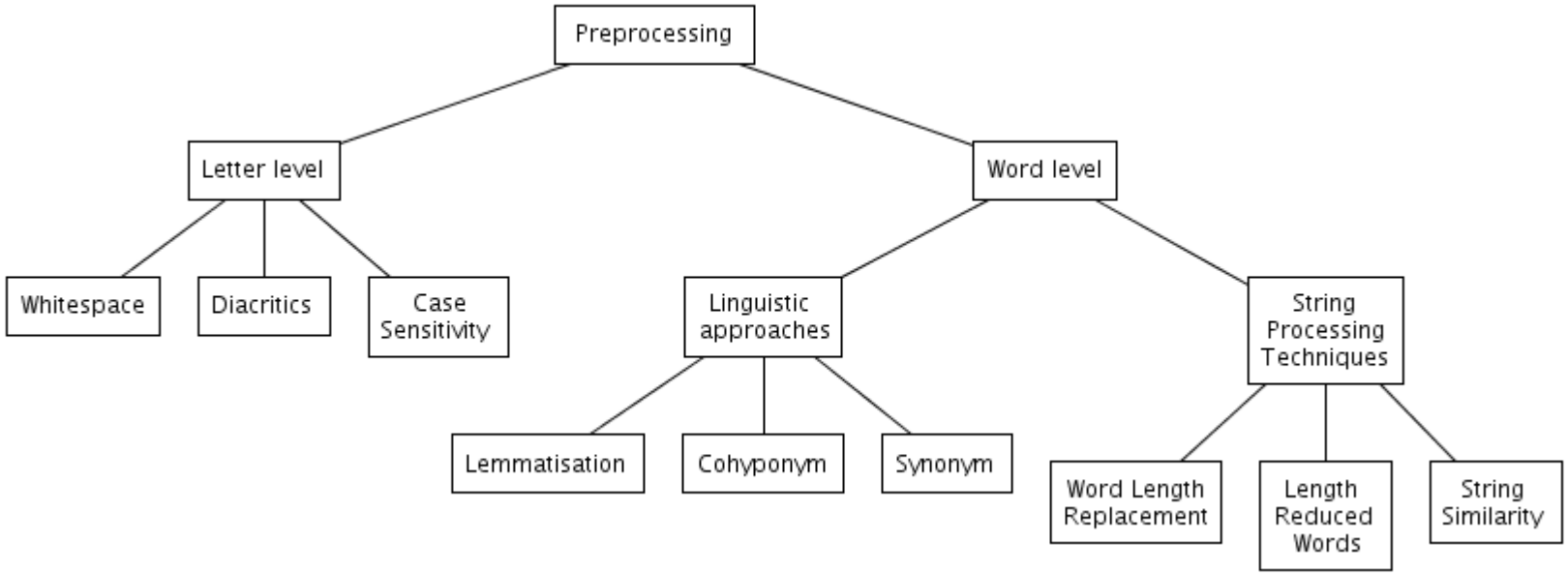
# Recent approach



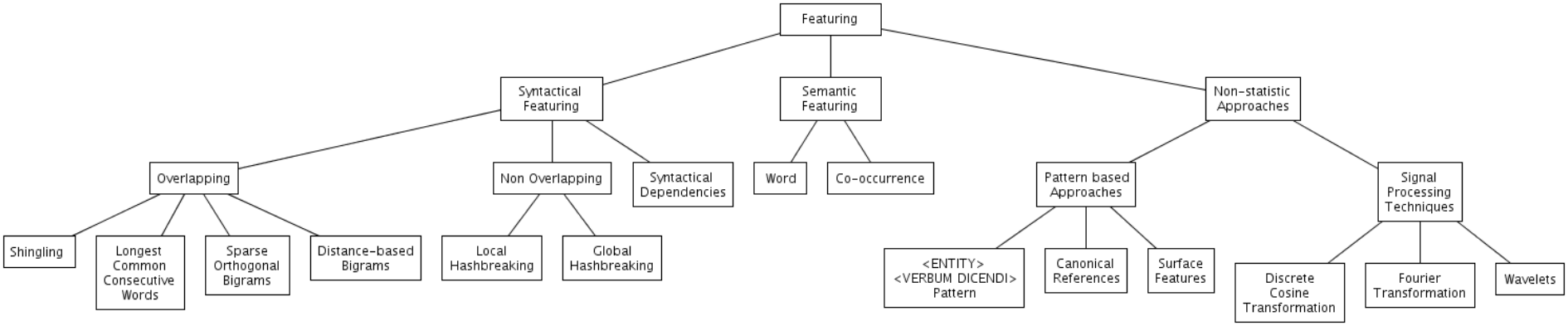
# Segmentation



# Preprocessing



# Preprocessing: Featuring



# 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 version 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

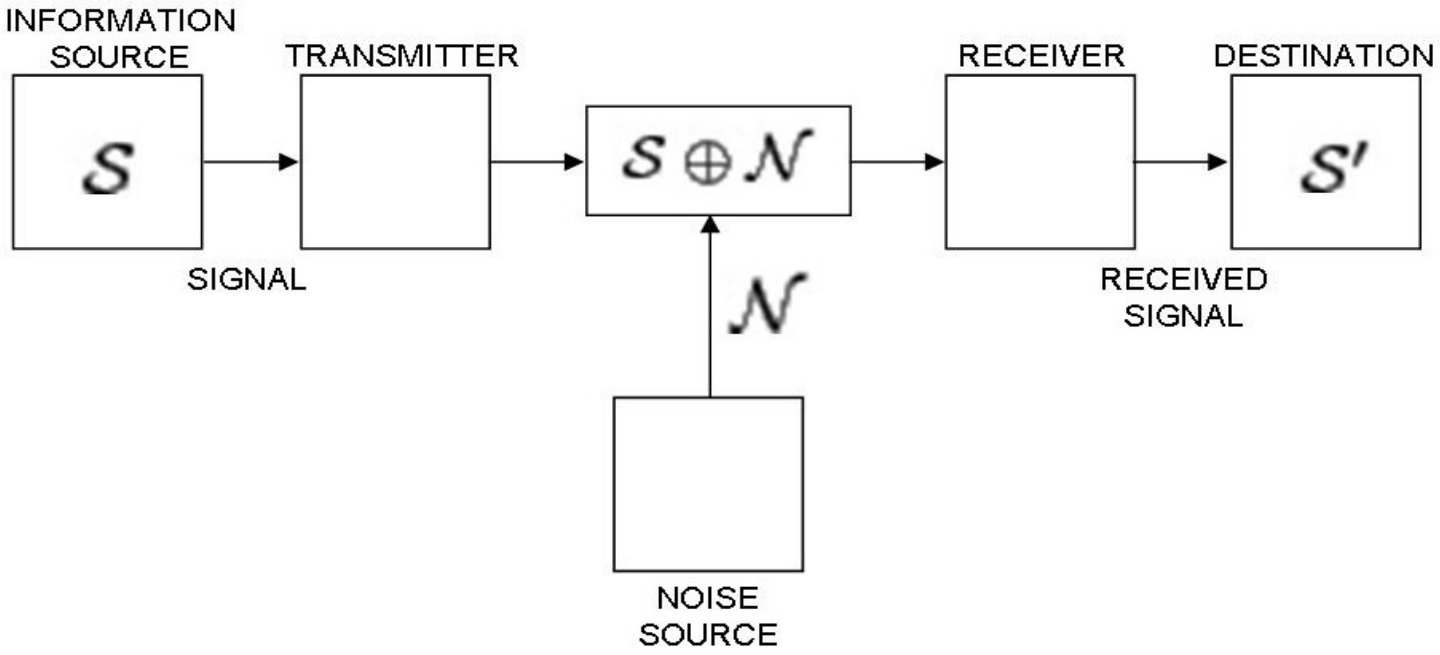
- **Example: book Genesis, chapter 1, verse 1**

| | |
|---|---|
| **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 Bibel versions contain “only” the 28632 verses contained in all seven editions.

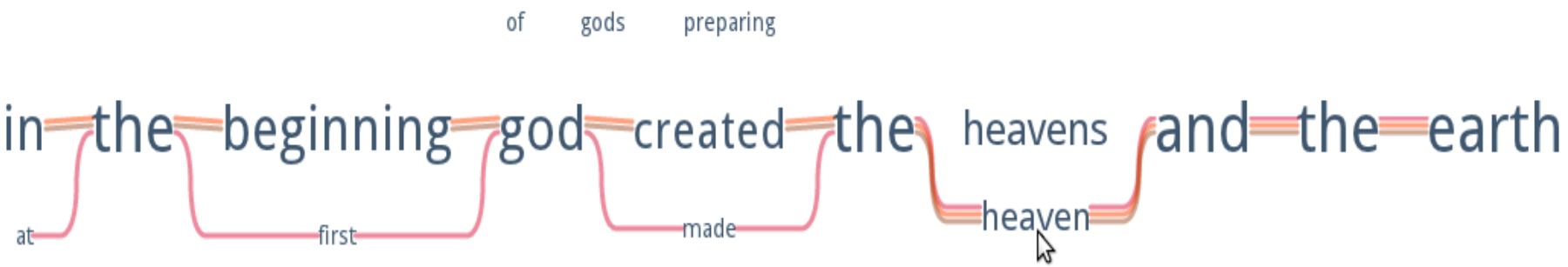
# Framework

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

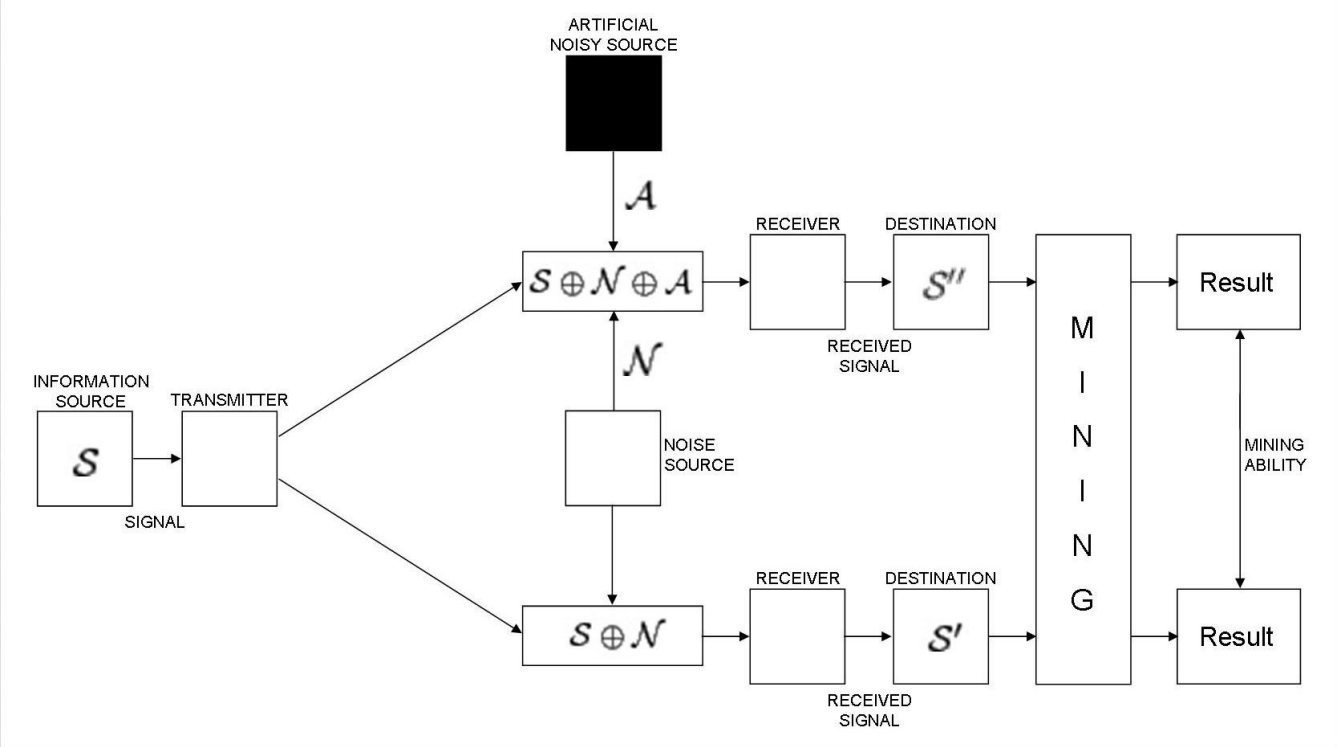


# Dealing and learning from variants

| | |
|---|---|
| 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 -- |



# 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

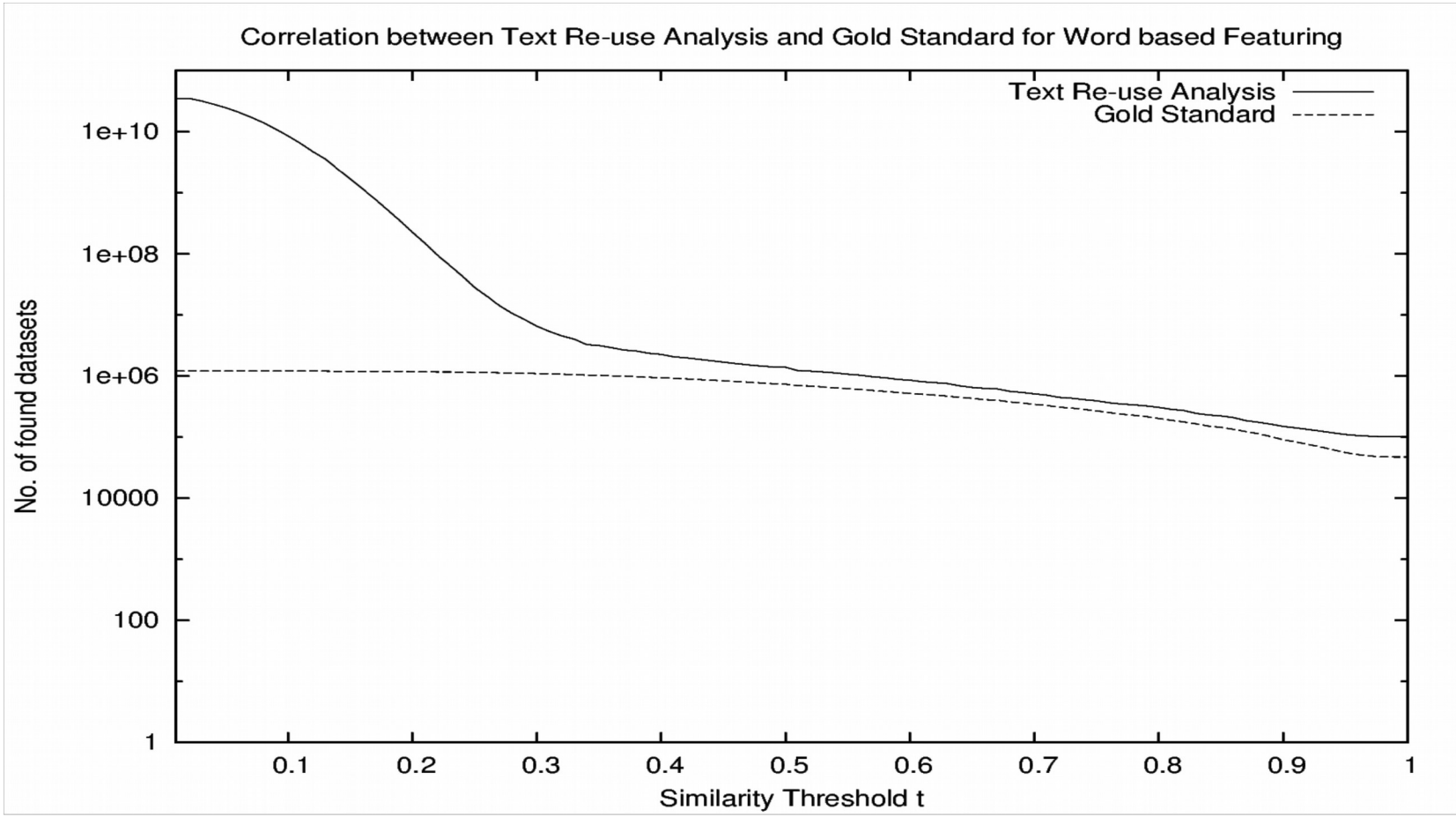
| | Trigram Shingling | | | | Bigram Shingling | | | | Word based Featuring | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_{11}$ | $S_{12}$ | $S_{13}$ | $S_{14}$ | $S_{21}$ | $S_{22}$ | $S_{23}$ | $S_{24}$ | $S_{31}$ | $S_{32}$ | $S_{33}$ | $S_{34}$ |
| ASV vs. BBE | 0.02 | 0.02 | 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 | 0.03 | 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 |

With

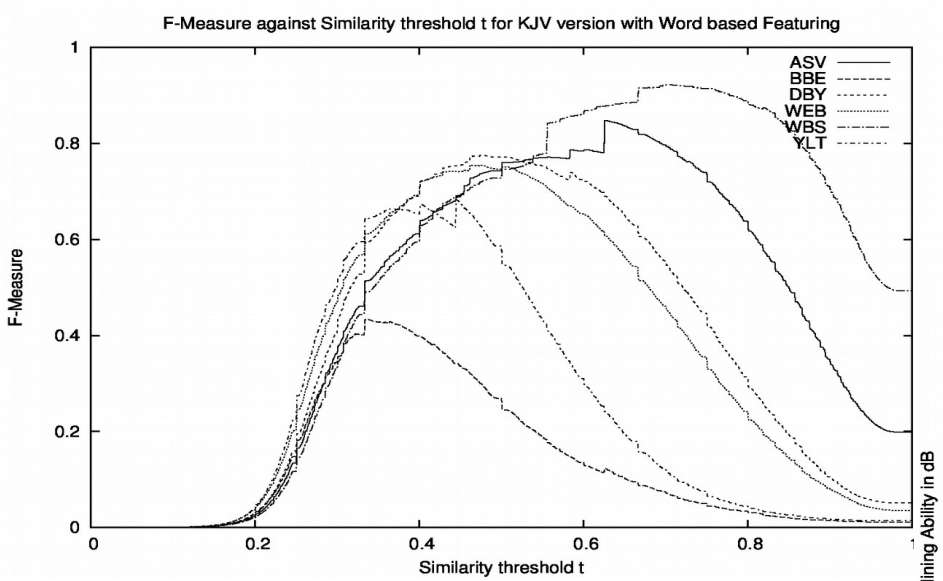
| | Trigram Shingling | | | | Bigram Shingling | | | | Word based Featuring | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_{11}$ | $S_{12}$ | $S_{13}$ | $S_{14}$ | $S_{21}$ | $S_{22}$ | $S_{23}$ | $S_{24}$ | $S_{31}$ | $S_{32}$ | $S_{33}$ | $S_{34}$ |
| ASV vs. BBE | 6.16 | 6.15 | 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.[illegible] | 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.[illegible] | 4.82 | 4.80 |
| DBY vs. WBS | 5.49 | 5.45 | 5.46 | 5.44 | 5.21 | 5.17 | 5.18 | 5.17 | [illegible] | 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[illegible] | 5.19 | 5.13 | 5.10 |
| KJV vs. ASV | 4.97 | 4.95 | 4.96 | 4.95 | 4.80 | 4.78 | 4.79 | [illegible] | [illegible] | 4.47 | 4.47 | 4.47 |
| KJV vs. BBE | 6.35 | 6.30 | 6.34 | 6.32 | 6.00 | 5.97 | 5.99 | 5.9[illegible] | 5.26 | 5.23 | 5.00 | 4.98 |
| KJV vs. DBY | 5.49 | 5.45 | 5.46 | 5.44 | 5.21 | [illegible] | [illegible] | [illegible] | 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 | [illegible] | [illegible] | 4.54 | 4.41 | 4.41 | 4.41 | 4.41 |
| KJV vs. YLT | 6.39 | 6.33 | 6.39 | 6.39 | [illegible] | 6.[illegible] | [illegible] | 6.14 | 5.41 | 5.33 | 5.28 | 5.26 |
| WEB vs. ASV | 5.03 | 5.00 | 5.02 | 5.02 | [illegible] | [illegible] | 4.86 | 4.86 | 4.60 | 4.59 | 4.59 | 4.59 |
| WEB vs. BBE | 6.17 | 6.16 | 6.17 | 6.1[illegible] | [illegible] | [illegible] | 6.00 | 6.01 | 5.30 | 5.27 | 5.26 | 5.22 |
| WEB vs. DBY | 5.69 | 5.65 | 5.67 | 5.65 | [illegible] | 5.39 | 5.40 | 5.38 | 4.85 | 4.82 | 4.82 | 4.80 |
| WEB vs. KJV | 5.57 | 5.52 | 5.55 | [illegible] | 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 |

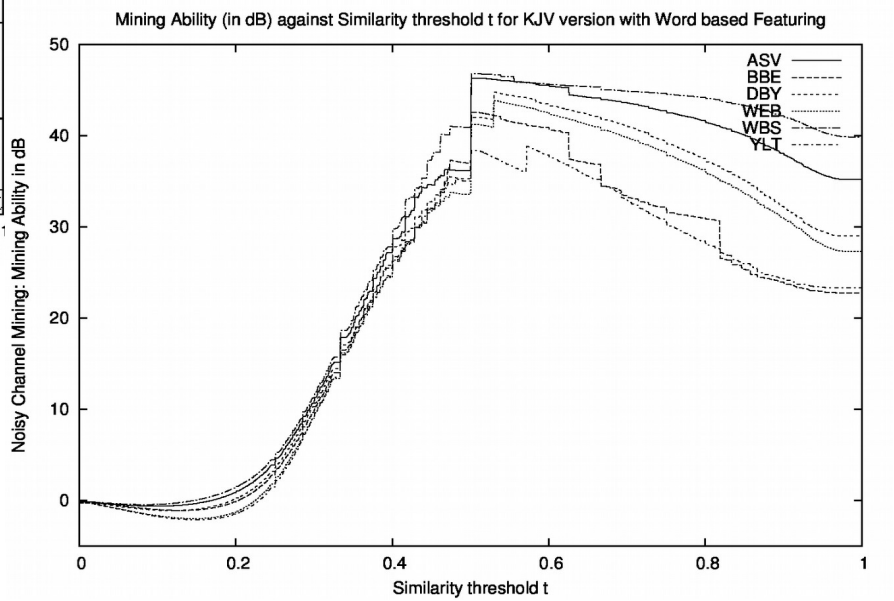
Without

# Dependency of recall and TR compression



# F-Measure and Noisy Channel Evaluation



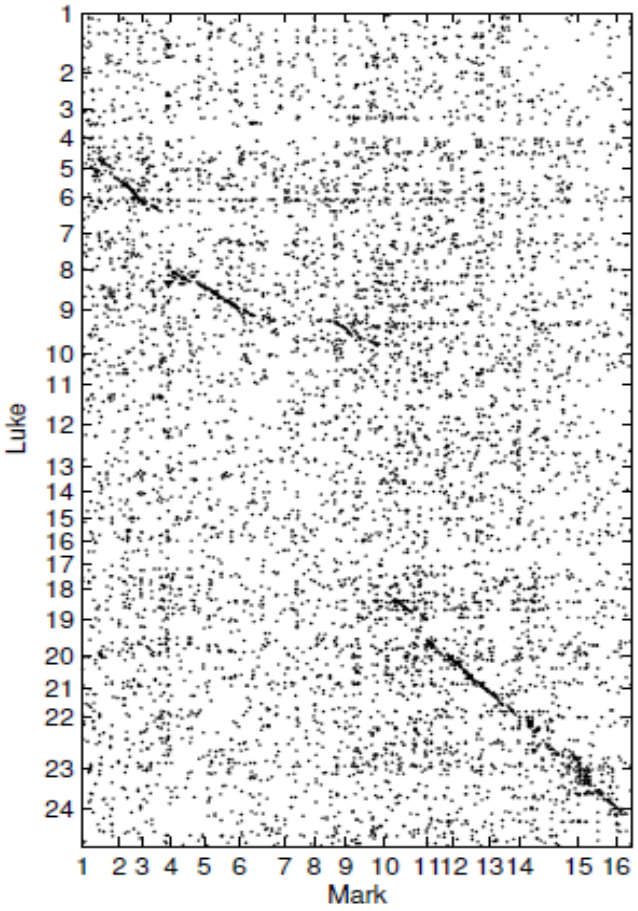

**F-Measure**: *WBS, ASV, DBY, WEB, YLT, BBE*
**NCE**: *WBS, ASV, DBY, WEB, BBE, YLT*

# Relation of preprocessing, featuring, and reuse style

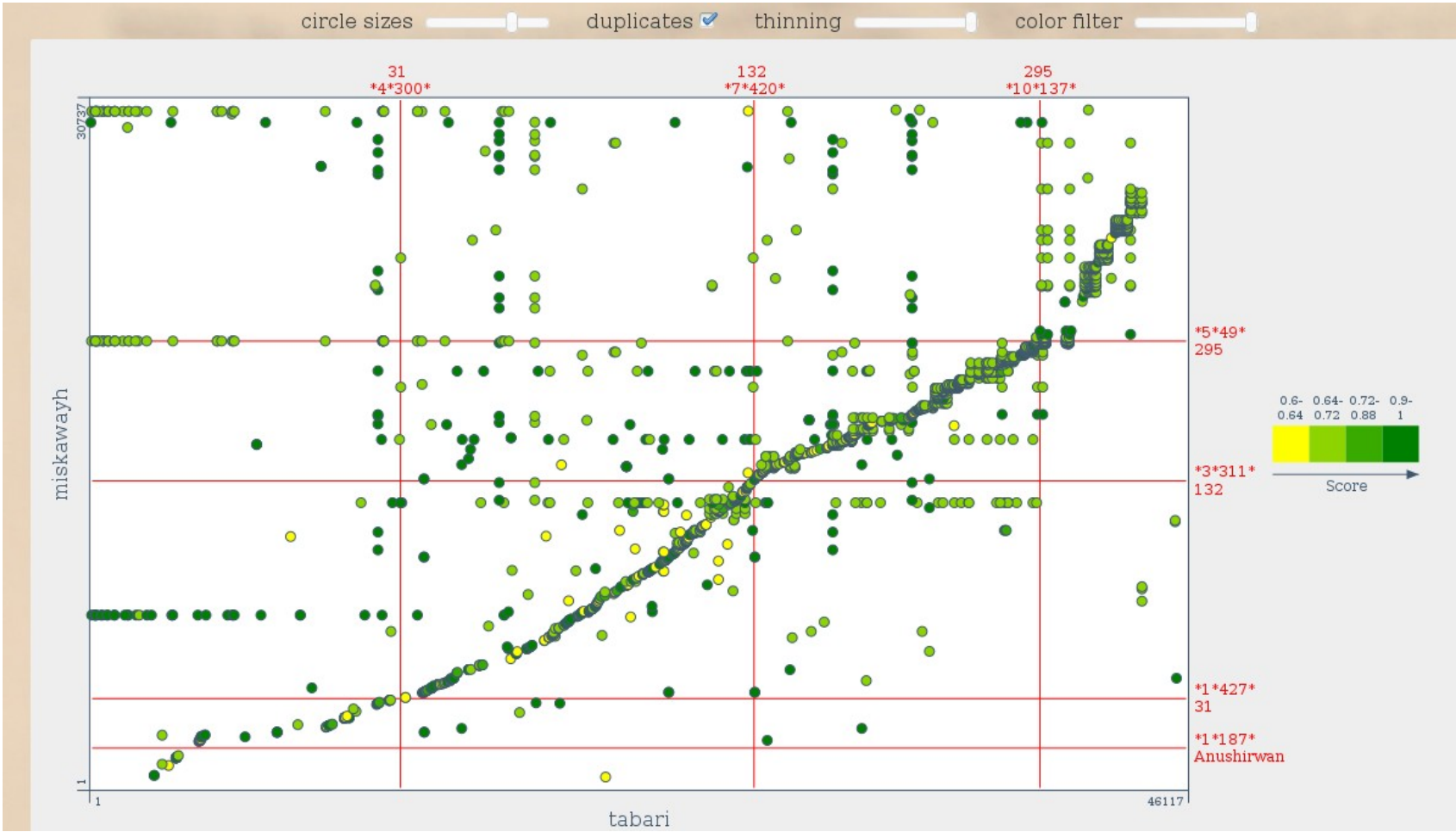
| | ASV | BBE | DBY | WEB | WBS | YLT |
|---|---|---|---|---|---|---|
| $\xi_1$ | (0.38, 0.77) | (0.08, 0.40) | (0.15, 0.74) | (0.15, 0.74) | (0.37, 0.91) | (0.08, 0.52) |
| $\xi_2$ | (0.38, 0.79) | (0.08, 0.40) | (0.16, 0.74) | (0.15, 0.74) | (0.40, 0.92) | (0.08, 0.53) |
| $\xi_3$ | (0.38, 0.78) | (0.08, 0.41) | (0.15, 0.74) | (0.15, 0.74) | (0.39, 0.91) | (0.08, 0.52) |
| $\xi_4$ | (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 |
|---|---|---|---|---|---|---|
| $\xi_1$ | (0.63, 0.84) | (0.34, 0.44) | (0.46, 0.77) | (0.46, 0.75) | (0.70, 0.92) | (0.34, 0.64) |
| $\xi_2$ | (0.63, 0.85) | (0.34, 0.45) | (0.48, 0.78) | (0.46, 0.76) | (0.70, 0.92) | (0.36, 0.65) |
| $\xi_3$ | (0.63, 0.85) | (0.34, 0.43) | (0.48, 0.78) | (0.46, 0.76) | (0.70, 0.92) | (0.44, 0.68) |
| $\xi_4$ | (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

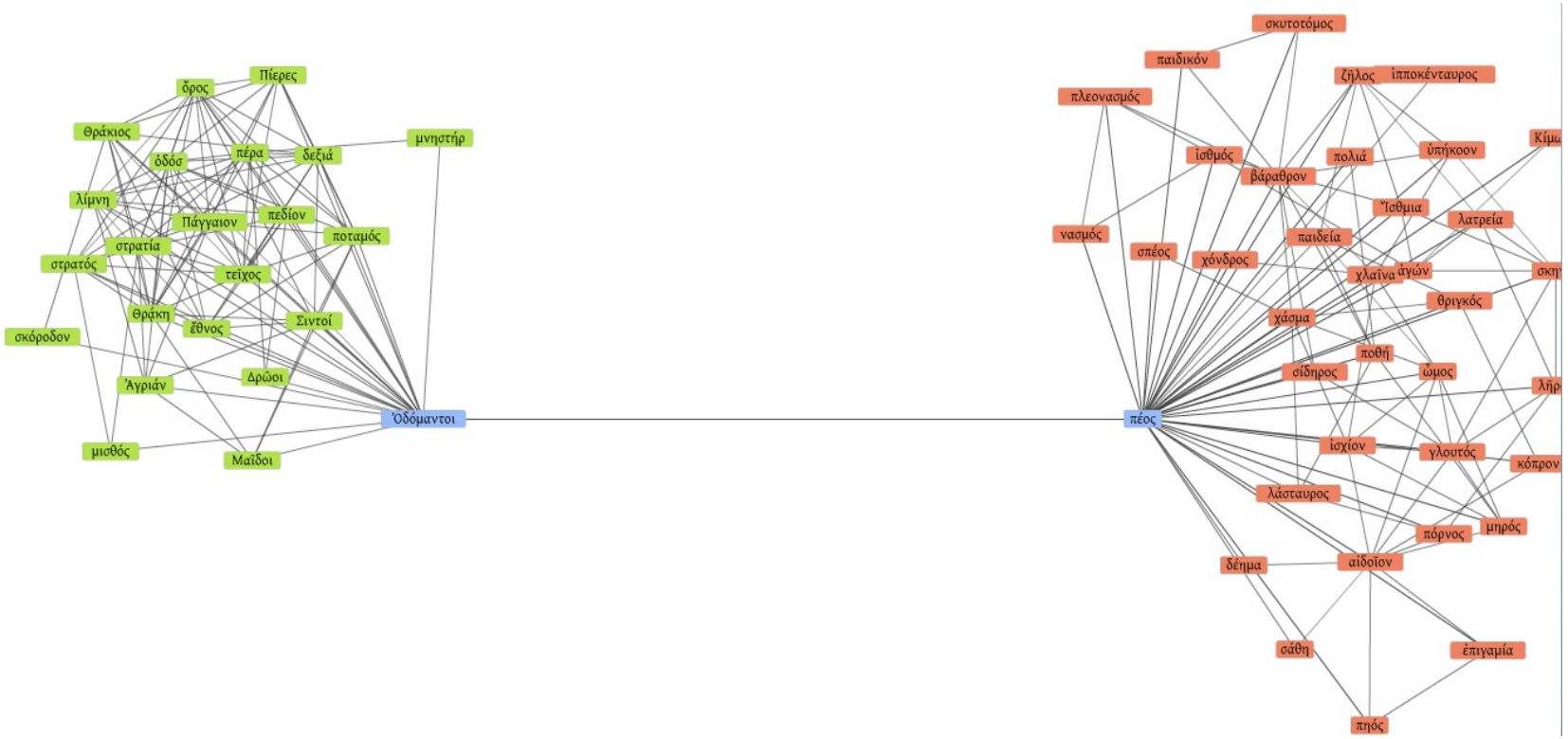


# Dotplot view



# Algorithmic problem: Contrastive semantics (recent work)

- Question: What are the common primitives in the re-use



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

# DH Estonia 2015: attend the eTRAP Workshop!

The one-day eTRAP **Text Reuse Workshop** builds on eTRAP's research activities, some of which deploy Marco Büchler's TRACER tool. TRACER is a suite of algorithms aimed at investigating text reuse in multifarious corpora, be those prose, poetry, in Arabic or Estonian. TRACER provides researchers with **statistical information about the texts under investigation and its integrated reuse visualiser**, the TRACER Debugger, displays occurrences of text reuse in a more readable format for further study.

This workshop seeks to **teach participants to independently understand**, use and run the TRACER tool on their own datasets. However, for the purpose of the workshop and to ensure its smoothest possible running, participants won't be able to investigate their own data but will all be working on the same dataset provided by eTRAP. For those participants who **wish to continue** using TRACER after the workshop on their own corpora, eTRAP will provide the necessary assistance remotely.

**October 21st!!!**

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

# Thank you!

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

eTRAP
Electronic Text Reuse Acquisition Project
INSTITUTE OF COMPUTER SCIENCE
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

SPONSORED BY THE
Federal Ministry
of Education
and Research

**Visit us via http://etrap.gcdh.de**