Computational Linguistics in the Netherlands Journal 15 (2026) 217-253 Submitted 01/2026; Published 06/2026

From Orthography to Semantics: Large-Scale Unsupervised Textual Similarity Detection in Historical Greek

Paulien Lemay[*] PAULIEN.LEMAY@UGENT.BE
Els Lefever[**] ELS.LEFEVER@UGENT.BE
Klaas Bentein[*] KLAAS.BENTEIN@UGENT.BE

[*] *Department of Greek Linguistics, Ghent University*

[**] *Language and Translation Technology Team (LT3), Ghent University*

Abstract

Computational methods for detecting textual similarity provide powerful tools for exploring linguistic patterns, formulaic language, and textual transmission in historical corpora. However, in Ancient Greek studies, these approaches have mostly been tested on small datasets or employed in targeted search tasks. In this paper, we investigate what insights emerge when similarity measures are applied across a large and diverse corpus of Greek texts spanning multiple periods and genres. Our approach is fully unsupervised and does not rely on prior assumptions or predefined queries. We make use of well-established approaches, applying MinHash with locality-sensitive hashing (LSH) to identify repeated orthographic patterns and transformer-based embeddings to capture semantic relationships across texts. We first explore our approaches on the Database of Byzantine Book Epigrams (DBBE), a curated dataset with verse- and epigram-level similarity clusters. Its relatively compact and structured nature makes it an ideal testbed for probing the behavior of the similarity algorithms. We then scale up to a larger, more heterogeneous corpus of Greek texts spanning roughly 400 BC to 1500 AD. Applying MinHash-LSH reveals repeated formulae across textual traditions, while clustering transformer-based embeddings uncovers conceptual and thematic relationships, highlighting recurring motifs and ideas despite orthographic variation. Our findings demonstrate how unsupervised methods suited to high-volume data can uncover structures and relationships that targeted, query-based studies may overlook.

1. Introduction

Computational methods for detecting similarity in ancient texts have proven highly effective for uncovering linguistic patterns, formulaic language, and textual reuse. One frequently used approach focuses on surface-level similarity, which relies on orthographic features such as character- or word-level n-grams, edit distance, or string alignment algorithms. This approach is often used to detect near-copy passages and to support stylistic research (Ochab and Essler 2019, Storey and Mimno 2020). Semantic similarity approaches, in contrast, are used to capture deeper meaning-based relationships between words or texts in order to identify paraphrases, semantic shifts over time, and conceptual connections that go beyond mere lexical overlap (D'Angelo et al. 2025, Krahn et al. 2023, Stopponi et al. 2024, Swaelens 2025).

While both surface-level and semantic approaches have demonstrated their effectiveness, they are often developed and evaluated for specific tasks or narrowly defined research questions, with their design and optimization tailored to particular corpora. As a result, these methods are typically applied in targeted search settings, where researchers already have hypotheses about the types of similarity they wish to detect. This focus limits their usefulness for exploratory analysis of very large corpora, where patterns of similarity may be diverse, subtle, or previously unknown. Moreover,

many similarity detection techniques become computationally demanding when applied at large scale, making it challenging to analyse corpora consisting of millions of textual units in an efficient and unsupervised manner.

To address this challenge, we propose a corpus-driven, task-agnostic methodology that integrates both surface-level and semantic similarity analyses. Surface-level similarity is captured using MinHash with locality-sensitive hashing (LSH), which efficiently detects repeated or highly similar sequences across large corpora (Broder 1997). To capture semantic similarity, we generate transformer-based sentence embeddings, which we use to construct a graph via nearest-neighbour retrieval. Together, these complementary approaches enable the exploration of large, unstructured corpora, revealing a broad range of textual patterns.

We begin by examining the Database of Byzantine Book Epigrams (DBBE), a manually curated collection of epigrams grouped into clusters of recurring formulaic expressions across orthographic, semantic, and thematic dimensions, based on expert scholarly judgment (Deforche et al. 2024, Swaelens 2025). This structured and interpretable reference provides a useful environment for exploring the behavior of our methods. Building on insights from this controlled corpus, we then scale the approach to a larger and more heterogeneous collection of non-literary Greek texts spanning the Archaic to Byzantine periods, investigating whether unsupervised methods can reveal broader patterns and regularities across centuries of textual production.

By adopting this approach, we aim to make the following contributions:

- We introduce a **corpus-driven framework** for detecting similarity, which minimizes prior assumptions about the types of similarity to be identified, thereby allowing patterns to emerge organically from the data.
- We advance an **unsupervised methodology** that does not rely on labeled training data, facilitating the analysis of large, heterogeneous corpora across diverse textual genres and historical periods.
- We demonstrate that **scalable similarity detection** can reveal latent structures of formulaicity, textual transmission, and intertextual relationships across centuries, providing new insights into patterns of literary composition and manuscript culture.

The remainder of this paper is organized as follows. In Section 2 (Related Work), we review prior applications of computational similarity techniques to ancient languages. Section 3 (Data) introduces the corpora employed for applying the proposed clustering methodology. In Section 4 (Computational Methods), we provide a detailed description of the selected algorithms. Section 5 (Experiment Setup) outlines the procedure to apply the proposed algorithms to the data. Finally, Sections 6 and 7 present the results, evaluate their significance, and discuss the broader implications of our findings.

2. Related Work

Recent work in machine learning for ancient languages has highlighted how computational methods are transforming the study of historical texts, enabling large-scale analyses across diverse corpora and tasks (Sommerschield et al. 2023). Within this broader landscape, similarity methods constitute an important field of study. In the following section, we discuss the potential of both surface-level and semantic similarity methods, reviewing key contributions and outlining the approaches adopted here, with technical details provided in Section 4.

2.1 Surface-Level Similarity

Surface-level similarity has been widely explored in historical stylometry, a field frequently employed to attribute authorship to disputed texts. For example, in *Stylometry of Literary Papyri* (Ochab and Essler 2019), distance-based clustering is applied to literary papyri by representing each text as a vector of its most frequent words, allowing patterns of writing style to emerge and enabling evaluation against known author labels.

Using a comparable representation, *Like Two Pis in a Pot* (Storey and Mimno 2020) builds on this approach but shifts the focus from patterns that characterise individual authors to patterns that recur across texts regardless of authorship. This makes it possible to detect recurring stylistic patterns that are not tied to any single writer, showing that texts by different authors and separated by centuries can nevertheless be closely related in stylistic terms. These findings suggest that surface-level similarity captures not only individual writing habits, but also broader literary conventions and practices.

Focusing on a different level of recurrence, recent work on Byzantine book epigrams (Giannikou et al. 2024) demonstrates that surface-level patterns can also be used to study formulaic structure within a textual tradition. Rather than focusing on overall stylistic similarity between texts, it identifies recurring sequences of words and shows how these formulaic expressions persist, vary, and are recombined across Byzantine book epigrams. This shifts the emphasis from similarity between whole texts to repetition at the level of phrases and conventional expressions, offering a complementary perspective on how textual reuse operates within a genre.

In parallel to these application-driven studies, recent methodological work (Lemay et al. 2026) explores how surface-level similarity detection can be scaled to large and heterogeneous corpora. Using character-level MinHash combined with locality-sensitive hashing, this line of work prioritises computational efficiency and robustness in identifying closely related textual segments, enabling similarity search in settings where exhaustive pairwise comparison would be impractical. Rather than targeting specific literary or philological questions, it provides a general-purpose framework for scalable similarity computation, supporting a wide range of downstream analyses in digital humanities research.

In this study, we build on the approach proposed by Lemay et al. (2026). While the proposed methodology demonstrates strong performance and scalability, it focuses on algorithmic quality rather than what insights can be gained from large, diverse corpora. Here, we extend this work by applying MinHash-LSH to a heterogeneous corpus, examining both scaling challenges and the qualitative patterns revealed in the resulting clusters.

2.2 Semantic Similarity

An early approach to semantic similarity detection is the work by Berger et al. (2016), who examine the reuse of Biblical texts. Their study leverages Ancient Greek WordNet[1] and lemma lists to identify synonyms, hypernyms, and part-of-speech changes between reused and source texts, complemented by manual analysis. The results show that reuse is often highly non-literal, highlighting the limitations of surface-level similarity measures.

More recent research has focused on building sentence embeddings that capture semantic meaning. However, generating high-quality embeddings requires substantial training data, which is often lacking for low-resource languages. One strategy to overcome this limitation is multilingual knowledge distillation, in which low-resource models are aligned with high-resource models through parallel

1. `https://greekwordnet.chs.harvard.edu/`

sentences, allowing semantic knowledge to be transferred across languages. Krahn et al. (2023) apply this approach to Ancient Greek and English, demonstrating that models trained in this way effectively support translation search, semantic textual similarity (STS), and cross-lingual retrieval tasks. Similarly, Riemenschneider and Frank (2023) apply multilingual knowledge distillation to align Latin, Ancient Greek, and English embeddings in order to identify translations and retrieve semantically similar sentences across these languages. Complementary strategies have also been proposed to improve sentence embeddings, including contrastive learning to sharpen semantic distinctions (D'Angelo et al. 2025) and genre-aware embeddings that integrate metadata into semantic representations (Perrone et al. 2019).

An alternative approach focuses on capturing linguistic structure at the word and subword level. Riemenschneider and Krahn (2024) address low-resource historical language processing by employing character-aware hierarchical language models (HLMs). These models integrate an intra-word encoder that captures fine-grained character-level patterns with an inter-word encoder that produces contextualized word embeddings, encoding both syntactic and semantic information. This architecture is particularly well-suited to languages with substantial spelling variation and rich morphology, while also supporting the learning of context-dependent meanings through distributional semantics.

Alongside these modeling advances, benchmark resources have been developed to systematically evaluate semantic similarity. Stopponi et al. (2024) created AGREE, a benchmark for distributional semantic models of Ancient Greek, while Swaelens et al. (2025) propose a benchmark for Byzantine Greek, providing frameworks for sentence-level evaluation in historical texts.

Building on these developments, we employ the open-source model from Riemenschneider and Krahn[2], which integrates a training mechanism based on multilingual knowledge distillation with a hierarchical language modeling architecture. While their experiments demonstrate strong performance on STS for Ancient Greek, they focus primarily on sentence-level evaluation and do not explore scaling to larger corpora or unsupervised clustering. In this study, we extend their work by investigating how these embeddings can support large-scale comparative analysis.

3. Data

To evaluate our approach, we assemble a dataset combining texts from multiple corpora across diverse geographical regions and historical periods. This allows us to evaluate the selected methods on a sufficiently large volume of data, ensuring that their scalability can be meaningfully assessed. For copyright reasons, we cannot publish the full corpus directly. However, we provide a detailed description of its composition and sources, enabling full reconstruction of the dataset. Table 1 summarizes the size of each of the data sources, which are introduced and discussed in the following paragraphs.

Source dataset	Total n words	Total n lines	Total n documents
DBBE	318 068	52 701	12 634
Papyri	5 231 385	839 394	55 178
PHI	7 053 539	1 454 517	194 803
BIE	21 642	4 391	794

Table 1: Total number of words, lines, and documents contained in every source dataset.

2. https://huggingface.co/kevinkrahn/shlm-grc-en

3.1 Database of Byzantine Book Epigrams

The first part of our corpus is the Database of Byzantine Book Epigrams (DBBE)[3]. This dataset contains over 12 000 epigrams dating from approximately 500 to 1500 AD, aggregating texts that were previously dispersed or unpublished. From the margins of the manuscripts, these epigrams organize content, recognize patrons, scribes, and authors, and provide guidance to readers (Ricceri et al. 2023).

A key feature of DBBE is its curated Verse Group and Type classifications. Verse Groups cluster similar verses across epigrams, whereas Types provide a normalized, reconstructed version of one or more Occurrences[4]. These groupings are manually curated by scholars, reflecting individual expertise, interpretive judgment, and editorial decisions, rather than any formal metric of similarity (Swaelens 2025). Despite this limitation, DBBE uniquely provides a controlled setting in which textual patterns are well-attested. Previous studies have demonstrated both semantic and orthographic formulaicity in these epigrams (Deforche et al. 2024, Swaelens 2025), meaning that we can assume repeated patterns exist across the dataset. This makes DBBE particularly suitable for examining how the selected clustering methods respond to both surface-level repetition and conceptual relationships. Its relatively small size and clear structure allow us to observe algorithmic behavior before extending the approach to the full dataset, where formulaicity is less pervasive and patterns are less predictable.

The data was obtained from the publicly available SQL dump on Zenodo (Demoen et al. 2023) and processed using a local containerized PostgreSQL setup. The identifiers used to refer to DBBE entities correspond to the identifiers as they can be found on the DBBE website.

3.2 Byzantine Inscriptional Epigrams

The four volumes of the series *Byzantinische Epigramme in Inschriftlicher Überlieferung* (Rhoby et al. 2009, Rhoby 2010, Rhoby 2014, Rhoby 2018) present a comprehensive corpus of Byzantine epigrams preserved on the material supports for which they were composed. The volumes cover a wide range of frescoes, mosaics, icons, metalwork, ivory, wood, textiles, and manuscripts, with entries organized by object type and the region where the objects are currently located. Each edition includes detailed object descriptions, the Greek text, critical apparatus, German translation, and commentary. Building on the work of Wolfram Hörandner, the project follows clearly defined inclusion criteria. Its chronological scope extends from approximately 600 to 1500 AD, encompassing the height of the Byzantine dodecasyllabic epigram tradition and complementing other projects focused on earlier periods (Delouis 2012).

As part of this study, we created a dataset of Byzantine Inscriptional Epigrams (BIE) by processing the four volumes of this series[5]. To extract epigrams, we combined Unicode-based detection of Greek characters with a RoBERTa-based language classification model (Liu et al. 2019). Unicode detection flags all text containing Greek letters, but many lines with Greek characters are not epigrams (e.g., citations or editorial notes). RoBERTa-based language identification considers the broader context and predicts whether a passage is predominantly Greek based on learned linguistic patterns. This distinction would otherwise require a large set of complex rules based on line length, punctuation, or layout. Using this approach we systematically digitized and organized the

3. `https://www.dbbe.ugent.be/`
4. Attested versions of epigrams as they appear in individual manuscripts.
5. We are grateful to Prof. Andreas Rhoby for providing the original Word documents of his edition, which significantly simplified the extraction and processing of the epigrams from the source material.

epigrams. Each epigram is identified by a composite key consisting of the chapter title and the individual number assigned in the original editions.

3.3 Integrating Digital Papyrology

With the *papyri.info* platform, the Integrating Digital Papyrology (IDP) project aims to unify various papyrological resources, linking digital transcriptions with metadata and images (Baumann 2013). The platform offers two core components. On the one hand, there is the Papyrological Navigator (PN), which is a robust search and browsing interface that aggregates major datasets such as the Advanced Papyrological Information System (APIS), the Duke Databank of Documentary Papyri (DDbDP), the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV) and the Bibliographie Papyrologique (BP). On the other, there is the Papyrological Editor (PE), which supports community-driven peer-reviewed editorial contributions in EpiDoc-encoded TEI XML. Launched in 2010, the site has brought together more than 50 000 digitized Greek and Latin papyri, mostly dated between the 4th century BC and the 8th century AD.

The dataset used in our experiments was created using the database of the Ghent University EVWRIT project (Bentein 2025). The identifiers used to refer to the papyrus texts are the identifiers used by Trismegistos (Depauw and Gheldof 2013).

3.4 PHI Greek Inscriptions

The Packard Humanities Institute's (PHI) Greek inscriptions database[6] provides searchable access to approximately 200 000 Greek inscriptions, spanning a chronological range from the 8th century BC to the 6th–7th century AD. The database emphasizes textual accessibility and breadth. It provides plain-text transcriptions without a critical apparatus and integrates material from major printed corpora, including Inscriptiones Graecae and regional epigraphic collections. Although its minimal editorial markup limits its direct use for critical editions, its structured format and consistent encoding make it an indispensable tool for digital epigraphy, paleography, historical linguistics, and the study of regional and chronological variation in the Greek language. Data was collected and formatted using the scripts shared in the I.PHI project (Sommerschield et al. 2021). The identifiers used to refer to the PHI texts are the identifiers used by Trismegistos (Depauw and Gheldof 2013).

4. Computational Methods

In this section, we provide in-depth background on the algorithms introduced in Section 2 and applied in Section 5, explaining how they support the construction of both surface-level and semantic clusters.

4.1 Surface-Level Similarity: MinHash-LSH

In this study, we measure textual resemblance at the surface level using MinHash with locality-sensitive hashing. These techniques efficiently approximate pairwise similarity while scaling to large corpora.

6. `https://inscriptions.packhum.org/`

4.1.1 Challenge: Scaling Pairwise Textual Comparison

A common method for quantifying similarity between documents is Jaccard similarity, which compares sets of n-grams by dividing the size of their intersection by the size of their union (Jaccard 1901). For example, consider two short passages represented as sets of character trigrams:

$$A = \{\text{aaa}, \text{bbb}, \text{ccc}\}, \quad B = \{\text{bbb}, \text{ccc}, \text{ddd}\}.$$

Their union is $\{\text{aaa}, \text{bbb}, \text{ccc}, \text{ddd}\}$ and their intersection is $\{\text{bbb}, \text{ccc}\}$, giving a Jaccard similarity of $2/4 = 0.5$.

Jaccard similarity is easy to compute for small datasets, but the number of required comparisons explodes as the dataset grows, as each document must be compared to every other document. This means that doubling the number of documents quadruples the work needed.[7] For example:

- 1 million documents require about 500 billion comparisons.
- 2 million documents require about 2 trillion comparisons.
- 4 million documents require about 8 trillion comparisons.

Given sufficient computational resources, a full pairwise comparison could in principle still be performed on the data considered in this study, effectively yielding a brute-force solution. However, our aim is to employ methods that remain computationally efficient as dataset sizes continue to grow. Therefore, we focus on probabilistic and hashing-based techniques that mitigate the scaling problem outlined above by avoiding explicit all-pairs comparison. These methods generate compact representations of documents and allow likely matches to be identified efficiently without exhaustively evaluating every pair (Navarro 2001).

4.1.2 From Textual Representation To MinHash

MinHash is a prominent example of a probabilistic approach to similarity clustering. It has a long history in large-scale information retrieval (Broder 1997) and remains a widely used component of modern similarity search pipelines (Khan et al. 2024). It generates compact signatures for each document by applying multiple hash functions to its shingles[8] and recording the minimum value per hash. The fraction of matching minima between two signatures approximates their Jaccard similarity.

Table 2 illustrates this principle. Each row represents a set of three 3-character shingles. Each column introduces a new hash function, which is applied to all shingles in each set before proceeding to the next. In the final column, we list the minima for every hash function, which form the document's signature. When comparing documents, we check how many values of their MinHash signatures match. For instance, both rows share 2 as a minimum value for the first hash function (representing *bbb*) and 1 for the second (representing *ccc*). These matches indicate that the corresponding documents are likely to contain overlapping shingles.

Intuitively, each individual hash function randomly *samples* the set, and may therefore miss shared elements. For example, Hash 3 does not capture the similarity between these two sets. In

7. The number of pairwise comparisons between n items is given by the binomial coefficient $\binom{n}{2}$, which counts all unique unordered pairs. This is equivalent to the formula $\frac{n(n-1)}{2}$. For example, with 10 000 documents, the number of comparisons would be $10000 \times 9999/2 \approx 50$ million.

8. A shingle is a contiguous sequence of tokens treated as a single element in a set for similarity comparison. We use the information-retrieval term "shingle" (common in MinHash/LSH) rather than "n-gram," which is more typical in language modeling.

this case, the probability that a hash function selects such a non-shared element (*aaa* or *ddd*) as the minimum is 2 out of 4, reflecting the fraction of elements outside the intersection. This is equal to the Jaccard similarity of the sets, which is 2/4=0.5. By applying more independent hash functions, these random misses balance out, and the fraction of matching minima across all hash functions converges toward the true Jaccard similarity. The proportion of hash functions that indicate a match becomes more stable as the number of hash functions increases. Most implementations allow the number of hash functions to be configured. Defaults such as 128–256 typically provide a good balance between accuracy and computational efficiency, although this can be adjusted depending on corpus size or desired precision.

2-grams	**Hash 1**	**Hash 2**	**Hash 3**	**MinHash Signature**
{aaa, bbb, ccc}	12, **2**, 7	8, 3, **1**	**4**, 9, 5	2 1 4
{bbb, ccc, ddd}	**2**, 7, 11	3, **1**, 6	9, 5, **2**	2 1 2

Table 2: MinHash signatures for two shingle sets. The fraction of matching minima approximates Jaccard similarity and becomes more accurate with additional hash functions.

4.1.3 LSH-Based Comparison

Once we have computed MinHash signatures for all documents in our dataset, a naive approach would be to compare every signature with every other signature to estimate pairwise similarities. This quickly becomes impractical for large corpora, because the number of comparisons would still grow quadratically with the number of documents. To address this, LSH is used as a prefiltering step. Each MinHash signature is split into several bands, and only documents that share at least one band are considered potential matches.

Table 3 illustrates this process. The sets *"aaa, bbb, ccc"* and *"bbb, ccc, ddd"* share identical values in two bands and are therefore retained as potential matches. By contrast, *"xxx, yyy, zzz"* does not share any band values with the other sets and is excluded from further comparison. Without LSH, we would still need to compare every element of this set against all elements of the other sets to confirm that no matches exist.

However, this signal is approximate: it provides no information about the specific overlap between shingles or the magnitude of similarity. Therefore, these smaller candidate subsets are subsequently examined in detail by analysing the original shingles. The actual number of pairs to analyse depends both on the choice of LSH parameters and on the structure of the corpus. Fewer comparisons are needed if we require documents to share a larger number of bands to be considered potential matches, while corpora with many similar documents will naturally produce more candidate pairs.

Text	**Band 1**	**Band 2**	**Band 3**
Set A: {aaa, bbb, ccc}	2	1	4
Set B: {bbb, ccc, ddd}	2	1	2
Set C: {xxx, yyy, zzz}	*5*	*6*	*7*

Table 3: LSH splits MinHash signatures into bands. Sets sharing bands form candidate pairs.

Finally, candidate pairs identified through LSH can be grouped into clusters using standard graph-based methods, where each text is treated as a node and edges are drawn between documents deemed similar. We then apply a union–find algorithm (Cormen et al. 2022), which efficiently merges nodes into the same cluster whenever they are connected by such edges, producing connected components that represent groups of similar texts.

4.2 Semantic Similarity: From Hierarchical Sentence Embeddings to Clustering

To cluster texts based on semantic similarity, we employ a hierarchical language model (HLM) as the encoding architecture, trained using a student–teacher framework. The resulting sentence embeddings are then clustered using an approximate nearest neighbor index, followed by grouping based on connected components in the resulting similarity graph. In this section, we review the theoretical foundations underlying both the modeling and the clustering methods.

4.2.1 Sentence-Level Hierarchical Language Modeling

The model used in this study, `kevinkrahn/shlm-grc-en`[9], is a Sentence-Level Hierarchical Language Model (SHLM) designed for Ancient Greek and English. While the Hugging Face model card references related publications (Krahn et al. 2023, Riemenschneider and Krahn 2024), the full training configuration is not reproduced, so the following summary is based on the available model card and configuration files.

The SHLM follows a **hierarchical language modeling (HLM)** approach (Sun et al. 2023), encoding text at multiple levels: characters, words, and sentences. In the first stage, an intra-word encoder processes each word as a sequence of up to 16 characters, generating character-informed word representations. These are then passed to an inter-word encoder, which contextualizes the words within the sentence and produces a final sentence embedding via CLS pooling. This hierarchical structure is particularly advantageous for morphologically rich or low-resource languages, as it captures both fine-grained linguistic patterns and overall compositional meaning, resulting in robust embeddings. The principal architectural parameters are summarized in Appendix A.

To enable cross-lingual embeddings, the model employs **multilingual knowledge distillation** (Reimers and Gurevych 2020, Krahn et al. 2023). During training, it learns to replicate the sentence representations of a high-performing English teacher model[10] using parallel Ancient Greek–English sentence pairs from the Perseus Digital Library[11] and First1KGreek[12]. This process aligns Greek and English texts in a shared semantic space, facilitating direct comparison and allowing the larger model to transfer semantic information to the student model.

4.2.2 Clustering

Once embeddings are obtained, we need an efficient way to search for similar items. To do this, we use an inverted-file (IVF) index, as implemented in the FAISS library (Douze et al. 2025)[13]. IVF first partitions the embedding space into coarse clusters using k-means clustering (MacQueen 1967). These clusters group nearby embeddings around centroids, providing an approximation of the underlying similarity structure. However, in practice, many genuinely similar items may lie near the boundary between clusters, meaning that strictly searching within a single cluster would miss relevant neighbors. To address this, IVF does not rely on a single cluster. Instead, for a given query, it identifies the nearest centroids and searches across several of the closest clusters. This allows the method to recover neighbors that fall just outside the primary cluster, while still avoiding a full search over all embeddings. As a result, IVF strikes a balance: clustering reduces the search space for efficiency, while probing multiple nearby clusters preserves accuracy by capturing cross-cluster similarities.

9. `https://huggingface.co/kevinkrahn/shlm-grc-en`
10. `https://huggingface.co/BAAI/bge-base-en-v1.5`
11. `https://github.com/PerseusDL/canonical-greekLit`
12. `https://github.com/OpenGreekAndLatin/First1KGreek`
13. FAISS (Facebook AI Similarity Search) is a library for efficient similarity search and clustering of dense vectors.

Once these candidate neighbors are identified, we construct a weighted graph where each text is a node and edges represent the cosine similarity between texts. To detect communities in this graph, we implement a union–find algorithm, which connects nodes into the same cluster whenever they are linked by edges exceeding a similarity threshold. This approach is analogous to the clustering step used with MinHash-LSH.

5. Experiment Setup

In the following section, we describe our approach to applying the surface-level and semantic similarity methods introduced in Section 4 to both the DBBE dataset and the full corpus described in Section 3. This allows us to compare clustering behaviour between a smaller, more structured dataset and a larger, more heterogeneous collection. We focus on the conceptual flow of the experiments, with detailed parameter settings provided in the corresponding appendices.

The full code used for these experiments is publicly available on GitHub[14], and an overview of the experiment pipeline can be found in Appendix B. All experiments were configured with a focus on scalability. Computational requirements are specified in Appendix C.

5.1 Database of Byzantine Book Epigrams

5.1.1 Surface-Level Clustering

For surface-level clustering, DBBE texts are **preprocessed** to reduce orthographic variation and mitigate lacunae and spelling errors that would otherwise fragment the clusters. At the **verse level**, we perform a grid search over shingle sizes and similarity thresholds to identify the configuration that best reconstructs the DBBE Verse Groups. Shingle size 1 was excluded because single-character shingles discard sequential structure and, given Greek's small alphabet, most verses have many letters in common, producing near-uniform Jaccard scores and little discriminative power. At the **poem level**, each epigram is represented as the set of verse-cluster identifiers assigned in the previous stage, and poem-to-poem similarity is computed via Jaccard overlap between these sets. Because expert annotations are available for the DBBE, both stages are evaluated against these groupings using precision, recall, and exact cluster recovery. However, the resulting optimum depends on the DBBE editorial groupings, which reflect a particular set of scholarly judgments. The reported metrics therefore indicate alignment with this perspective rather than absolute clustering quality. The full configuration is given in Appendix D.

5.1.2 Semantic Clustering

For semantic-level clustering, DBBE texts are again **preprocessed** to reduce orthographic variation. Additionally, NFKC normalisation is applied to align with the conventions used during model training[15], and only texts with at least six words (above the lower percentile) are retained to ensure sufficient contextual information for semantic modeling. At the **verse level**, embeddings are then computed using SHLM and structured into a graph based on cosine similarity. Verses are linked when they exceed a similarity threshold, and connected components are considered clusters. Multiple threshold values are evaluated to identify configurations that best align with the clustering structure in DBBE. At the **poem level**, poems are represented as sets of predicted verse clusters,

14. `https://github.com/PaulienLem/code-clin2026`
15. NFKC normalization not only brings decomposed sequences (e.g. α with breathing and accent marks) into their composed form (ἄ), but also reduces typographic or variant letterforms (such as the lunate sigma ϲ).

and Jaccard similarity is computed between these sets, as in the surface-level pipeline. Evaluation is done using precision, recall and exact cluster recovery using the DBBE ground truth. The full configuration is given in Appendix E.

5.2 Full Dataset

For the large-scale corpus, no ground truth annotations exist, and the texts are more varied in style, length, and formulaicity than the DBBE material. DBBE-derived parameter settings cannot therefore be reused directly. Instead, parameters are selected by evaluating cluster structure using a composite unsupervised criterion combining Davies–Bouldin index (Xiao et al. 2017)[16], silhouette scoring (Rousseeuw 1987)[17], and a non-singleton cluster ratio[18]. Combined, these metrics capture compactness, separation, and the avoidance of singleton clusters. Because the corpus is too large for exhaustive parameter search, grid searches are conducted on a stratified sample of 10 000 documents, and the best-performing configuration is then applied to the full dataset.

5.2.1 Surface-Level Clustering

For surface-level clustering on the full corpus, texts are **preprocessed** following the same pipeline as for the DBBE, with two additions: elaborated gap-marker stripping to handle editorial lacuna notation common in documentary sources, and a word-count filter to exclude lines that are too short or too long for reliable comparison. At the **line level**, we search over shingle sizes and Jaccard similarity thresholds on a stratified sample. We select the configuration that scores best on the composite criterion described above, restricting the search to shingle sizes greater than 1. The best-performing parameters are then used to cluster the full dataset at line level. At the **document level**, each document is represented as the set of line cluster identifiers from the previous stage, and Jaccard similarity over these sets is used to group documents together. A range of thresholds is explored, again selected according to the same composite criterion. The full configuration is given in Appendix F.

5.2.2 Semantic Clustering

For semantic clustering on the larger corpus, we apply a similar **preprocessing** pipeline as we did on the DBBE dataset. We then generate the embeddings and perform **line level** clustering, where thresholds are selected via a grid search on a stratified sample of lines from 10 000 documents. Since no ground truth labels are available, we evaluate configurations using the same composite criterion as in the surface-level approach. The best-performing parameters are then used to cluster the full dataset at line level. In a subsequent step, we aggregate at the **document level** by representing each document as the set of line-cluster identifiers it contains, and compute poem-to-poem similarity via exact containment over these sets. The full configuration is summarised in Appendix G.

16. The Davies–Bouldin index captures how clearly clusters are separated while remaining internally cohesive.
17. The silhouette score provides a per-item measure of how well an item fits its assigned cluster compared to neighbouring clusters.
18. We define the non-singleton cluster ratio as a simple regularisation term that measures the proportion of items assigned to clusters containing more than one element, discouraging solutions with excessive singleton clusters and favouring structurally informative groupings.

6. Results

6.1 Database of Byzantine Book Epigrams

As a first step, we focus on the DBBE dataset, whose structured nature provides a basis for validating our methods. Clustering is performed first at the surface level, followed by semantic level clustering, with the results assessed through both quantitative metrics and qualitative inspection.

6.1.1 Surface-Level Clustering

Figure 1: Verse-level grid search perfectly reconstructed 86.9% of DBBE Verse Groups (shingle size 2, Jaccard 0.6

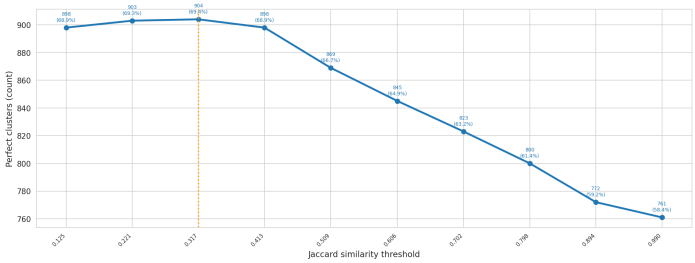

Figure 2: Epigram-level grid search perfectly reconstructed 69% of DBBE Types (Jaccard 0.317).

Quantitative As a first step, we applied MinHash-LSH to build surface-level clusters on the DBBE dataset. At the verse level, a shingle size of 2 and a Jaccard similarity of 0.6 perfectly reconstructed 86.9% of the DBBE Verse Groups (Figure 1). This configuration means that segmenting each verse into overlapping two-character sequences and requiring that 60% of these sequences match is sufficient to reconstruct most clusters. At the epigram level, a Jaccard threshold of 0.317 yielded 69% perfect reconstruction (Figure 2), suggesting that DBBE considers epigrams similar even when only a small subset of their verses are orthographically aligned.

Qualitative To gain a deeper understanding of the algorithm's behaviour, we examine a selection of representative cluster examples. Appendix H presents one illustrative case for each of the four observed outcome types: perfect reconstruction, splitting, merging, and mixed clusters.

When we look at the **verse-level** clusters, we notice that the algorithm produces **perfect** reconstructions in the majority of cases (86.9%). It groups verses in the same way as DBBE, despite differences in word order, accentuation, and morphological form. The most common divergence from the DBBE classification is **splitting**: in the example shown, the presence of entirely different words (πόθου δρόσον instead of βίβλον τήνδε) leads the algorithm to separate verses that DBBE groups based on their identical opening. **Merged** clusters (2.1%) arise when surface similarity overrides finer editorial distinctions: variants combining Χριστοῦ, Χρυσοστόμου, στόμα, and Παύλου are collapsed into a single cluster that DBBE distinguishes more finely. In the **mixed** cluster example, the algorithm does not treat ἀμήν as a separating factor between two ground truth groups, merging them together. However, it does exclude the verse δόξα σοι κ(ύρι)ε πάντων ἕνεκεν, whose unique

phrase σοι κ(ύρι)ε produces a sufficiently distinct shingle profile to prevent it from merging with the other four verses.

At the **poem level** (Appendix I), the approach reconstructs around 69% of DBBE Types. Even when some similarities are missed, partial overlap in predicted Verse Groups can still be sufficient for **perfect** clustering, as illustrated by the first example. In the **split** example, four epigrams that should cluster together share only one predicted Verse Group (4032 - λέων ὁ μάρκος εὑρεθεὶς ἐκ τῶν λόγων); other similar verses are not grouped due to missing text fragments, leaving the overlap too limited for correct clustering. In the **merged** example, three two-verse epigrams share one similar verse; a 50% overlap that is sufficient for the algorithm to combine them, despite DBBE distinguishing them based on differences in the second verse. The **mixed** example shows partial recovery of Type 3227 based on variants of ἡ βίβλο(ς) αὕτη τ(ῆς) μον(ῆς) γαλησίου. However, one near-identical epigram from Type 2492 was also included, despite having πέλη instead of μον(ῆς)). Conversely, epigram 19059, which DBBE also includes in Type 3227, is excluded because its incomplete verse provides insufficient overlap for detection.

6.1.2 Semantic Clustering

In the second experiment, we continue to use the DBBE dataset to benefit from its structured organization, while shifting the focus from orthographic similarity to semantic similarity. This allows us to evaluate whether semantically informed representations can recover meaningful groupings beyond surface-level variation, and to compare their behaviour against the previously explored orthography-based approach.

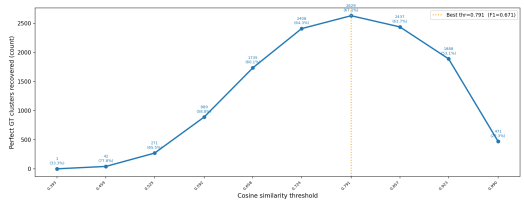

Figure 3: Verse-level grid search perfectly reconstructed 67.1% of DBBE Verse Groups (cosine similarity 0.791)

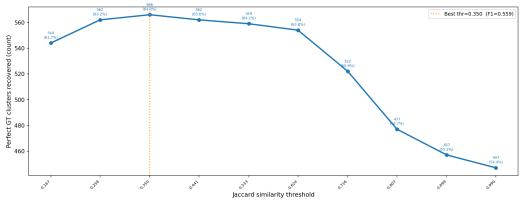

Figure 4: Epigram-level grid search reconstructed 64% of DBBE Types (Jaccard 0.35)

Quantitative As shown in Figure 3, a relatively high cosine similarity threshold of 0.791 yields the best results for verse-level clustering. At this setting, semantic clustering perfectly reconstructs 67% of the DBBE Verse Groups. At the poem level, a Jaccard threshold of 0.35 yields the best results, reconstructing 64% of the DBBE Type clusters (Figure 4). However, when measuring the overlap with the orthographic clusters, we find that 95% of the semantically clustered verses are also clustered orthographically. Similarly, 88% of the semantically clustered poems are already captured by the orthographic clustering. These results suggest two main findings.

Firstly, the semantic approach achieves lower coverage than the orthographic method. This is partly due to the nearest-neighbour retrieval step, which restricts verse-level candidate comparisons: 33% of missed verses show high cosine similarity but fall outside the set of retrieved nearest neighbours. The remaining 67% lie below the cosine similarity threshold. This latter group is likely explained by two interacting factors: the strict similarity cutoff and limitations of the embedding

space in capturing semantic similarity, potentially influenced by domain shift between the Classical Greek training data and the more orthographically variable Byzantine Greek texts.

Secondly, the strong overlap with orthographic clusters may either indicate that surface form is a dominant organising principle in the ground truth or reflect a residual sensitivity of the semantic model to surface-level similarity. However, the small number of clusters found exclusively through the semantic approach show higher orthographic variance at the verse level (Table 18, Appendix J) and at the poem level (Table 20, Appendix K), and can be seen as evidence that the model captures at least some degree of semantic similarity beyond shared surface forms.

Qualitative We examine the resulting clusters in greater detail to gain insight into how they are formed, distinguishing between perfect, split, merged, and mixed clusters. Full examples can be found in Table 19 (Appendix J).

At the **verse level**, the **perfect** cluster example confirms the strong alignment observed in the quantitative analysis between semantic and orthographic clustering. The verses in this cluster differ mainly in word order and minor lexical variation. In the **split** cluster example, a single DBBE Verse Group is only partially reconstructed: two variants are excluded because they omit γῆν or reverse the ordering of γῆν and θάλασσαν; differences apparently sufficient for the algorithm to withhold cluster membership. The **merged** cluster example unites verses from two ground truth groups sharing the theme of *rich gifts*. While this suggests broader semantic sensitivity, partial lexical overlap (e.g. πλουσίαις, δωρεαῖς) indicates a mixed signal rather than a purely semantic effect, though still beyond what surface-based methods such as MinHash-LSH would capture. The **mixed** cluster example (3392) combines verses containing near-synonymous theological adjectives (θείων vs. θεοσόφων) within the formula τὴν καλλονήν τε τῶν ... λόγων. At the same time, it excludes verses that complete the same opening with more specific attributions such as τοῦ Χριστοῦ μου or τοῦ προφήτου. This indicates a mixed signal in which strong formulaic and lexical similarity coexists with limited semantic variation.

At the **poem level**, the examples can be found in Table 21 (Appendix K). The **perfect** cluster example again follows directly from verse level behaviour, showing high orthographic similarity. In the **merged** cluster example, two poems are joined despite one verse differing substantially (ἀρετῆς πάσης ξένος vs. ἐν μοναστaῖς καὶ θύτης). Their other shared verse is similar enough to push the containment score above the threshold. The **split** cluster example shows a poem that is assigned to a separate cluster from an otherwise coherent group: despite expressing comparable thematic elements such as references to the hand, burial, and the persistence of writing, its verse-level structure and lexical choices diverge enough at the verse level to prevent alignment. In the **mixed** cluster example, the algorithm recovers a complete Type while also absorbing a single-verse poem whose sole verse closely resembles the opening lines of the clustered epigrams. The presence or absence of πέφηκε is not treated as sufficient grounds for separation, in contrast to the DBBE editorial classification.

6.2 Full Dataset

Having validated our approach on the DBBE dataset, we next turn to the full corpus. The same two-step clustering workflow is applied: first at the surface level, then at the semantic level. For each stage, optimal parameters are identified on representative subsets and subsequently applied to the complete dataset. Outcomes are evaluated both quantitatively, using cluster quality metrics, and qualitatively, by examining recurring patterns, textual parallels, and thematic groupings across the corpus.

6.2.1 Surface-Level Clustering

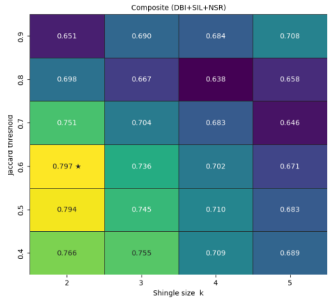

Figure 5: Line-level grid search: Optimal quality (0.797) obtained at shingle size 2, Jaccard 0.6

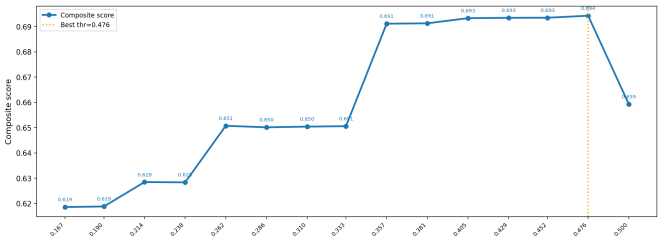

Figure 6: Document-level grid search: Optimal quality (0.694) obtained at Jaccard threshold 0.48.

Quantitative Analysis of a stratified sample indicated that optimal cluster quality was achieved at a similarity threshold of 0.6 combined with a shingle size of 2. In practical terms, this requires that two lines share roughly 60% of their two-character shingles to be considered similar. As a next step, we applied this threshold to cluster the full line-level dataset. At the document level, the same procedure was followed: a stratified sample was used for grid search, which identified an optimal similarity threshold of 0.48, meaning that documents sharing at least 48% of their line-level clusters were grouped together. This threshold was subsequently applied to cluster the full corpus.

Qualitative By analysing specific cases, we explore how the clustering captures patterns, and what insights emerge when moving from a curated corpus to a broad, heterogeneous collection.

Firstly, we notice several clusters that seem to point towards a broader formulaic tradition. For example, among the textual parallels between the DBBE and the corpus of Byzantine Inscriptional Epigrams is the epigram shown in Table 4. This epigram is discussed by Rhoby (2009), who notes its appearance in the dome of the church of Timios Stavros in Pelendri and draws attention to several other churches in which similar or nearly identical texts occur. The version from the Timios Stavros church is dated to the late fourteenth century. The version preserved in the DBBE[19] is dated to around 1100 and is based on a transcription found in *Duke University Library, Kenneth W. Clark Collection, Greek MS 25*, where it appears as a prefatory text to the Gospels of Luke and John.

Original	**Translation**
Ἐγὼ κριτής τε καὶ Θ(εὸ)ς πάντων πέλω·	I am both judge and God of all;
[ἰδού, προ]κύψας ὑψόθεν πρὸ τῆς δίκης	behold, leaning down from on high before the judgment,
παρεγγυῶμαι τοὺς ἐμοὺς τηρεῖν νόμους	I command to observe my laws,
ὅσοις θελητὸν ἐκφυγεῖν τὰς βασά[νους].	to those who wish to escape the torments.

Table 4: BIE - Epigramme auf Fresken 251

Another example of formulaic tradition are the verses in Table 5. The BIE text is preserved as an inscription on a Byzantine copper lamp dating to the tenth or eleventh century, as recorded in Rhoby (2010). In the verse, Christ is asked for the forgiveness of sins, a prayer formula that, as

19. `https://www.dbbe.ugent.be/occurrences/19024`

noted by Rhoby, occurs quite frequently, particularly at the end of epigrams inscribed on objects commemorating a donation. A closely related variant dated to the twelfth century also appears in DBBE[20]. The presence of this variant in DBBE illustrates Rhoby's observation that such prayer formulae occur frequently and confirms that this formula was not confined to a single artifact but circulated more broadly across different media and contexts.

ID	Original	Translation
BIE Me115	Σῶτερ, παράσχου λύσιν τῶν ὀφλημάτων	Savior, grant release from sins
DBBE 21585	Χριστέ, παράσχου λύσιν τῶν ὀφλημάτων	Christ, grant release from sins.

Table 5: Comparison of parallel epigrams BIE Me115 and DBBE 21585

A second type of similarities we see appearing in the results, seems to point to shared textual traditions. One example of this is the recurrence of the following epigram, inscribed on a statue commemorating Oppian of Anazarbus, author of the *Halieutica*:

Original	Translation
ὀππιανὸς κλέος ἔσχον ἀοίδιμον. ἀλλά με μοιρ(ῶν)	I bore the undying glory of Oppian, but the envious thread of
βάσκανο(ς) ἐξήρπαξε μῖτο(ς) κρυερὸς δ' αἴδης τε·	fate
κ(αὶ) νέον ὄντα κατέσχε τό(ν) εὐεπίης ὑποφήτην,	snatched me away, and cold Hades with it;
εἰ δὲ πολὺν χρόν(ον) ζω(ῆς) μίμν(ειν) φθόνο(ς) αἰνὸ(ς)	even in youth it claimed the herald of fine speech.
ἤθελ(εν), οὐκ ἄν μοι τίς ἶσον γέρ(ας) ἔλαχε φωτῶν	Had dreadful envy been willing to grant me a long span of life, no mortal would have obtained an equal share of honour.

Table 6: DBBE 19110

In DBBE, the epigram is linked to *Heidelberg, Universitätsbibliothek, Pal. gr. 40*[21] (14th century) and *Vatican City, Biblioteca Apostolica Vaticana, Urb. gr. 148*[22] (15th century), which contain the text of the *Vita Oppiani*. It also surfaces in the PHI database[23], where the source is given as Westermann (1845). However, Westermann's edition ultimately draws on the Heidelberg manuscript cited in DBBE. This demonstrates how similarity search not only exposes formulaic traditions but also allows us to trace the transmission of texts across catalogues and editions, enabling deeper comparative analysis of their content and metadata.

Thirdly, we frequently observe the repetition of standardized phrases that serve specific functional purposes. Examples include the use of religious phrases such as Ἰησοῦ Χριστοῦ τοῦ θεοῦ καὶ σωτῆρος (ex. bgu.3.725[24], DBBE 18281[25]), and dating formulae. The recurring nature of these phrases not only aids in the identification and linking of texts across different sources but also offers insights into administrative, religious, and social practices.

6.2.2 Semantic Clustering

Quantitative For the semantic pipeline, a grid search was conducted on a stratified sample of the corpus to determine appropriate similarity thresholds. The results of the line-level analysis, shown in Figure 7, indicate that a high cosine similarity threshold is required for the algorithm to provide meaningful clusters. This is expected given the diversity of the corpus. Setting a high threshold ensures that only line pairs with strong semantic alignment are considered, preventing weak or noisy

20. https://www.dbbe.ugent.be/occurrences/21585
21. https://www.dbbe.ugent.be/occurrences/24924
22. https://www.dbbe.ugent.be/occurrences/19110
23. https://epigraphy.packhum.org/text/286719
24. https://papyri.info/ddbdp/bgu;3;725
25. https://www.dbbe.ugent.be/occurrences/18281

associations from forming clusters and thereby enhancing cluster stability. The settings resulting from this sample set analysis were applied to cluster the full dataset on line level.

A grid search at the document level indicated that the highest quality score was achieved at a relatively low Jaccard similarity threshold, suggesting that even partial overlap in predicted verse clusters carries sufficient signal for grouping documents together. This may reflect the selectivity already imposed by the high cosine threshold at the verse level.

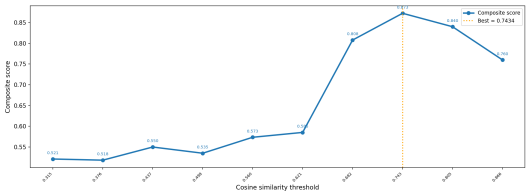

Figure 7: Line-level: Optimal cluster quality (87%) at cosine similarity 0.74

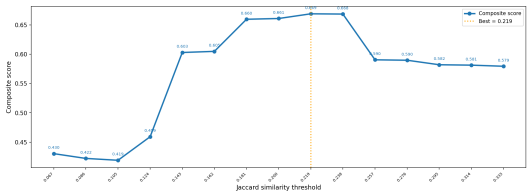

Figure 8: Document-level: Optimal cluster quality (67%) at Jaccard 0.219

Qualitative From a qualitative point of view, the line-level clusters capture relatively loose connections based on recurring motifs and shared vocabulary rather than tightly defined themes. For example, consider the following three clusters (see Appendix L for the full list):

- **Cluster 2311 — Measuring land:** This cluster gathers lines discussing measurements of land. While all lines revolve around the theme of assessing territory, the phrasing varies significantly. Some verses from DBBE emphasize abstract or poetic measurements, whereas papyri examples include more concrete surveying language.

- **Cluster 2452 — No fire:** Here, lines share the theme of avoiding or preventing fire. Internally, there is a mix of poetic hypothetical statements and more pragmatic or legalistic descriptions from papyri of fire prevention.

- **Cluster 2602 — Crying & mother:** This cluster groups lines centered on maternal grief and tears. The majority of lines share this theme clearly, ranging from DBBE's concise poetic expressions to PHI and papyri texts with narrative or legal context, illustrating the diversity of expression. A few lines, however, are only loosely connected, as some appear because they include a matronymicon rather than expressing grief directly, highlighting that not all cluster members are equally representative of the theme.

When these line-level clusters are aggregated at the document level, the thematic links become even more generic, which is a predictable consequence of moving from smaller textual units to larger ones. The documents in Table 7, for example, were clustered together likely because they share several distinctive semantic features. Both PHI TM920177 and BIE TR95 are different editions of the same source, but they do share with the DBBE epigram a distinctive specification of time, specifying years, cycles, and other chronological markers. Other than that, all documents convey a sense of ending or completion. The DBBE text marks the conclusion of a written work, while the PHI and BIE texts describe the culmination of a life. Additionally, all passages exhibit a formal and ceremonial tone, with structured phrasing and references to lineage or provenance. These shared thematic and stylistic patterns might explain why the semantic clusterer grouped these texts together.

Greek	Translation
DBBE 18750	
Τέλος ὧδε πέφθακεν, τῆσδε τῆς δέλτου.	The end of this tablet has come.
Μιχαὴλ δὲ ταῦτα σοι, προσφωνεῖ ἄνερ·	Michael addresses you with these words, O man.
Ὄστις τοὐπίκλην, ὁ μεσοποταμίτης·	Known by the surname, the Mesopotamian.
Κληρικοῦ δ' αὖθις, ἐπισκοπῆς εὑρίπου·	Of a cleric, again of episcopal office.
Συνάψας οὖν ἔγραψα πόνῳ καὶ μόχθῳ·	Having compiled, I wrote with toil and effort.
Εἰς θεοῦ τε αἴνεσιν καὶ ὑμνωδίαν·	For the praise and hymnody of God.
Πεπληρωτο δὲ ἐν χρόνοις τοῖς τοῦ κόσμου,	It was completed in the times of this world.
Χιλιὰς παρέδραμεν, ἐξαπλουμένη·	A millennium has passed, multiplied.
Καὶ ἑκατοντὰς ἐννακοσιοστῆ τε·	And the nine-hundredth century.
Σὺν τέσσαρες μόνοις τε, νῦν καὶ δεκάτοις·	With four and ten more now.
Ἰνδικτιόνης, δεκάτης καὶ τετάρτης.	Of the indiction, the tenth and fourteenth.
PHI TM920177	
Ὄναρ τὰ πάντα καὶ σκιᾶς οὐδὲν πλεόν·	All is a dream and nothing more than a shadow.
τὸ γὰρ φέρον τὰ πρῶτα τοῦ λαμπροῦ γένους	He who bore the foremost of the noble line.
καὶ βασιλικῆς ἐκ φυλῆς κατηγμένον	And sprung from royal lineage.
Κομνηνόβλαστον κλῆμα τῶν ἐκ πορφύρου	A Komnenian shoot from the purple-born.
ἐκ γῆς ἀνίσχον τῇδε Σικελοκράτους	Rising from the land of Sikelokrates.
κόνις καλύπτει καὶ θαλαμεύει σκότος·	Dust covers him and darkness encloses him.
καὶ τάφον οἰκεῖ τῆς κατάρας χωρίον	He inhabits the tomb, a place of curse.
πρίγκιψ Μανουὴλ τῶν χαρίτων ἡ βρῦσις,	Prince Manuel, fountain of graces.
τὸ τριακοστόπεμπτον ἀνύων ἔτος·	Completing the thirty-fifth year.
δεκὰς σὺν τῇ ἑπτάδι τοῦ Ἰουνίου	On the seventeenth of June.
φυτὸν μαραίνει τῆς ἀγαθοβρυσίας,	The plant of goodness withers.
κατὰ δισεπτάριθμον ἰνδίκτου κύκλον	Within the septennial indiction cycle.
καὶ κατὰ χιλιοστὸν ἑξαπλοῦν ἔτος	And in the sixfold thousandth year.
προσλαβὸν ἐκτὸς καὶ χρόνου περιόδους	Beyond the cycles of time.
ἑπτακοσίας ἐννεὰ πρὸς ταῖς δέκα.	Seven hundred and nine plus ten.
BIE Epigramme Auf Stein - TR95	
Ὄναρ τὰ πάντα καὶ σκιᾶς οὐδὲν πλέον·	All is a dream and nothing more than a shadow.
τὸ γὰρ φέρον τοῦ λαμπροῦ γένους	He who bears the noble lineage.
καὶ βασιλικῆς ἐκ φυλῆς κατηγμένον	And sprung from royal lineage.
Κομνηνόβλαστον κλῆμα τῶν ἐκ πορφύρου	A Komnenian shoot from the purple-born.
ἐκ γῆς ἀνίσχον τῇδε Σικελοκράτους	Rising from the land of Sikelokrates.
κόνις καλύπτει καὶ θαλαμεύει σκότος·	Dust covers him and darkness encloses him.
καὶ τάφον οἰκεῖ τῆς κατάρας χωρίον	He inhabits the tomb, a place of curse.
πρίγκιψ Μανουὴλ τῶν χαρίτων ἡ βρῦσις,	Prince Manuel, fountain of graces.
τὸ τριακοστόπεμπτον ἀνύων ἔτος·	Completing the thirty-fifth year.
δεκὰς σὺν τῇ ἑπτάδι τοῦ Ἰουνίου	On the seventeenth of June.
φυτὸν μαραίνει τῆς ἀγαθοβρυσίας,	The plant of goodness withers.
κατὰ δισεπτάριθμον ἰνδίκτου κύκλον	Within the indiction cycle.
καὶ κατὰ χιλιοστὸν ἑξαπλοῦν ἔτος	And in the sixfold thousandth year.
προσλαβὸν ἐκτὸς καὶ χρόνου περιόδους	Beyond temporal cycles.
ἑπτακοσίας ἐννεὰ πρὸς ταῖς δέκα.	Seven hundred and nine plus ten.

Table 7: Thematic similarity in DBBE 18750, PHI TM920177, BIE TR95

In the second example, displayed in Table 8, both documents share a spiritual and ethical focus, guiding the reader or listener toward moral or divine understanding. DBBE 18534 describes the soul's journey through wisdom and proper conduct, using imagery of light, fragrance, and guidance. PHI TM140971 similarly addresses the human soul navigating challenges, highlighting correct action and divine oversight. The use of natural metaphors and the recurring theme of moral or spiritual direction link the two texts semantically: both concern how a person should live in accordance with divine or ethical law, making them conceptually and thematically related.

Greek	Translation
DBBE 18534	
ἡ βίβλος ἥδε τῶν θεοπνεύστων λόγων:	This book of divinely inspired words
λειμὼν πέφυκε ψυχικῶν ἀρωμάτων:	has become a meadow of spiritual fragrances
ταύτην διελθὼν νουνεχῶς, ψυχὴν φίλε	Having passed through this wisely, dear soul
εὐωδιάσεις πν(εύματο)ς εὐωδία:	you will smell the fragrance of the Spirit
ἡ βίβλος αὕτη τῶν θ(εο)ῦ προσταγμάτων,	This book of God's commandments
λύχνος πέφυκε τοῖς ποσὶ καὶ φῶς τρίβοις:	has become lamp for the feet, light for paths
ταύτην διελθὼν νουνεχῶς, τοὺς σοὺς πόδ(ας)	Having passed through this wisely, your feet
κατευθυνεῖς ἄν(θρωπ)ε πρὸς θεί(ας) τρίβους:	you will direct, O man, toward divine paths
PHI TM512665	
ἀλλ' ὁπόταμψυχὴ προλίπῃ φάος ἀελίου,	But when the soul leaves the light of the sun
δεξιὸν εἴσιθι, ὡς δεῖ τινα πεφυλαγμένον	enter on the right, as one must, being guarded
εὖ μάλα πάντα· χαῖρε παθὼν τὸ πάθη-	fully in all things; rejoice in suffering
μα· τὸ δ' οὔπω πρόσθε ἐπεπόνθεις· θεὸς ἐγ-	but what you had not yet suffered before, God,
ένου ἐξ ἀνθρώπου, ἔριφος ἐς γάλα	born from a human, like a kid to milk
ἔπετες. χαῖρε, δεξιὰν ὁδοιπορ[ῶν]	you followed. Rejoice, walking the right path
λειμῶνάς τε ἱεροὺς κατ<ά τ'> ἄλσεα	and sacred meadows among groves
Φερσεφονείας.	of Persephone

Table 8: Comparison of DBBE 18534 and PHI TM512665 showing thematic similarity.

7. Discussion & Future Work

Experiments show that both surface-level and semantic clustering are not only scalable but also worthwhile: both produce strong quantitative results on the DBBE corpus and yield orthographic and thematic insights when applied to larger datasets. These findings raise broader questions about similarity detection in historical corpora.

When looking at the DBBE corpus, a striking observation is that both orthographic and semantic clustering surface predominantly orthographically similar material. This suggests that Byzantine epigram transmission operated primarily through close textual copying rather than through paraphrase, and that DBBE's editorial groupings implicitly reflect this orthographic logic. At the same time, the experiments reveal a smaller layer of semantic variation, where similar meaning is expressed through different vocabulary, suggesting that at least some degree of adaptive transmission is present in the corpus, and that this dimension may repay further philological investigation even if it remains subtle relative to the dominant orthographic signal. Interpreting this semantic layer is not straightforward, however. In practice, surface-form similarity and semantic similarity often co-occur, and the embedding model may encode both simultaneously, making it difficult to separate their respective contributions to clustering behaviour. This is further complicated by the domain shift between the Classical Greek data on which the model was trained and the more orthographically variable Byzantine material, as well as by general model limitations in capturing fine-grained semantic distinctions. Future work could address this more systematically, for example by using datasets in which semantic and orthographic similarity are more clearly decoupled, or by explicitly quantifying orthographic overlap within semantically derived clusters.

When the analysis scales to the larger, unlabeled corpus, the relative sparsity of clusters is to be expected given the heterogeneity of the material, but the clusters that do emerge are informative in qualitatively different ways depending on the method. Surface-level clustering via MinHash produces tight, high-precision groups that offer direct evidence of shared manuscript traditions and formulaic conventions. The cross-corpus parallels identified in our experiments illustrate how surface-level similarity search can trace lines of textual transmission across institutional boundaries and media.

The recurrence of standardised religious and administrative formulae across papyri and epigrams further demonstrates that certain fixed expressions functioned as broadly shared cultural currency, not confined to a single tradition or genre. Semantic clustering, by contrast, produces broader thematic groupings that operate at a different level of granularity and serve a different research purpose: the clusters identified in the full corpus group texts that are conceptually related without sharing wording, tracing topical and imagistic affinity rather than copying. These two levels of similarity are thus not competing but complementary, each addressing a distinct historical question.

At the same time, clustering inevitably involves methodological choices. While we opted for a combined quality score to evaluate the unlabeled dataset, different research aims, corpus characteristics, and embedding models could require different scoring mechanisms, thereby highlighting other facets of textual similarity. This does not contradict a data-driven stance; rather, it highlights that exploration and interpretation are intertwined and that selecting among plausible configurations without unintentionally steering the analysis remains a genuine methodological challenge. Against this backdrop, human-in-the-loop clustering offers a particularly promising direction for future work. Starting from the structures surfaced by the data, researchers can iteratively refine clustering in a human-in-the-loop setting, for instance by adjusting similarity thresholds, revisiting feature choices, or re-evaluating borderline cluster assignments, thereby aligning results more closely with interpretive questions while preserving mathematical consistency.

Beyond this, the study highlights the potential of working with large corpora to situate textual traditions within a broader context. Future research could build on this by creating larger-scale datasets, particularly for historical Greek where data remain relatively scattered. This would enable cross-corpus analyses that reveal broader patterns of linguistic, thematic, and formulaic variation.

In addition, future work could investigate hybrid methods combining surface-level and semantic similarity measures. For example, semantic similarity could be used to construct thematic clusters, within which surface-level similarity metrics are then applied to analyse orthographic variation among closely related expressions.

8. Conclusion

This study set out to explore the opportunities and challenges of applying large-scale, unsupervised similarity techniques to Greek corpora. Whereas most existing work remains tied to narrow research questions and small, domain-specific datasets, which limits both computational scalability and generalisability, our approach examined what becomes possible when assumptions are minimised and data-driven patterns are allowed to emerge. By developing two complementary pipelines, one based on MinHash-LSH for surface-level similarity and one based on transformer-based embeddings for semantic similarity, we demonstrated that unsupervised methods can reveal latent structural patterns without relying on predefined hypotheses.

A combination of targeted tests and quantitative analyses showed that these methods can detect meaningful relationships across texts at different levels of granularity. The surface-level pipeline efficiently identifies repeated forms and near-duplicate passages, while the semantic pipeline uncovers higher-level conceptual affinities.

Overall, this study provides both a scalable technical framework and a conceptual foundation for future work on interactive, exploratory analysis of large textual corpora. By integrating human judgment with computational detection of linguistic and semantic structure, forthcoming research has the potential to support new forms of philological inquiry and to illuminate broader patterns across expanding and increasingly interconnected historical Greek datasets.

Appendix A. SHLM Structure

Component	Parameter	Value
Intra-word encoder	Transformer layers	4
Intra-word encoder	Attention heads	12
Intra-word encoder	Hidden dimension	768
Intra-word encoder	Maximum characters per word	16
Intra-word encoder	Character vocabulary size	512
Inter-word encoder	Transformer layers	12
Inter-word encoder	Attention heads	12
Intra-word encoder	Hidden dimension	768
Inter-word encoder	Feedforward dimension	2048
Inter-word encoder	Maximum sequence length	256
Sentence embedding	Output dimension	768

Table 9: Architectural parameters of the SHLM model used in this study

Appendix B. Experiment Pipeline

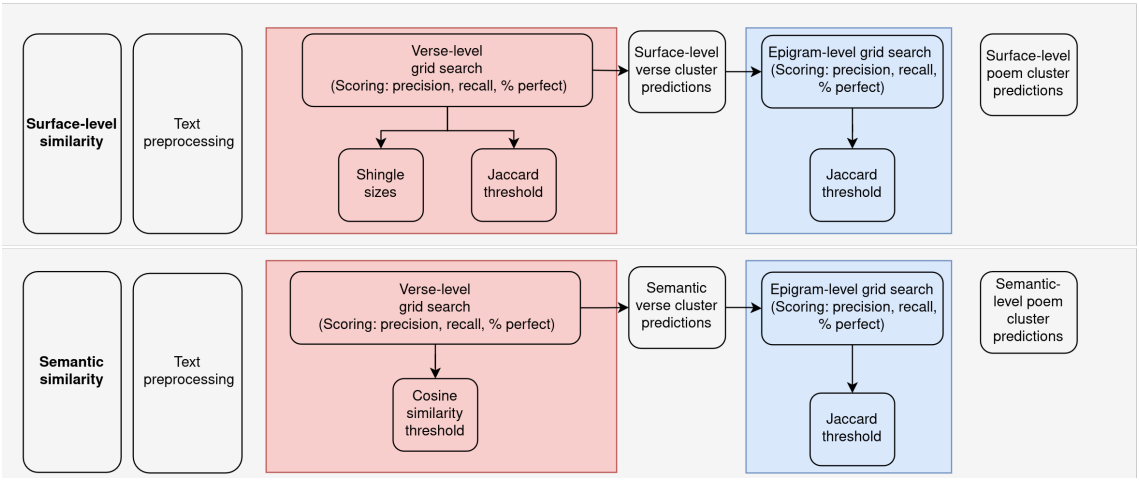

Figure 9: Schematic representation of text processing for the DBBE data

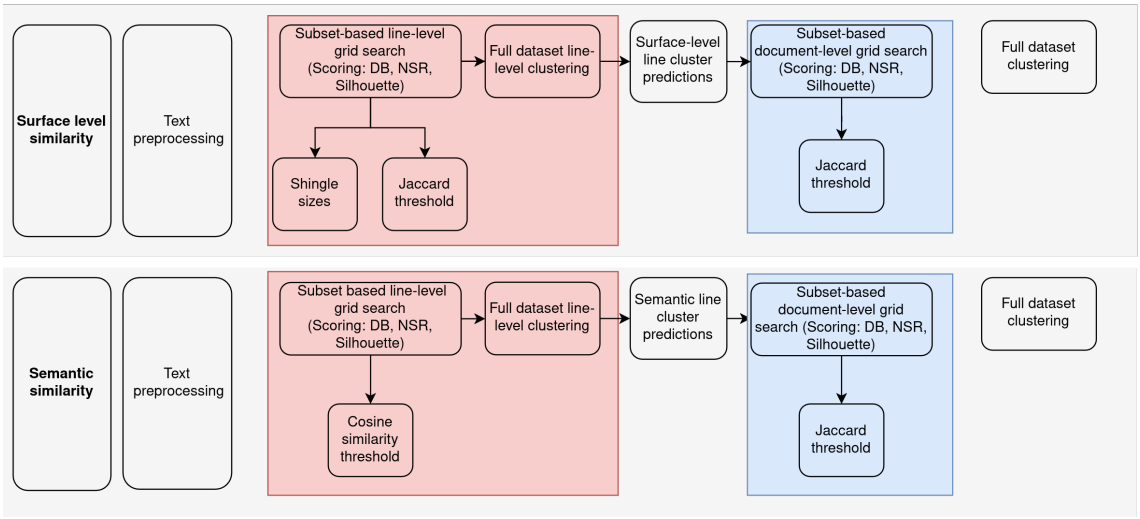

Figure 10: Schematic representation of text processing for the full dataset

Appendix C. Computational Requirements

Component	Specification
Python Version	3.11.3
CPU Cores	32 physical / 32 logical
System RAM	376 GB
GPU Model	NVIDIA Tesla V100-SXM2-32GB
GPU Memory	32 GB HBM2
Driver / CUDA	Driver 580.95.05, CUDA 13.0

Table 10: Summary of the HPC execution environment.

Procedure	Step	Time	Peak RAM	Peak GPU
Surface-Level Clustering	Line-level	6 min 42 s	1.99 GB	0 GB
	Document-level	6 min 48 s		
Semantic Clustering	Embedding generation	71 min	26.89 GB	12 GB
	Line-level	11 min		
	Document-level	24 s		

Table 11: Processing time and peak resource usage of full dataset experiments.

Appendix D. Methodology - DBBE Orthographic Clustering

Component	Configuration	Motivation
Preprocessing (applied to all stages)		
Lowercasing	Applied	Reduces orthographic variation due to case inconsistencies
Diacritics removal	Applied	Normalizes inconsistent or noisy accentuation
Punctuation removal	Applied	Normalizes inconsistent or noisy punctuation
Stage 1 — Verse-level clustering		
Shingle size	2–5 character grid search	Tests sensitivity of similarity detection across increasing character-context windows.
MinHash signatures	128 permutations	Standard trade-off between accuracy and efficiency
LSH bands	16 bands (8 rows each)	Efficient candidate generation while limiting false negatives
Similarity metric	Jaccard similarity	Measures overlap between LSH-prefiltered pairs
Clustering method	Union–Find	Cluster with verses (nodes) and Jaccard similarity percentages (edges)
Threshold selection	40–90% (grid search)	Explore optimal similarity percentage required to draw edges.
Evaluation	Precision, recall, perfect clusters	Measures alignment with DBBE Verse Groups
Stage 2 — Poem-level clustering		
Poem representation	Set of Stage 1 predicted Verse Group IDs	Lifts verse-level evidence to poem level
Similarity metric	Jaccard similarity	Exact similarity computation on predicted Verse Group sets
Threshold selection	10 Jaccard similarity thresholds between 1st and 99th percentile	Thresholds chosen from the observed range of poem-to-poem Jaccard similarities
Evaluation	Precision, recall, perfect clusters	Measures alignment with DBBE Type groupings

Table 12: Preprocessing and grid-search configuration for surface-level clustering.

Appendix E. Methodology - DBBE Semantic Clustering

Component	Configuration	Motivation
Preprocessing (applied to all stages)		
Lowercasing	Applied	SHLM is case-sensitive: we assume treating Μιχαήλ and μιχαήλ differently would not improve results.
Unicode	NFKC	Standard used during model training
Diacritics removal	Applied	Normalizes inconsistent or noisy accentuation
Punctuation removal	Applied	Normalizes inconsistent or noisy punctuation
Min verse length	6 tokens ($>$ lower quartile)	Minimal context for semantic interpretability
Stage 1 — Verse-level clustering		
Embedding model	SHLM	See Section 4
Indexing	IVF + nearest-neighbour retrieval (200 neighbours)	See Section 4
Clustering method	Union–Find	Cluster selected neighbours with verses as nodes and cosine similarities as edges.
Threshold selection	10 cosine values between the 50th and 99th percentile of similarities within the selected neighbours	Targets similarity levels where true verse relations emerge, excluding noise.
Evaluation	Precision, recall, F1, perfect cluster recovery	Measures alignment with DBBE Verse Groups
Stage 2 — Poem-level clustering		
Poem representation	Set of Stage 1 predicted verse-cluster IDs	Lifts verse-level evidence to poem level
Similarity metric	Jaccard similarity	Exact similarity computation on predicted Verse Group sets
Threshold selection	10 Jaccard similarity thresholds between 1st and 99th percentile	Thresholds chosen from the observed range of poem-to-poem similarities
Evaluation	Precision, recall, F1, perfect cluster recovery	Measures alignment with DBBE Type groupings

Table 13: Preprocessing and grid-search configuration for semantic clustering.

Appendix F. Methodology - Full Dataset Orthographic Clustering

Component	Configuration	Motivation
Preprocessing (applied to all stages)		
Lowercasing	Applied	Reduces orthographic variation due to case inconsistencies
Diacritics removal	Applied	Normalizes inconsistent or noisy accentuation
Punctuation removal	Applied	Normalizes inconsistent or noisy punctuation
Gap-marker removal	Applied	Strips editorial lacuna markers (`[ ]`, `< >`, `{ }`, `...`, `--`, etc.)
Word-count filter	6-50 words	Removes incomplete lines and outliers with insufficient or excessive content
Sampling		
Document eligibility	2-10 lines	Make sure we compare similar documents
Sample size	10 000 documents	Balances coverage with computational tractability
Stage 1 — Line-level clustering		
Shingle size	2-5 character grid search	Tests sensitivity of similarity detection across increasing character-context windows.
MinHash signatures	128 permutations	Standard trade-off between accuracy and efficiency
LSH bands	16 bands (8 rows each)	Efficient candidate generation while limiting false negatives
Similarity metric	Jaccard similarity	Measures overlap between LSH-prefiltered line-pairs
Clustering method	Union–Find	Clusters lines (nodes) connected by edges exceeding the Jaccard threshold
Threshold selection	40–90% (grid search)	Explore optimal similarity percentage required to draw edges.
Scoring criterion	Composite (DBI, silhouette proxy, NSR)	Unsupervised objective balancing cluster quality and graph sparsity
Stage 2 — Document-level clustering		
Document representation	Set of Stage 1 line-cluster IDs	Lifts line-level evidence to document level
Similarity metric	Jaccard similarity	Exact similarity computation on predicted line group sets
Threshold selection	15 Jaccard similarity thresholds between 5st and 95 percentile	Thresholds chosen from the observed range of document-to-document similarities
Scoring criterion	Composite (DBI, silhouette proxy, NSR)	Same unsupervised criterion as Stage 1

Table 14: Preprocessing and grid-search configuration for surface-level clustering on the unsupervised corpus.

Appendix G. Methodology - Full Dataset Semantic Clustering

Component	Configuration	Motivation
Preprocessing (applied to all stages)		
Lowercasing	Applied	Reduces orthographic variation due to case inconsistencies
Diacritics removal	Applied	Normalizes inconsistent or noisy accentuation
Punctuation removal	Applied	Normalizes inconsistent or noisy punctuation
Gap-marker removal	Applied	Strips editorial lacuna markers (`[ ]`, `< >`, `{ }`, `...`, `--`, etc.)
NFKC normalisation	Applied	Consistent with SHLM training convention; also normalises variant letterforms (e.g. lunate sigma)
Token-count filter	6–50 tokens	Removes lines too short for reliable semantic similarity and outliers dominated by noise
Sampling		
Document eligibility	2-10 lines	Ensures comparability across documents
Sample size	Up to 10 000 documents	Balances corpus coverage with computational tractability
Stage 1 — Line-level clustering		
Embedding model	SHLM	See Section 4
Indexing	IVF + nearest-neighbour retrieval (200 neighbours)	See Section 4
Similarity metric	Cosine similarity	Measures angular distance between L2-normalised embeddings
Clustering method	Union–Find	Clusters lines (nodes) connected by edges exceeding the cosine threshold
Threshold selection	10 thresholds, 50th–99th percentile of KNN similarity distribution	Sweep range derived from the observed similarity distribution; mirrors MinHash grid-search design
Scoring criterion	Composite (DBI, silhouette proxy, NSR)	Unsupervised objective balancing cluster quality and graph sparsity
Stage 2 — Document-level clustering		
Document representation	Set of Stage 1 line-cluster IDs	Lifts line-level evidence to document level
Similarity metric	Jaccard similarity	Exact similarity computation on predicted line-level clusters
Threshold selection	10 Jaccard similarity thresholds between 1st and 99 percentile	Thresholds chosen from the observed range of document-to-document similarities
Scoring criterion	Composite (DBI, silhouette proxy, NSR)	Same unsupervised criterion as Stage 1

Table 15: Preprocessing and grid-search configuration for semantic clustering on the unsupervised corpus.

Appendix H. DBBE Verse-Level Orthographic Clusters

Cluster Type	Ground truth	Verse	Epigram (Type ID)
PERFECT (86,9%)			
Perfect	19323	Κἂν τῆς [μερί]δος, τύχω τῆς τῶν ἐ[ρίφων]·	22737 (5516)
Perfect	19323	κἂν τῶν ἐρίφων τῆς μερίδος τυγχάνω;	35965 (35964)
SPLIT (10,3%)			
Split	16669	ὡς ἂν ἔχης ἥδυσμα κ(αὶ) βίβλον τήνδε·	34810 (34800)
Split	16669	ὡς ἂν ἔχης ἥδυσμα καὶ βίβλον τήνδε,	34807 (34800)
Split	16669	**Missed:** ὡς ἄν ἔχης ἥδυσμα καὶ πόθου δρόσον·	21587 (4918)
MERGED (2,1%)			
Merged	5452	στόμα δὲ παύλου τὸ χρυσοστόμου στόμα ·:·	17130 (4312)
Merged	18624	※ χ(ριστο)ῦ στόμα πέφυκε τοῦ παύλου στόμα·	17130 (4312)
Merged	18624	+ χριστοῦ στόμα πέφυκε τὸ παύλου στόμα	21413 (4828)
Merged	18624	Χριστοῦ δὲ στόμα τοῦ Παύλου πέλει στόμα,	22845 (5586)
Merged	18624	Χριστοῦ δὲ στόμα τοῦ Παύλου πέλει στόμα,	35527 (5586)
Merged	18623	Εἰ Παύλου στόμα τοῦ Χρυσοστόμου στόμα,	22845 (5586)
Merged	18623	Τὸ στόμα Παύλου στόμα τοῦ Χρυσοστόμου	25422 (5862)
Merged	18623	Εἰ Παύλου στόμα τοῦ Χρυσοστόμου στόμα,	35527 (5586)
Merged	18625	τοῦ Χρυσοστόμου Χριστοῦ και Παύλου	35527 (5586)
Merged	18625	τοῦ Χρυσοστόμου Χριστοῦ καὶ Παύλου στόμα.	22845 (5586)
MIXED (0,7%)			
Mixed	15377	**Missed:** δόξα σοι κ(ύρι)ε πάντων ἕνεκεν +	21267 (4692)
Mixed	15377	δόξα τῶ θ(ε)ῶ παντων | ἕνεκεν·	30654 (5642)
Mixed	15377	+ δοξα τώ θ(ε)ω ἡμ(ῶν) παντ(ων) ἕνεκεν	21974 (5642)
Mixed	4210	δόξα τῶ Θ(ε)ῶ πάντων ἕνεκα ἀμήν:-	23987 (6208)
Mixed	4210	+ δόξα τῶ θ(ε)ῶ πάντων ἕνεκα· ἀμή	25714 (5642)

Table 16: Each subsection shows one predicted verse cluster: it either perfectly matches the ground truth (Perfect), misses part of it (Split), misses some and merges other ground truth groups (Mixed), or fully merges multiple ground truth clusters (Merged). Type links provide high-level translations complementing the verses shown here.

Appendix I. DBBE Poem-Level Orthographic Clusters

Cluster Type	Pred Verse Group	Poem Verses	Epigram (Type ID)	Pred Cluster
PERFECT (69.4%)				
Perfect	6963 1402 -1	+ ἀσυμπαθὴς ἄν(θρωπ)ε κ(αὶ) φθόνου γέμων· τί γὰρ βλαβήσῃ πρὸς χάριν τὸν οἰκέτ(ην) σώζειν θ(εο)ῦ θέλοντος ἐξ εὐσπλα(...)χνίας:-	35328 (33768)	442
Perfect	6963 1402 3166	ἀσυμπαθὴς ἄν(θρωπ)ε | (καὶ) φθόν(...) γέμων· | (...) βλαβήση πρὸ(ς) | χάρ(ιν) τ(ὸν) οἰκέτ(ην) σώζ(ειν) | θ(εο)ῦ θέλοντος ἐξ εὐσπλαγ|χνίας	35329 (33768)	442
Perfect	6963 1402 3166	+ ἀσυμπαθὴς ἄν(θρωπ)ε, (καὶ) | φθόνου γέμων· | τί γὰρ βλαβήση πρὸς | χάριν τὸν οἰκέτην, | σώζειν θ(εο)ῦ θέλοντος | ἐξ εὐσπλαγχνί(ας):· +	35330 (33768)	442
SPLIT (0.6%)				
Split	4032 1151 4014 2146	(...)άρκος εὑρεθεὶς ἐκ τῶν λόγ(ων), (...)οὺς βρυχηθμ(οὺς) ἐξερεύγεται κτίσει· (...)τίσης πτύξασα δορκάδος δίκην, (...)νοὺς εκπέφευγε τῆς πλανης βρόχους·	23744 (5810)	484
Split	4032 1151 6774 2146	(...)άρκ(ος), εὑρεθ(εῖς) ἐκ τῶν λόγ(ων), φρικτοὺς βρυχηθμ(οὺς) | (...)εύγεται κτίσει· οὓς ἡ κτείσης πτύξασα δορκάδ(ος) | (...)ην, τοὺς δεινοὺς ἐκπέφευγε τῆς πλανης βρόχ(ους)·|	26247 (5810)	484
Split	4032 4048 6932 6478	+λέων ὁ μάρκος εὑρεθεὶς ἐκ τ(ῶν) λόγ(ων) φρικτ(ῶς) βρυχηθμ(ῶν) ἐξερε(...) κτύπ(ους) | οὓς ἡ κτίσ(ις) πτήξασα δορκάδ(...) (...)κην, τ(οὺς) δυσπλόκ(ους) πέφευγε τ(...)	25870 (5810)	894
Split	4032 4048 6932 1418	λέων ὁ μάρκος εὑρεθεὶς ἐκ τῶν, λόγων φρικτῶν βρυχηθμῶν ἐξερεύγεται κτύπους οὓς ἡ κτίσις πτήξασα δορκάδος δίκην τοὺς δυσπλόκους πέφευγε τῆς πλάνης βρόχους.	23223 (5810)	894
MERGED (18.1%)				
Merged	5618 1859	Ψαλτῆρος ἐξήγησις ἀκριβεστάτη ἑρμηνείας ἔχουσα πολλῶν πατέρων	18258 (2659)	912
Merged	3013 5618	θεοφυ|λάκτου ποιμένος βουλγαρί(ας), | ἥδ' (ἐστιν) ἐξήγησις ἀκριβεστάτη:	20998 (4506)	912
Merged	5618 1859	+ψαλτῆρος ἐξήγησις ἀκριβεστάτ(η)· ἑρμηνεί(ας) ἔχουσα (...) π(ατ)έρων:-	22513 (2659)	912
MIXED (11.9%)				
Mixed	3833 -1	ἡ βίβλο(ς) αὕτη τ(ῆς) μον(ῆς) γαλησίου· τ(ῆς) κειμ(ένης) ἔγγιστα τ(ῆς) ἐφεσίων +	24032 (3227)	239
Mixed	3833	βίβλος αὕτη μονῆς τοῦ Γαλησίου:	25391 (3227)	239
Mixed	3833	·:· βίβλ(ος) ἱερὰ, τῆς μονῆς Γαλησίου	17249 (2211)	239
Mixed	3833	ἡ βίβλος αὕτη πέλη τοῦ γαλησίου	17990 (2492)	239
Mixed	4839	ἡ βίβλος αυτη της μονῆς (...)	19059 (3227)	1132

Table 17: Example poem-level clusters. Perfect clusters reconstruct ground truth, split clusters miss part of a ground truth, merged clusters merge multiple ground truth groups into one cluster, mixed clusters combine some and split other ground truth clusters. Type IDs indicate ground truth groupings. Type links provide high-level translations complementing the poems shown here.

Appendix J. DBBE Verse-Level Semantic Clusters

Cluster Type	Ground truth	Verse	Epigram (Type ID)
Perfect	10886	στίθη τρέμων, | (...)ε, και (...)	24363 (6352)
Perfect	10886	στίθη τρέμων ἄν(θρωπ)ε καὶ νέβων κάτω	31129 (6352)
Perfect	10629	+ μανουὴλ πέφυκα πυκτὶς τοῦ βουλωτοῦ·	20866 (2909)
Perfect	10629	+ θεοδώρου πέφυκα. πυκτὶς | τοῦ ἀθηναίου.	22623 (2909)
Perfect	10583	ἀλλὰ τὸ κρίμα τοῦ τέμνειν ἐστὶ μέγα	20479 (4241)
Perfect	10583	ἀλλὰ τὸ κρίμα τομῆς ἔστι βία.	30844 (4241)
Perfect	18516	εὔχου μοι θῦτα γαβριὴλ τῶ ἀθλίω:	23838 (6141)
Perfect	18516	+ εὔχου μοι θῦτα τῶ τάλα μαλαχία· |	20336 (4186)
Perfect	14796	ἑπτα δὲ χιλιαδῶν ἔτι καὶ πέντε	20402 (4210)
Perfect	14796	ἑπτὰ χιλιάδων τε καὶ τριῶν ἔτι·	19858 (36300)
Perfect	13821	μερισμὸν ἐκ τέσσαρ(ας) ἀρχ(ας) λαμβάν(ειν). |	17794 (2340)
Perfect	13821	μερισμ(ὸν) εἰς τέσσαρ(ας) ἀρχ(ὰς) εἰσφέρων: |	32932 (2340)
Perfect	7561	[εἴληφ]εν τέρμα, βίβλος ἴδε [σὺν πό]θω·	22737 (5516)
Perfect	7561	ἤλιφε τέλ(ος) βίβλος ἴδε σὺν πόθω·	17613 (2189)

Table 18: Examples of verse pairs that were discovered exclusively based on the semantic similarity search. Type links provide translations complementing the verses shown here.

Cluster Type	Predicted ID	GT (ID, Size)	Verse	Epigram (Type ID)
PERFECT (67.1%)				
Perfect	661	18364 (5)	ἐξ ἧς κάτω πίπτουσιν ἄφρονες μόνοι	20157 (3401)
Perfect	661	18364 (5)	ἐξ ἧς κάτω | πίπτουσιν| ἄφρονες| μόνοι:-	21943 (3401)
Perfect	661	18364 (5)	κάτο πίπτωσι· ἄφρονες μόνοι:	22079 (3401)
Perfect	661	18364 (5)	ἄφρονες μόνοι κάτω πίπτουσιν, ἀμὴν.	35393 (35391)
Perfect	661	18364 (5)	ἐξ ἧς κάτω πίπτουσιν ἄφρονες μόνοι.+	19288 (3401)
SPLIT (20.3%)				
Split	1023	16163 (4)	γῆν γὰρ ἅπασαν καὶ θάλασσαν ἂν δράμῃς	18745 (3023)
Split	1023	16163 (4)	γῆν γὰρ ἅπασαν καὶ θάλασσαν ἂν δράμης |	25066 (3023)
Split	3275	16163 (4)	**Missed:** Τ(ὴν) γὰρ ἅπασαν (καὶ) θάλασσαν ἄν δράμης	20597 (3023)
Split	2523	16163 (4)	**Missed:** θά|λασσ(αν) καὶ γῆν τ(ὴν) ἅπασαν εἰ δράμοις	21487 (34498)
MERGED (2.6%)				
Merged	2538	8924 (1)	τοῖς πλουσί(οις) σου δώρ(οις) εὐημερί(αν)	18877 (3092)
Merged	2538	11260 (1)	ὃν ἀντ(ά)μειψαι δωρεαῖς σου πλουσίαις : -	16926 (3665)
MIXED (10%)				
Mixed	3392	5396 (2)	τὴν καλλονήν τε τῶν ὧδε θείων λόγων·	21025 (1955)
Mixed	3392	9904 (4)	Τὴν καλλονήν τε τῶν θεοσόφων λόγων:	24980 (5884)
Mixed	1784	5396 (2)	τὴν καλλονήν δὲ, τῶν λόγων τοῦ χριστοῦ μου·	18806 (1955)
Mixed	2998	9904 (4)	Τὴν καλλονήν δὲ τῶν τοῦ προφήτου λόγων:	17094 (1955)
Mixed	2998	9904 (4)	τὴν καλλονὴν δὲ τῶν τοῦ προφήτου λόγων·	20470 (1955)
Mixed	7726	9904 (4)	τ(ὴν) (...)λλονὴν (δὲ) τ(ῶν) τοῦ προφήτ(ου) λόγω(ν):	24540 (1955)

Table 19: Examples of predicted verse clusters: Perfect clusters match a single ground truth group; Split clusters fragment a true group; Merged clusters combine multiple groups; Mixed clusters exhibit both splitting and merging behavior. Type links provide translations complementing the verses shown here.

Appendix K. DBBE Poem-Level Semantic Clusters

Poem ID	Type ID	Excerpt (Shown verses / Total verses)
21499	3481	ίνδαρον ὑψαγόρην καδμῆϊδον οὔδεϊ θήβ(ης)· κλειδικη εὐνηθεῖσα μενεπτο(...) (2/10)
20222	3481	ίνδαρον ὑψαγόρην καδμῆϊδον οὔδεϊ θήβ(ης)· κλειδικη εὐνηθεῖσα μενεπτο(...) (2/10)
20262	4049	Ἐννέα τῶν πρώτων λυρικῶν πάτρην γενεήν τε ἐστὶ καὶ ἐκ Σπάρτης, Δωρίδος ἁρμονίωης (2/3)
20129	4049	ἐννέα τῶν πρώτων λυρικ(ῶν) πάτρ(ην) γενε(ήν) τε μάνθανε καὶ πατέρας καὶ διάλεκτον ἄθρει (2/12)
21092	2013	Ὅ(...) (...)ν λ(...) πη(...) γλυκὺν τῶν ἐγκυκλίων (2/20)
24541	2013	Ὅμηρος ὃν λέγουσι πηγὴν τῶν λόγων Στιχοπλόκον γλυκὺν τῶν ἐγκυκλίων (2/22)

Table 20: Example poem pairs that were discovered exclusively based on semantic similarity.

Cluster	Pred Verse Group	Poem ID	Verses	Type ID	Pred
PERFECT (64%)					
Perfect	12397 7779 19498 12398	23756	πυκτὶς, τίνος σύ· γνῶθι πραξαποστόλου· κτήτωρ δὲ, ὑάκινθος ἐν μονοτρόποις· γραφεὺς, ἰωνὰς, ἀλήτης ξένος· τὰ πάντα καινὰ. θαύματος παντὸς πέρα:	5879	192
Perfect	12397 13946 13947 12398	24987	Πυκτὶς, τίνος σύ· γνῶθι πραξαποστόλου: Κτήτωρ δὲ, νικόδημος ἐν μονοτρόποις: Γραφεὺς, ὁ αὐτός· ἔξοχος καλλιγράφοις: Τὰ πάντα καινά. θαύματος παντὸς πέρα:	5879	192
MERGED (15.4%)					
Merged	6835 6836	20383	Ματθαῖος οἰκτρὸς ἀρετῆς πάσης ξένος, πίνακα τοῦτον ὠργάνωσε κανόνων	4201	659
Merged	6835 6836	23835	Ματθαῖος οἰκτρὸς ἀρετῆς πάσης ξένος πίνακα τούτον ὠργάνωσε κανόνων.	4201	659
Merged	9696 6836	22028	Ματθαῖος οἰκτρὸς ἐν μοναστᾶις καὶ θύτης πίνακα τούτων ὠργάνωσε κανόνων.	5229	659
SPLIT (14.7%)					
Split	7165 7180 13613	21481	Ἔγραψε ταυτ(α) χεῖρ | μιχαῆλ (μον)αχ(οῦ)· (καὶ) χεῖρ | μὲν σίπετε τάφω· | γραφῆ (δὲ) ἑως τέλους | διαμένει ·:·	1974	-1
Split	265 310	17015	ἡ μὲν χεὶρ ἡ γράψασα, σήπετε τάφω· τὸ δὲ γράμμα μένη | εἰς χρόνους πληρεστάτους :-	1974	857
Split	265 14508	17080	ἡ χεὶρ μὲν ἡ γράψασα σύπετε τάφω, γραφεὶ δὲ | μένη· προς χρώνους πολλοῦς·	1974	857
… 22 additional poems of Type 1974 in predicted cluster 857 …					
MIXED (6%)					
Mixed	2614	18299	[Α]ὕτη ἡ βήβλος τὴς μονὴς Χορταΐτου	2679 (2)	471
Mixed	2614 6315 6316	12520	αὐτη ἡ βήβλος πέφηκε τῆς μονῆς χορταΐτου ἀρχομένης πρωτῆς τε, τῆς τεσσαρακοστῆς γε τοῦ χρισοστῶμου πανηγέ, τοῦ φιλοσοφωτάτου.	2732 (2)	471
Mixed	2614	18383	+ αὐτη ἡ βηβλος πέφικεν τῆς μονὶς χορταΐτου :-	2732 (2)	471

Table 21: Each subsection shows one poem cluster. Type links provide high-level translations.

Appendix L. Line-Level Semantic Clusters Full Dataset

Cluster 2311 (15 verses)

BIE - Epigramme auf Fresken 24	ἂν γῆς μετρήσῃς καὶ θαλάσσης τὰ βάθη, If you were to measure the depths of the land and the sea,
DBBE 21116	γῆν μετροῦντα πρόπασαν, ἀπειρεσίῃσιν ἐρωαῖς· ‖ Surrounding as a girdle all the earth with infinite flows ‖
DBBE 22570	Γῆν μετροῦντα πρόπασαν ἀπειρεσίῃσιν ἐρωαῖς. Surrounding as a girdle all the earth with infinite flows.
PHI TM 942442	ταὐτὸ μέτρον γαίης πρὸς The same measure of the land towards
PHI TM 885588	ἢ μέρος ἐξ αὐτῆς ἢ γῆς μέτρον ὢν ἀπὸ αὐτῆς Either a part of it or the measure of the land from it
PHI TM 47715	μετρῆσαι Ἀριστοδικίδηι καὶ παραδεῖξαι γῆς To measure for Aristodikides and show the land
Papyri TM 4998	γε[ν]ό̣μεν[ον ἐν τῆι γῆι μέτρωι χοῖ][-ca.?-][μετρήσει. δικαίαι] Having been in the land's measure, he will measure. Justly
Papyri TM 3183	καταμετρήσει τῆς γῆς πε[-ca.?-] He will measure the land's part…
Papyri TM 3686	τὰ ἐν τῆι ἑαυτοῦ γῆι μέρη τοῦ σημαινομένου The parts in his own land of what is being indicated
Papyri TM 2666	λον τὸ[ν] γενόμ[ενον ἐν τῆι γ]ῆι μέτρωι χόϊ τῶι Εὐπόλεως μετρή- …the one that occurred in the land, measured by the soil for Eupoles…
Papyri TM 22234	δὴ ὐ(*)πολογεῖται ἐκ τοῦ μέτρου τ[ῆς γ]ῆς τοῦ κατὰ πε- …thus it is calculated from the measure of the land according to…
Papyri TM 22467	θωμε(*) τὴν γῆν καὶ μετρήσω εἰς τὸ δη[μό-] …we see the land and I will measure it for the public…
Papyri TM 5313	μετρῆσαι ἑκάστωι οὗ ἡ γῆ ἐστι [διὰ τῶν] To measure for each where the land is [through the…]
Papyri TM 1832	νίου γῆς μεμετρῆσθαι ἀπὸ …of the young land to be measured from…
Papyri TM 1869	γῆν. ἐγεωμέτρησα οὖν αὐ- The land. Therefore, I measured it…

Cluster 2452 (11 verses)

BIE Epigramme auf Fresken 159	Βάτον καιομένη(ν) κ(αὶ) μὴ φλεγομένην A bramble burning and yet not aflame
DBBE 34882	ὡς ἄν γε μὴ τὴν ἄλωνα τὸ πῦρ προσαναλώσει: So that the fire does not consume the threshing floor:
DBBE 19293	ὡς ἄν γε μὴ τὴν ἄλωνα τὸ πῦρ προσαναλώσει: So that the fire does not consume the threshing floor:
DBBE 26356	Ὡς ἄν γε μὴ τὴν ἄλωνα τὸ πῦρ προσαναλώσει: So that the fire does not consume the threshing floor:
PHI TM 918925	αὐτῶν τὸ μὴ ἀποκαῦσαι, γενέσθαι τὸ That it does not burn, it happens that…
PHI TM 775339	[ματι τῶι] ζεφυρίωι(?) μὴ κάειν πῦρ μηδ[ὲ π]ρὸς τ– – – – – […at the west wind] not to light fire nor towards …
PHI TM 814037	ἀναθεῖναι μηθέν, μηδὲ σκανοῦν μηδὲ π[ῦρ ἀνάπτειν ἐντὸς ἢ ἐκτὸς] Do not place anything, nor sprinkle, nor light fire inside or outside

PHI TM 129604	τοῦτον παρὰ τόπον μὴ ὑπρισ- …this in the place, do not ignite…
PHI TM 495238	τῷν ̣τ[οι]ο̣ύ̣τ̣ω̣ν [ἐ]ν [ταῖς] π[υροκαύσε]σ̣ι̣ν̣ εἰδότες μήτε ὁ ἐπὶ …of such things in the fire, knowing neither he who is over…
Papyri TM 10229	[μ]ξενος πῦρ προσβαλεῖν καὶ μὴ ἰσ̣χύ- …to attack with fire and not have power…
Papyri TM 11630	πάσηι βεβαιώσει ἐπὶ [τὸν ἄ]παντα χρόνον κ(αὶ) [μ]ὴ ὑπερυποκειμένην He will secure it for all time, and not let it be superseded or alienated by anyone in authority…

Cluster 2602 (9 verses)

BIE Epigramme auf Fresken 193	μ(ήτ)ηρ δὲ θρηνεῖ καὶ μαθητὴς δακρύων And the mother laments, and the student weeps
DBBE 22746	θρηνεῖ τεκοῦσα καὶ μαθητὴς δακρύοις. The mother laments and the student with tears.
PHI TM 782805	[μ]ατρί τε παμπληθὺν θρηνῶν γόον· ἐγ δὲ λοχείας And to the mother, full of lamentation, a cry; and I of childbirth…
PHI TM 887398	[ΑΣ]ΚΥΡΙΕΩΣ καὶ ἐθήκατο δάκρυα μητρὶ …and he placed tears to the mother…
PHI TM 763150	ᾳὐτά τοι φ[θ]ιμένωι μάτηρ ἐπὶ δάκρυ χέασ[α] …to the deceased, the mother poured out tears…
PHI TM 773280	οἱ δ’ αὖ ματρὶ πόθους σὺ[ν] δάκρυσι λ[εῖβον ἔνηβοι], And they again poured desires to the mother with tears, young men…
PHI TM 127802	μητέρι καὶ γενέτῃ στοναχὰς καὶ δάκρυ λιπόντι. | To the mother and father, moans and tears being shed.
Papyri TM 10447	Μαρρῆς Ἥρω̣ν̣ος τ̣οῦ Πεθέως μητ(ρὸς) Τεφορᾶ(τος) ἱερε(ύς) Marrēs, son of Hērōn of Petheus, priest of Teforatos
Papyri TM 20886	ἔθη τέκνων δι[καίῳ, Αὐρ]ηλίῳ Μ̣ε̣ν̣ω̣νι Θέωνος μ[η]τρὸς Κλαυ- ..in accordance with Roman law concerning children, (declares this) to Aurelius Menon, son of Theon, whose mother was Claudia

References

Baumann, Ryan (2013), The son of suda on-line, *The Digital Classicist* pp. 91–106, JSTOR.

Bentein, Klaas (2025), Socio-semiotic, multimodal annotation of documentary sources : digital infrastructure in the everyday writing project, *in* Reggiani, Nicola, editor, *Digital papyrology III*, De Gruyter, pp. 221–256. http://doi.org/10.1515/9783111070162-014.

Berger, Maria, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler (2016), Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1849–1859.

Broder, Andrei Z (1997), On the resemblance and containment of documents, *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, IEEE, pp. 21–29.

Cormen, Thomas H, Charles E Leiserson, Ronald L Rivest, and Clifford Stein (2022), *Introduction to algorithms*, MIT press.

D'Angelo, Caterina, Andrea Taddei, and Alessandro Lenci (2025), Detecting semantic reuse in Ancient Greek literature: A computational approach, *in* Bosco, Cristina, Elisabetta Jezek, Marco Polignano, and Manuela Sanguinetti, editors, *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, CEUR Workshop Proceedings, Cagliari, Italy, pp. 327–336. https://aclanthology.org/2025.clicit-1.34/.

Deforche, Maxime, Ilse De Vos, Antoon Bronselaer, and Guy De Tré (2024), A hierarchical orthographic similarity measure for interconnected texts represented by graphs, *Applied Sciences* **14** (4), pp. 30, MDPI. https://doi.org/10.3390/app14041529.

Delouis, Olivier (2012), Andreas Rhoby, Byzantinische Epigramme in Inschriftlicher Überlieferung. I, Byzantinische Epigramme auf Fresken und Mosaiken (veröffentlichungen zur byzanzforschung 15 - philosophisch-historische klasse. denkschriften 374), 2009. Andreas Rhoby, Byzantinische Epigramme in Inschriftlicher Überlieferung. II, Byzantinische Epigramme auf Ikonen und Objekten der Kleinkunst. Nebst Addenda zu Band I "Byzantinische Epigramme auf Fresken und Mosaiken" (Veröffentlichungen zur Byzanzforschung 23 - Philosophisch-historische Klasse. Denkschriften 408), 2010, *Revue des études byzantines* **70** (1), pp. 316–317. https://www.persee.fr/doc/rebyz_07665598_2012_num_70_1_4982_t9_0316_0000_1.

Demoen, Kristoffel, Gilbert Bentein, Klaas Bentein, Floris Bernard, Julián Bértola, Julie Boeten, Mathijs Clement, Cristina Cocola, Eline Daveloose, Sien De Groot, Pieterjan De Potter, Ilse De Vos, Krystina Kubina, Hanne Lauwers, Paulien Lemay, Renaat Meesters, Marjolein Morbé, Delphine Nachtergaele, Marthe Nemegeer, Joachim Nielandt, Mace Ojala, Lisa-Lou Péchillon, Raf Praet, Rachele Ricceri, Anne-Sophie Rouckhout, Jeroen Schepens, Febe Schollaert, Lev Shadrin, Nina Sietis, Dimitrios Skrekas, Colin Swaelens, Maria Tomadaki, Sarah-Helena Van den Brande, Merel Van Nieuwerburgh, Lotte Van Olmen, Noor Vanhoe, and Nina Vanhoutte (2023), Database of byzantine book epigrams. https://doi.org/10.5281/zenodo.7682523.

Depauw, Mark and Tom Gheldof (2013), Trismegistos: An interdisciplinary platform for ancient world texts and related information, *International Conference on Theory and Practice of Digital Libraries*, Springer, pp. 40–52.

Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou (2025), The faiss library, *IEEE Transactions on Big Data*, IEEE.

Giannikou, Kyriaki, Colin Swaelens, Ilse De Vos, Els Lefever, and Klaas Bentein (2024), Decoding byzantine book epigrams: an exploration of machine-assisted extraction of formulaic material, *Workshop on Data-driven Approaches to Ancient Languages*, Language & Translation Technology Team, pp. 22–32.

Jaccard, Paul (1901), Etude comparative de la distribution florale dans une portion des alpes et des jura, *Bull Soc Vaudoise Sci Nat* **37**, pp. 547–579, UNIL.

Khan, Arham, Robert Underwood, Carlo Siebenschuh, Yadu Babuji, Aswathy Ajith, Kyle Hippe, Ozan Gokdemir, Alexander Brace, Kyle Chard, and Ian Foster (2024), Lshbloom: Memory-efficient, extreme-scale document deduplication, *arXiv preprint arXiv:2411.04257*.

Krahn, Kevin, Derrick Tate, and Andrew C. Lamicela (2023), Sentence embedding models for Ancient Greek using multilingual knowledge distillation, *in* Anderson, Adam, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti, editors, *Proceedings of the Ancient Language Processing Workshop*, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, pp. 13–22. https://aclanthology.org/2023.alp-1.2/.

Lemay, Paulien, Els Lefever, and Klaas Bentein (2026), Corpusclues: Scalable unsupervised similarity search for historical texts using minhash-lsh, *Proceedings of the Fifteenth Language Resources and Evaluation Conference*. Under review.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019), Roberta: A robustly optimized bert pretraining approach. https://arxiv.org/abs/1907.11692.

MacQueen, James (1967), Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, Oakland, CA, USA, pp. 281–297.

Navarro, Gonzalo (2001), A guided tour to approximate string matching, *ACM computing surveys (CSUR)* **33** (1), pp. 31–88, ACM New York, NY, USA.

Ochab, Jeremi K and Holger Essler (2019), Stylometry of literary papyri, *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pp. 139–142.

Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray (2019), GASC: Genre-aware semantic change for Ancient Greek, *in* Tahmasebi, Nina, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, Association for Computational Linguistics, Florence, Italy, pp. 56–66. https://aclanthology.org/W19-4707/.

Reimers, Nils and Iryna Gurevych (2020), Making monolingual sentence embeddings multilingual using knowledge distillation, *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 4512–4525.

Rhoby, Andreas (2010), *Byzantinische Epigramme auf Ikonen und Objekten der Kleinkunst*, Vol. 23, Verlag der Österreichischen Akademie der Wissenschaften.

Rhoby, Andreas (2014), *Byzantinische Epigramme in inschriftlicher Überlieferung. Band 3, Teil I: Byzantinische Epigramme auf Stein nebst Addenda zu den Bänden 1 und 2*, Verlag der Österreichischen Akademie der Wissenschaften.

Rhoby, Andreas (2018), *Object*, Verlag der österreichischen Akademie der Wissenschaften.

Rhoby, Andreas, Wolfram Hörandner, Anneliese Paul, et al. (2009), *Byzantinische Epigramme auf Fresken und Mosaiken*, Vol. 1, Verlag der Österreichischen Akademie der Wissenschaften Vienna.

Ricceri, Rachele, Klaas Bentein, Floris Bernard, Antoon Bronselaer, Els De Paermentier, Pieterjan De Potter, Guy De Tré, Ilse De Vos, Maxime Deforche, Kristoffel Demoen, Els Lefever, Anne-Sophie Rouckhout, and Colin Swaelens (2023), The database of byzantine book epigrams project: principles, challenges, opportunities, *Journal of Data Mining and Digital Humanities* p. 41. http://doi.org/10.46298/jdmdh.10244.

Riemenschneider, Frederick and Anette Frank (2023), Graecia capta ferum victorem cepit. detecting latin allusions to ancient greek literature, *Proceedings of the Ancient Language Processing Workshop*, pp. 30–38.

Riemenschneider, Frederick and Kevin Krahn (2024), Heidelberg-boston @ SIGTYP 2024 shared task: Enhancing low-resource language analysis with character-aware hierarchical transformers, *in* Hahn, Michael, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Yulia Otmakhova, Jinrui Yang, Oleg Serikov, Priya Rani, Edoardo M. Ponti, Saliha Muradoğlu, Rena Gao, Ryan Cotterell, and Ekaterina Vylomova, editors, *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, Association for Computational Linguistics, St. Julian's, Malta, pp. 131–141. https://aclanthology.org/2024.sigtyp-1.16/.

Rousseeuw, Peter J (1987), Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics* **20**, pp. 53–65, Elsevier.

Sommerschield, Thea, Yannis Assael, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas (2021), I.PHI dataset: ancient greek inscriptions.

Sommerschield, Thea, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas (2023), Machine learning for ancient languages: A survey, *Computational Linguistics* **49** (3), pp. 703–747, MIT Press, Cambridge, MA. https://aclanthology.org/2023.cl-3.5/.

Stopponi, Silvia, Saskia Peels-Matthey, and Malvina Nissim (2024), Agree: a new benchmark for the evaluation of distributional semantic models of ancient greek, *Digital Scholarship in the Humanities* **39** (1), pp. 373–392, Oxford University Press.

Storey, Grant and David Mimno (2020), Like two pis in a pod: Author similarity across time in the ancient greek corpus, *Journal of Cultural Analytics*, Center for Digital Humanities, Princeton University.

Sun, Li, Florian Luisier, Kayhan Batmanghelich, Dinei Florencio, and Cha Zhang (2023), From characters to words: Hierarchical pre-trained language model for open-vocabulary language

understanding, *in* Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, pp. 3605–3620. https://aclanthology.org/2023.acl-long.200/.

Swaelens, Colin (2025), *Verse by Verse: Modelling Semantic Similarity in Byzantine Greek Poetry*, PhD thesis, Ghent University.

Westermann, Anton (1845), *Biographoi: vitarum scriptores Graeci minores*, sumptum fecit G. Westermann.

Xiao, Junwei, Jianfeng Lu, and Xiangyu Li (2017), Davies bouldin index based hierarchical initialization k-means, *Intelligent Data Analysis* **21** (6), pp. 1327–1338, SAGE Publications Sage UK: London, England.