How relevant is part-of-speech information to compute similarity between Greek verses in a graph database?

Colin Swaelens†, Maxime Deforche‡, Guy Detré‡, Ilse De Vos*, Els Lefever†
†Language and Translation Technology Team, Ghent University, Belgium
‡ Dpt. of Telecommunications and Information Processing, Ghent University, Belgium
*VAIA - Flanders AI Academy, Heverlee, Belgium
{colin.swaelens,maxime.deforche,guy.detre,els.lefever}@ugent.be, ilse.devos@kuleuven.be

Abstract

This paper presents the automatic linguistic analysis of the Database of Byzantine Book Epigrams (DBBE) on the one hand, and its representation and integration in a graph database on the other hand. Firstly, we provide a comprehensive description of the DBBE data we want to provide with a complete morphological analysis. The presented methodology explores the possibilities of fine-tuning the DBBErt transformer-based language model, which was trained on pre-Modern and Modern Greek. Secondly, the automatically annotated epigrams are integrated in a graph database, a new way to represent the relatedness of this entangled corpus. With the graph database, we can compute similarity between words, verses and epigrams. Given the scope of this paper, we computed a complete orthographic similarity between the verses, a similarity based on the automatically assigned part-of-speech information and a final similarity measure that combines both orthography and part-of-speech information. The results of these similarity measures provide scholars with new visual representations of relations between (parts of) texts, which is beneficial for new critical editions and commentaries.

Keywords: Ancient Language Processing, Graph Database, Similarity Search

1. Introduction

The traditional way of making a historical text accessible to the general public typically involves the production of a critical edition. Through a critical edition, an editor, viz. a philologist, presents their interpretation of what the original text (*Uhrtext*) likely was, drawing from the manuscripts that have survived over time. Beneath the main text of a critical edition, the apparatus displays all variants of a given word within the text, as found in the manuscripts. The greater the number of manuscripts included, the more *critical* the edition becomes. The question arises as to whether research would benefit from more dynamic systems in contrast to to the static system inherent to a critical edition. A more dynamic system could, for instance, store linguistic information for each word within the text, offering better insight into the variation of textual readings. In this paper, we propose such a dynamic system built upon a graph database framework, which facilitates the grouping of similar words, verses and even complete chunks of text. The similarity measure may rely solely on orthographic criteria, or it can take into account variation in spelling as well as flexible word order. Incorporating linguistic information, such as part-of-speech tags, morphological features, or semantic labels, enables the utilisation of the most fine-grained queries to identify related textual segments. Such a tool empowers philologists to make more robust and comprehensive critical editions as well as commentaries.

In this paper, we introduce the first version of this system, which incorporates various orthography-based similarity measures alongside automatically tagged part-of-speech information. The corpus we work with comprises Byzantine book epigrams, which are poems typically inscribed in the margins of manuscripts by the scribe of the manuscripts themselves. Editions of these book epigrams do exist (Rhoby, 2018), but the Database of Byzantine Book Epigrams (DBBE) (Ricceri et al., 2023) has the unique benefit of storing both the verbatim transcription of the epigrams as well as their edition-like variants, called *Occurrence* and *Type* respectively. As the DBBE *Occurrences* present the epigrams exactly as they appear in the manuscripts, they exhibit quite some inconsistencies, including variations in orthography, punctuation, and metre. This stems from the epigrams being predominantly autographs, a sharp contrast to classical texts that have been copied and edited over centuries.

2. Literature Review

The visual grouping tool presented in this paper integrates the focal points of this literature review: orthographic similarity and linguistic annotation of Greek.

2.1. Orthographic Similarity

Orthographic similarity measures seek to calculate a similarity score between two texts, purely based on the likeness between individual tokens or characters comprising the text, without considering contextual information or semantics. Character-based orthographic measures such as N-grams (Kondrak, 2005), Jaro(-Winkler) (Winkler, 1990; Jaro, 1995), and Damerau(-Levenshtein) (Damerau, 1964; Levenshtein, 1966) compute string similarity by comparing sequences of individual characters. Likewise, token-based similarity measures, like the Overlap Coefficient, the Cosine Similarity or the Jaccard Similarity (Jaccard, 1901; Gomaa et al., 2013) produce a similarity score by comparing between sets or sequences of complete tokens. Few techniques that combine both token- and character-based methods, have been investigated. These hybrid techniques ascertain the similarity between two tokens by considering the underlying character-based similarity score of those two tokens (Bronselaer and De Tré, 2009; Gali et al., 2019). Traditional orthographic methods typically aim to compute a single, comprehensive similarity score, without taking into account the underlying structural intricacies of the texts. However, when assessing the similarity among (the components of) Byzantine book epigrams, which constitute highly interconnected semi-structured texts, these methods prove inadequate.

Deforche et al. (2024) have proposed a new, innovative orthographic similarity measure: it supersedes the notion of simply merging character- and token-based measures and instead deals with texts in a more structured manner. This novel method breaks down texts into hierarchical discourse units, like words or verses, and, commencing from the smallest units, proceeds to compute similarities between all elements belonging to the same discourse unit. These hierarchical similarity calculations draw inspiration from the Damerau-Levenshtein distance (Damerau, 1964), and the computations for a specific discourse unit will integrate the precomputed similarity scores between the lower units of discourse. Furthermore, the hierarchical breakdown of texts, coupled with the similarity scores between the elements of each discourse unit, can be stored in a graph database (Angles and Gutierrez, 2008). By leveraging the advanced and/or visual querying capabilities of these databases, new methods and tools for exploring and analysing textual corpora can be devised. This hierarchical method has yielded promising results in computing orthographic similarities among (segments of) Byzantine book epigrams, where each epigram is represented by a hierarchical decomposition of tokens, verses, and complete texts (Deforche et al., 2023, 2024).

2.2. Part-of-Speech Tagging

Part-of-speech tagging involves assigning a part-of-speech label to each token in a text. While this task might be fundamental in natural language processing, it becomes non-trivial when applied to historical languages. The initial algorithms devised for part-of-speech tagging in Greek texts, combined a rule-based approach with a dictionary look-up (Packard, 1973; Crane, 1991). Given that the to-be-tagged text is edited to a classical standard, Crane's algorithm, Morpheus, remains competitive compared to more recent developments, such as RNN Tagger (Schmid, 2019). This neural-based part-of-speech tagger represents the first Greek-specific tagging algorithm introduced since Morpheus. In the three decades between Morpheus and RNN Tagger, existing part-of-speech taggers have been (re-)trained on classical Greek data, ranging from HMM-based (Halácsy et al., 2007) and statistical models (Bohnet and Nivre, 2012), over decision-tree based models (Schmid, 1994; Schmid and Laws, 2008) to Conditional Random Fields (CRF) (Müller et al., 2013).

When tagging morphologically rich languages like Greek, Latin, or Sanskrit, the part-of-speech tag is typically supplemented with the token's morphological features. In the case of Greek, the initial algorithms mentioned above (Packard, 1973; Crane, 1991) provided a complete morphological analysis in addition to their part-of-speech tag. None of those algorithms, however, disambiguate ambiguous word forms, which are quite common in Greek; instead, they provide all possible analyses of a word form. To illustrate, the Morpheus algorithm was unable to provide a single morphological analysis of 47.37% of our test set (cf. Section 3.1). Building upon the survey articles by Celano et al. (2016) and Keersmaekers (2019) which focused on classical and papyrological Greek respectively, Swaelens et al. (2023b) conducted a comparison between RNN Tagger and transformer-based part-of-speech taggers on unedited Byzantine Greek. Drawing inspiration from the exploratory research of Singh et al. (2021), they developed a pipeline that utilises contextualised token embeddings from the DBBErt model[1] as input for a bi-directional Long Short-Term Memory (LSTM) encoder and a CRF decoder, made available by the FLAIR framework (Akbik et al., 2019). As a second approach, they undertook fine-tuning of the contextualised token embeddings directly for part-of-speech tagging. This approach yielded results comparable to those achieved with the combination of a bi-LSTM encoder with a CRF decoder.

[1] https://huggingface.co/colinswaelens/DBBErt

3. Linguistic Annotation

3.1. Data

The majority of NLP techniques outlined in Section 2.2 are trained and evaluated on Classical Greek data sourced from editions. These editions are based on manuscripts, but any inconsistencies encountered are adjusted to fit a Classical Greek model. However, our focus lies in original, unedited texts which are gaining prominence thanks to the growing interest in optical character recognition (OCR) and handwritten text recognition (HTR) (Bhunia et al., 2021; Nockels et al., 2022; Retsinas et al., 2022). Regrettably, the available quantity of unedited Greek data containing linguistic annotation is currently insufficient to compile both a training and test set. At present, we have annotated a test set comprising approximately 10,000 tokens of unedited Byzantine Greek sourced from the DBBE *Occurrences*. We manually provided this test set with part-of-speech tags, morphological features, and lemmas. Further details are comprehensively reported by Swaelens et al. (2023b).

The training data used for the experiment that we present in Section 3.2, is a combination of PROIEL (Haug and Jøhndal, 2008), the Ancient Greek Dependency Treebanks (Celano, 2019; Bamman and Crane, 2011), the Gorman treebanks (Gorman, 2020), the texts provided by Trismegistos (Keersmaekers and Depauw, 2022), and the Pedalion trees (Keersmaekers et al., 2019). From these treebanks, we extracted the part-of-speech tag, morphological analysis, and lemma of each token. Lemmas are not yet taken into account for the experiments presented in this paper because the development of a lemmatiser for unedited Greek is still in progress (Swaelens et al., 2023a, 2024).

3.2. Method

Our initial objective is to offer a full morphological analysis of some 8,000 unedited Byzantine Greek tokens. The tag for this morphological analysis consists of nine slots, each corresponding to one of the following features, as put forward by the universal dependencies framework (Nivre et al., 2016): part-of-speech, person, number, tense, mood, voice, case, gender, and degree of comparison. Previous research adopted a two-step approach: initially predicting only the part-of-speech, followed by a second step where a single label encompassing all morphological features was predicted. Figure 1 depicts the results of two transformer-based approaches for both labelling part-of-speech and conducting morphological analysis (Swaelens et al., 2023b). These results are compared against a most-frequent-label baseline on the one hand, and the RNN Tagger on the other.

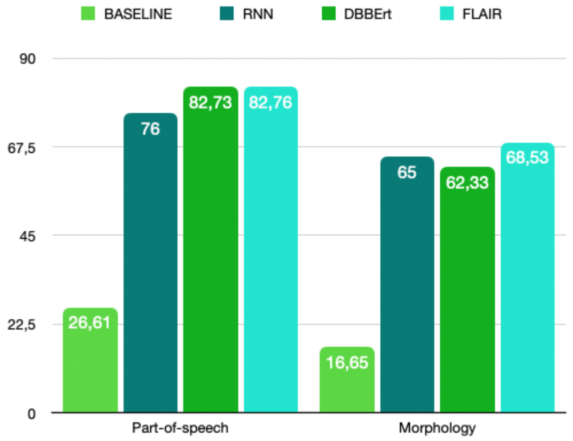

Figure 1: Results of existing transformer-based linguistic annotation of Byzantine Greek.

For the task of part-of-speech tagging they yield accuracy scores of 26.61% and 76.97% respectively. Fine-tuning the Greek transformer-based language model for part-of-speech tagging yielded an accuracy of 82.73%. The second approach, in which the transformer embeddings are processed in the FLAIR framework (cf. Section 2.2), in its turn, resulted in a tagging accuracy of 82.76%. For the task of morphological analysis, the baseline score is 16.65%, while analysis by the RNN Tagger resulted in 65.59%. With an accuracy of 62.33%, the DBBErt model fine-tuned on morphology performs 3 pp. less than the RNN Tagger. When the transformer embeddings are utilised within FLAIR, the output slightly outperforms the RNN Tagger by 3 pp., achieving an accuracy of 68.53%.

Previous research has highlighted that the drop in performance between part-of-speech labelling and morphological analysis may be attributed to the magnitude of the morphological label set. This label set comprises 1,057 possible labels, whereas the part-of-speech labels amount to 14. However, it is noteworthy that the training for both tasks is conducted on the same, relatively modest training set. Nevertheless, we aim to elevate the performance of the automatic morphological analysis. Therefore, both a more novel and a more traditional approach are trained and evaluated for this classification task.

3.2.1. Transformer-based Approach

In our first experiment, we fine-tuned the DBBErt model for each of the nine features outlined in Section 3.2. Except for the feature 'part-of-speech', our biggest label set counts only 9 labels, while the smallest comprises no more than 4. The accuracy of each classifier ranged from 82.73% for case to 96.24% for person. We have excluded the scores for degree of comparison, since the classifier labelled all tokens with '-', which indicates this

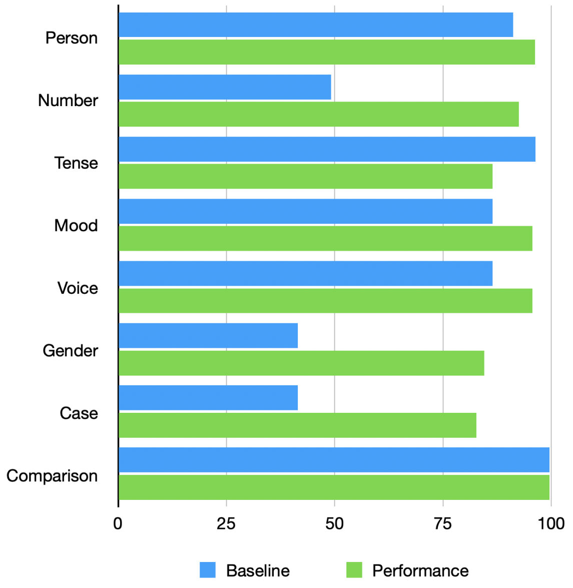

Figure 2: The performance of each of the fine-tuned DBBErt classifiers for the morphological features.

feature lacks labels. To verify that the other classifiers learned more than the one for the degree of comparison, Figure 2 displays the baseline of the most frequent label for each classifier alongside its performance. Despite these promising results, upon assembling the output of all classifiers, the accuracy of the combined label was only 58.4%. A recurring problem with the assembly method is the assignment of redundant features, such as nouns being labelled with ‘present’ for tense instead of ‘-’. To address this issue and prevent the assignment of redundant features, our second experiment employs a cascaded approach.

In this cascaded approach, the first step involves assigning the part-of-speech to a given token. Subsequently, only the classifiers of the features specific to the part-of-speech are employed to predict the label of that feature. For instance, if the part-of-speech is a noun, adjective, or pronoun, only the number, case, and gender classifiers predict a label, while the other features are automatically labelled as ‘-’. If the part-of-speech is a verb, we first determine the mood of the verb to predict the correct features. All verbs share the features voice, mood, and tense. An infinitive has no additional features, so the other slots are labelled as ‘-’. The indicative, imperative, subjunctive, and optative share the additional features person and number. The participle, on the other hand, has the additional features case, gender, and number. This cascaded approach, which combines rules with transformer-based classifiers, yielded an accuracy score of 58.29%. Contrary to our hypothesis, our cascaded approach did thus not outperform the assembly method.

These experiments suggest that transformer-based classifiers may not be suitable for the automatic morphological analysis of unedited Greek. Consequently, we explored a more traditional classification approach: support vector machines (SVM). These SVM classifiers are fed the transformer embeddings from the DBBErt model as input, a method known to be quite efficient for classification tasks (De Geyndt et al., 2022).

3.2.2. SVM

Typically, more traditional feature-based machine learning algorithms like SVM rely on manually crafted features, such as local context (preceding or next word) or linguistic information like part-of-speech. However, for this experiment, we generated transformer embeddings with the DBBErt_pos2024 model[2]. Since DBBErt_pos2024 is fine-tuned on part-of-speech tagging, these embeddings contain not only contextual information but also part-of-speech information. We adopted an approach similar to the one presented in Section 3.2.1.

Firstly, we trained an SVM classifier for the complete morphological tag, which resulted in an accuracy score of 39.43%. However, it classified practically all tokens as the punctuation label (u--------). When predicting the complete label at once, the SVM exhibited a drop in performance of almost 30 pp. compared to the best algorithm of Figure 1.

Secondly, we trained distinct SVM classifiers, similar to the approach with the nine transformer classifiers. This time, however, to conserve computational resources, we began with the morphological features of nouns, adjectives, and pronouns: case, gender, and number. These classifiers yielded accuracies of 75.34%, 90.94%, and 72.83% respectively. When these labels were assembled and redundant slots were assigned ‘-’, this classification approach yielded an accuracy score of 58.07%. As the morphological features of the Greek verbal system are much more complex than those of the nominal system (more relevant features with more options), we decided not to train classifiers for the remaining morphological features for verbs, as they would likely perform even worse than the classifiers for nominal features.

4. Similarity Detection

Given the scope of this exploratory paper, the detection of similar texts is limited to identifying similar verses of unedited Byzantine Greek. The similarity detection relies not only on the orthographic similarity measures, as described in Section 2.1, but

[2] https://huggingface.co/colinswaelens/DBBErt_pos2024

also on the combination of these methods with automatically provided linguistic information. In an ideal scenario, the linguistic information consisted of both the part-of-speech tag and a full morphological analysis. However, since the tool for automated morphological analysis requires further improvement, the linguistic information integrated into the pipeline is limited to part-of-speech tags. The subsequent sections offer a detailed description of the workflow outlined in Figure 3.

4.1. Graph Database

In graph databases data are organised by means of graphs, unlike traditional relational databases where data are structured in tables (Angles and Gutierrez, 2008). Such graphs consist of nodes and relationships (or edges) connecting these nodes. Due to their structure, graph databases excel in handling highly interconnected data (Batra and Tyagi, 2012), making them an ideal tool for storing a large number of similarity relationships between texts, with each text represented by a node. Furthermore, graph database systems allow for advanced and visual analysis of the numerous interconnections between nodes, providing an ideal instrument for detecting and analysing similar texts.

For this paper, we have established such a graph database to store verses of Byzantine book epigrams. Before importing the texts into the graph, the verses undergo preprocessing to standardise them and reduce noise, thereby facilitating similarity calculation in later steps of the process. The preprocessing involves converting uppercase characters to lowercase and removing punctuation and diacritics. Subsequently, these preprocessed verses are stored in dedicated verse nodes in the graph. However, verses that maintain the exact same spelling after preprocessing, are stored in a single node.

Not only complete verses but also individual words are stored in the graph. Words are tokenised by splitting up the preprocessed verses based on white spaces, and like verses, these words are stored in dedicated word nodes, where identical words are – again – represented by a single node. Nodes representing words are also connected to the verse node in which they appear. These relationships include information about both the rank and the part-of-speech tag of that word within the connected verse.

4.2. Method

Utilising the preprocessed verses and tokens already stored in the graph database, our objective is to compute three similarity scores between each pair of verse nodes: orthographic similarity, part-of-speech-based similarity, and a combination of both. The outcome of these similarity calculations is a score between 0 and 1, denoting the degree of similarity between two verses based on the specific similarity measure employed. A score of 0 indicates complete dissimilarity between two verses, whereas a score of 1 signifies complete similarity. The remainder of this section provides a succinct description of these three similarity measures.

4.2.1. Orthographic Similarity

The orthographic similarity between verse nodes is determined by employing an implementation of Deforche et al. (2024), utilising the default parameters of the algorithm[3]. This similarity measure firstly calculates the similarity between all word pairs, then utilises these word-level similarities to ascertain the similarity between all verse pairs. The process of determining the similarity between two words begins by computing the Damerau-Levenshtein edit distance (Damerau, 1964), which represents the minimal cost required to transform one word into another using one of the four supported edit operations: insertion, deletion, replacement of a single character, or the transposition of two consecutive characters. In this paper, we assume the cost of all mentioned edit operations to be equal to 1. The word-level similarity score is then obtained by dividing the resulting edit distance by the length (in characters) of the longest of the two words and subtracting this number from 1. In the case of Byzantine texts, this word-level similarity is computed without penalising either the itacism[4] nor diacritics. This means that, for example, the similarity between ξένοι and ξενη is 1, indicating that these words are treated as identical.

Next, a similar process is repeated to calculate the orthographic similarity scores between all pairs of verse nodes. In this case, the edit distance is calculated between two verses using the same four edit operations, but rather than considering individual characters, entire words are taken into account. Once again, all edit operations are presumed to have a cost of 1, except for the replacement operation between two words. In the case of replacements, the cost equals the dissimilarity between the word and its potential replacement, which can be determined by subtracting the precomputed similarity between those words from 1. Lastly, the edit distance between two verses needs to be converted into a similarity score. This is accomplished by dividing the resulting edit distance by the length (in words) of the longest verse and subtracting this number from 1. The resulting orthographic similarity score between two verse nodes is stored in the

[3] https://github.com/MaximeDeforche/DBBESimilarity

[4] The itacism is a phonetic shift of ει, η, ι, οι, υ into [i].

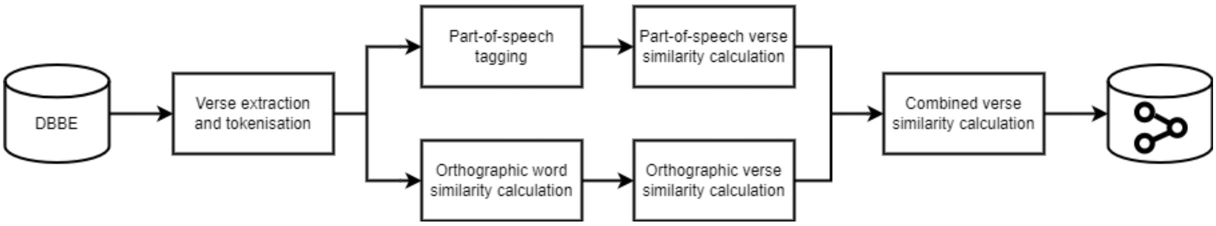
DBBE
Verse extraction and tokenisation
Part-of-speech tagging
Orthographic word similarity calculation
Part-of-speech verse similarity calculation
Orthographic verse similarity calculation
Combined verse similarity calculation

Figure 3: Workflow from relational database with plain text to a linguistically annotated graph database, including similarity scores between texts.

graph database by means of a relationship between those two nodes.

Further details on the implementation and customisation options of this orthographic similarity measure are comprehensively reported by Deforche et al. (2024).

4.2.2. Part-of-speech Similarity

Next, we compute a similarity measure based on the part-of-speech tags assigned to each word. For this similarity measure, we draw inspiration once again from the Damerau-Levenshtein edit distance (Damerau, 1964) to compute a part-of-speech similarity score between all verse nodes. For each verse node, we concatenate the part-of-speech tags of all words in a verse into a single string, maintaining the same order as the appearance of the words in that verse. Subsequently, the edit distance between these strings is determined by calculating the minimal cost required to transform one part-of-speech representation of a verse into the other. In this paper, the supported edit operations all have a cost of 1 and include the insertion, deletion, and replacement of a single part-of-speech tag, as well as the transposition between two consecutive part-of-speech tags. The resulting edit distance is then transformed into a similarity score by dividing it by the length (in words) of the longest verse and subtracting this result from 1. Finally, the resulting part-of-speech similarity score is stored in a similarity relationship connecting the two verse nodes for which this similarity is calculated.

As an example, we consider two verses that are represented by the part-of-speech tags of each word they consist of. The first verse consists of the tags: adverb (`d`), adjective (`a`), verb (`v`), noun (`n`) and verb (`v`), and the second verse of the tags: adverb (`d`), interjection (`i`), verb (`v`), verb (`v`), noun (`n`) and noun (`n`). First, the edit distance between `davnv` and `divvnn`, which are the concatenated part-of-speech tags of both verses, is determined. The edit operations to transform the concatenated tags from one verse into the other are visualised by Figure 4 and consist of a replacement (orange), a transposition (crossing arrows), and a insertion/deletion (green/red), resulting in a total edit distance of 3. Using this edit distance, the part-of-speech similarity between the verses is calculated using the method described above and results in similarity score of 0.5.

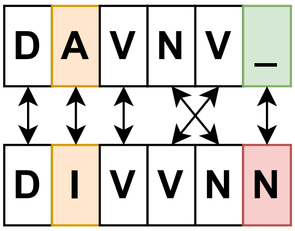

Figure 4: Edit operations between two concatenations of part-of-speech tags.

4.2.3. Combined Similarity

As a third and final measure, we aim to compute a similarity score between each pair of verse nodes that considers both the orthographic information and the automatically provided linguistic information. We calculate this score by averaging the orthographic and part-of-speech similarities already determined for each pair of verse nodes. In future research, we plan to explore more advanced and customisable options like the Ordered Weighted Average (OWA) operators (Yager, 1988; Yager and Kacprzyk, 2012) or the Logic Scoring of Preference (LSP) method (Dujmovic, 2018). This combined measure results in a balanced similarity score that considers both orthographic and part-of-speech similarities.

In a theoretical illustration, let us consider that the orthographic similarity between two verses is 0.6, while the part-of-speech similarity is to 0.8. Through the amalgamation of these two scores, we arrive at a combined similarity score of 0.7.

In parallel with the other similarity scores, these results are stored on the relationship between verse nodes in the graph database, allowing us to analyse the relations between all verses based on this hybrid similarity measure.

4.2.4. Visual Grouping of Verse Nodes

Upon computing the similarity scores between each pair of verse nodes, we harness the querying capabilities of the graph database to identify and thoroughly analyse verses that exhibit (dis)similarities. Utilising a specified similarity threshold, the graph database can be queried to reveal all verse nodes and their associated similarity relationships of a specific similarity measure scoring equal to or exceeding the specified threshold. Through visual representation of such query results, we observe the emergence of groups of verse nodes that demonstrate at least the specified level of similarity according to the chosen similarity measure. Opting for a high similarity threshold yields numerous groups of highly similar texts, whereas a lower similarity threshold produces fewer groups of texts with lower degrees of similarity. Although initially counterintuitive, selecting a lower similarity threshold can be interesting, particularly when examining texts rife with spelling variations or orthographic inconsistencies, such as Byzantine book epigrams.

The ability to select the similarity measure and threshold provides researchers with the flexibility to analyse texts in myriad ways. The similarity measures outlined in this paper offer the capability to visually identify similar verses based on their orthographic properties, linguistic information, or a blend of both.

5. Case Study

5.1. Data

For our case study, we will compute similarities between verses linked to *Types* 2148, 2150, and 4245 from the DBBE (Demoen et al., 2023). These *types* group 154 DBBE *occurrences*, resulting in a set of 410 verses. Given that identical verses are stored only once, this set is stored as 286 unique nodes in the graph database. No duplicate values are stored. Among these verses, 1a and 1c are shared across all three Types. Presumably, the number of verses gave rise to three distinct Types. The *occurrences* grouped under Type 2150, for instance, all comprise three verses, whereas Type 2148 encompasses *occurrences* consisting of only two verses; Conversely, Type 4245 links *occurrences* totalling six verses.

(1) a. Ὥσπερ ξένοι χαίρουσιν ἰδεῖν πατρίδα,
Hōsper xenoi chairoysin idein patrida,
Just like travellers rejoice by seing their homeland,

b. καὶ οἱ θαλαττεύοντες εὑρεῖν λιμένα,
kai hoi thalatteuontes eurein limena,
and sailors by finding a harbour,

c. οὕτως καὶ οἱ γράφοντες βιβλίου τέλος.
houtōs kai hoi grafontes bibliou telos.
so do scribes at the end of a book.
DBBE Type 2150

5.2. Orthographic Similarity

To showcase the capabilities of this dynamic system, we provide a visual render of verse grouping based on orthographic similarity exceeding 85% (Figure 5). The computation of this similarity measure involves two main steps. First the similarity between two words is computed without penalising either the itacism nor diacritics, as they appear arbitrary throughout the corpus (cf. Section 4.2.1). Then the similarity score of the verses is computed by combining the word similarity scores.

The group highlighted within the yellow frame in Figure 5, represents variants of verse 1a. This visual shows minimal outliers, indicating a high level of similarity between the verses. Notably, the word that causes most ‘dissimilarity’ is the third word of verse 1a, χαίρουσιν. Despite not penalising the itacism, the participle χαίροντες still displays a 55% similarity to the indicative χαίρουσιν, accounting for one-fifth of the verse’s overall similarity score. Verses within the blue frame are variants of Example 1a differing only in the use of the infinitive βλέπειν *blepein* (to look at) instead of ἰδεῖν *idein* (to see). Although semantically nearly identical, the variant using βλέπειν shows no similarity with the majority using ἰδεῖν.

The red frame encompasses verses like Example 1b. Surprisingly, 4 of the 43 verse variants contain a participle of the word κινδυνεύω *kinduneuō* (run risk) instead of the expected θαλαττεύω thalatteō (to be at sea). Despite them being unrelated, the similarity between these two participles is still 54%, which again accounts for one fifth of the verse similarity.

Verses grouped in the orange frame represent Example 1c. However, this group consists of two distinct parts connected by what we would call *bridge verses*. The left part lacks the verb ἰδεῖν preceding βιβλίου τέλος *bibliou telos* (the end of the book), including Example 1c. The right part, on the other hand, does have ἰδεῖν before βιβλίου τέλος. Additionally, this group has a variant that is not linked with this similarity measure: the green group. These variants do not display as subject the more common nominative οἱ γράφοντες *hoi grafontes* (the writers) as in Example 1c, but use instead a dative construction with τοῖς γράφοντοις *tois grafontois* (to the writers). Figure 6 provides a detailed visual of the differences between the dative construction of the verse variant on the right and the nominative construction of the verse variant on the left . The orthographic dissimilarity of both the article and the noun results in an orthographic similarity of 81.5% between the two verses.

The pink frame encompasses nine verse variants all counting more than three verses. The structure of the sentence follows that of Example 1b: καὶ *kai* (and) [placeholder] εὑρεῖν *heurein* (finds) [placeholder]. The first of the two placeholders is either a noun or a participle, wihle the second is most often a noun. If the only difference within one verse is the use of a participle of a different verb, as in the κινδυνεύω/θαλαττεύω example supra, the similarity score is still quite high. In these verses, for example, the second placeholder is filled with τὸ κέρδος *to kerdos* (profit), λιμένα *limena* (harbour) or νήκος *nèkos* (victory). These last two, display 0% and 33% similarity respectively to τὸ κέρδος, and 0% to each other. Combined with the dissimilarity in the first placeholder, results in these verse variants being grouped separately for this similarity measure.

The remaining verse variants will not be elaborated upon as these verses are not connected to more than two other verse variants. Most of them are incomplete verses due to lacunae.

It is important to keep in mind that Figure 5 provides a static representation, reflecting groups with a similarity score equal to or higher than 85%. However, the underlying system is dynamic, allowing adjustments to the similarity threshold which can be set lower or higher, and considerations for the itacism or other phonetic changes which can or cannot be penalised.

5.3. Implementation Part-of-Speech

This system could become even more dynamic with the implementation of linguistic annotation. Depending on your query, linguistic annotation could either refine search results by limiting them to specific parts-of-speech within verses, or, on the other hand, it could broaden the scope to include verses that display similarity based solely on part-of-speech information. As discussed in Section 3.2.2, in this paper the linguistic annotation is restricted to automatically labelled part-of-speech tags.

Once the similarity scores, as described in Section 4.2.3, are computed between verses in our dataset, the results are visualised in Figure 7. Notably, there are fewer verses that do not belong to any group compared to Figure 5. Another observation is the absence of verses from the green group in Figure 5. This is because the combination of part-of-speech information and orthography in a single similarity measure mitigates orthographic dissimilarities caused by the dative suffix resulting in the inclusion of those verses in the orange group. In Figure 6, the edge between the yellow verse nodes not only displays the orthographic similarity (81.5%) but also their combined similarity (90.7%), based on the part-of-speech labels visible on the edges between the yellow verse nodes and the green word nodes, representing their part-of-speech within that specific verse.

Similarly, one might anticipate the variants of Example 1a within the blue frame to integrate into the yellow group. However, despite the addition of part-of-speech information, these variants remain isolated. This suggests that part-of-speech information alone does not offset the penalisation of orthography and word order. Notably, the verses in the yellow group end with ἰδεῖν πατρίδα *idein patrida*, while those in the blue group end with πατρίδα βλέπειν *patrida blepein*.

6. Conclusion

We set out to explore the potential of a dynamic tool to assist scholars in their philological research endeavours. Our system operates in two main parts: first, the data is annotated with linguistic information; subsequently, users can select a similarity measure and define a threshold for similarity computation within the graph database. Currently, the linguistic information is limited to the automatic assignment of part-of-speech tags. The similarity measures presented include a purely orthographic measure, one based solely on part-of-speech, and a combined measure that integrates both aspects. Users have the flexibility to adjust the similarity measure and its threshold, tailoring the results to be either broad (with a lower similarity threshold) or specific (with a higher similarity threshold). With sufficient data in the graph database, scholars can uncover new relevant text segments to incorporate into their analysis or discover allusions to other authors for commentary purposes.

In future work, we plan to expand the relaxation rules of the itacism to include other phonetic changes in Byzantine Greek. We will also implement automatic morphological analysis, resulting in additional combined similarity measures. Furthermore, our focus will extend from orthographic to semantic similarity measures, exploring how these methods can be both flexibly and effectively combined in a manner that is specific to the field of study. We anticipate close collaborations with philologists to conceptualise a demo that will make this technology accessible to to the wider academic community.

7. Bibliographical References

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Renzo Angles and Claudio Gutierrez. 2008. Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1–39.

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Shalini Batra and Charu Tyagi. 2012. Comparative analysis of relational and graph databases. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(2):509–512.

Ayan Kumar Bhunia, Shuvozit Ghose, Amandeep Kumar, Pinaki Nath Chowdhury, Aneeshan Sain, and Yi-Zhe Song. 2021. Metahtr: Towards writer-adaptive handwritten text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15830–15839.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics.

Antoon Bronselaer and Guy De Tré. 2009. A possibilistic approach to string comparison. *IEEE Transactions on Fuzzy systems*, 17(1):208–223.

Giuseppe G. A. Celano, Gregory Crane, and Saeed Majidi. 2016. Part of speech tagging for ancient greek. *Open Linguistics*, 2(1).

Giuseppe GA Celano. 2019. The dependency treebanks for ancient greek and latin. *Digital Classical Philology*, page 279.

Gregory Crane. 1991. Generating and parsing classical greek. *Literary and Linguistic Computing*, 6(4):243–245.

Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Ellen De Geyndt, Orphée De Clercq, Cynthia Van Hee, Els Lefever, Pranaydeep Singh, Olivier Parent, and Veronique Hoste. 2022. Sentemo : a multilingual adaptive platform for aspect-based sentiment and emotion analysis. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 51–61. Association for Computational Linguistics (ACL).

Maxime Deforche, Ilse De Vos, Antoon Bronselaer, and Guy De Tré. 2023. An orthographic similarity measure for graph-based text representations. In *Flexible Query Answering Systems*, pages 206–218, Cham. Springer Nature Switzerland.

Maxime Deforche, 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).

Kristoffel Demoen, 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 Pechillon, 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jozo Dujmovic. 2018. *Soft computing evaluation logic: The LSP decision method and its applications*. John Wiley & Sons.

Najlah Gali, Radu Mariescu-Istodor, Damien Hostettler, and Pasi Fränti. 2019. Framework for syntactic string similarity measures. *Expert Systems with Applications*, 129:169–185.

Wael H Gomaa, Aly A Fahmy, et al. 2013. A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18.

Vanessa B Gorman. 2020. Dependency treebanks of ancient greek prose. *Journal of Open Humanities Data*, 6(1).

Péter Halácsy, Andras Kornai, and Csaba Oravecz. 2007. Hunpos: an open source trigram tagger. *Proceedings of the 45th Annual Meeting of the*

Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 209–212.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Matthew A Jaro. 1995. Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7):491–498.

Alek Keersmaekers. 2019. Creating a richly annotated corpus of papyrological Greek: The possibilities of natural language processing approaches to a highly inflected historical language. *Digital Scholarship in the Humanities*, 35(1):67–82.

Alek Keersmaekers and Mark Depauw. 2022. Bringing together linguistics and social history in automated text analysis of greek papyri. *Digital Classics III: Re-Thinking Text Analysis (Classics@)*.

Alek Keersmaekers, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, enriching and valorizing treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 109–117, Paris, France. Association for Computational Linguistics.

Grzegorz Kondrak. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Vladimir I et al. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joe Nockels, Paul Gooding, Sarah Ames, and Melissa Terras. 2022. Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of transkribus in published research. *Archival Science*, 22(3):367–392.

David W. Packard. 1973. Computer-assisted morphological analysis of Ancient Greek. In *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*.

George Retsinas, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. 2022. Best practices for a handwritten text recognition system. In *International Workshop on Document Analysis Systems*, pages 247–259. Springer.

Andreas Rhoby. 2018. *Ausgewählte byzantinische Epigramme in illuminierten Handschriften*. Verlag der österreichischen Akademie der Wissenschaften, Wien.

Rachele Ricceri, 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 & Digital Humanities*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Helmut Schmid. 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, pages 133–137.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, page 777–784, USA. Association for Computational Linguistics.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for bert language modelling and morphological analysis for ancient and medieval greek. In *The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, co-located with EMNLP 2021*, pages 128–137. Association for Computational Linguistics.

Colin Swaelens, Ilse De Vos, and Els Lefever. 2023a. Evaluating existing lemmatisers on unedited byzantine Greek poetry. In *Proceedings of the Ancient Language Processing Workshop*, pages 111–116, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Colin Swaelens, Ilse De Vos, and Els Lefever. 2023b. Linguistic annotation of byzantine book epigrams. *Language Resources and Evaluation*.

Colin Swaelens, Ilse De Vos, and Els and Lefever. 2024. Lemmatisation of medieval greek: Against the limits of transformers' capabilities? In *Proceedings of the Fourteenth Language Resources and Evaluation Conference*, Turin, Italy. European Language Resources Association.

William E Winkler. 1990. *String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage.* ERIC.

Ronald R. Yager. 1988. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190.

Ronald R. Yager and Janusz Kacprzyk. 2012. *The ordered weighted averaging operators: theory and applications*. Springer Science & Business Media.

Appendices: Figures

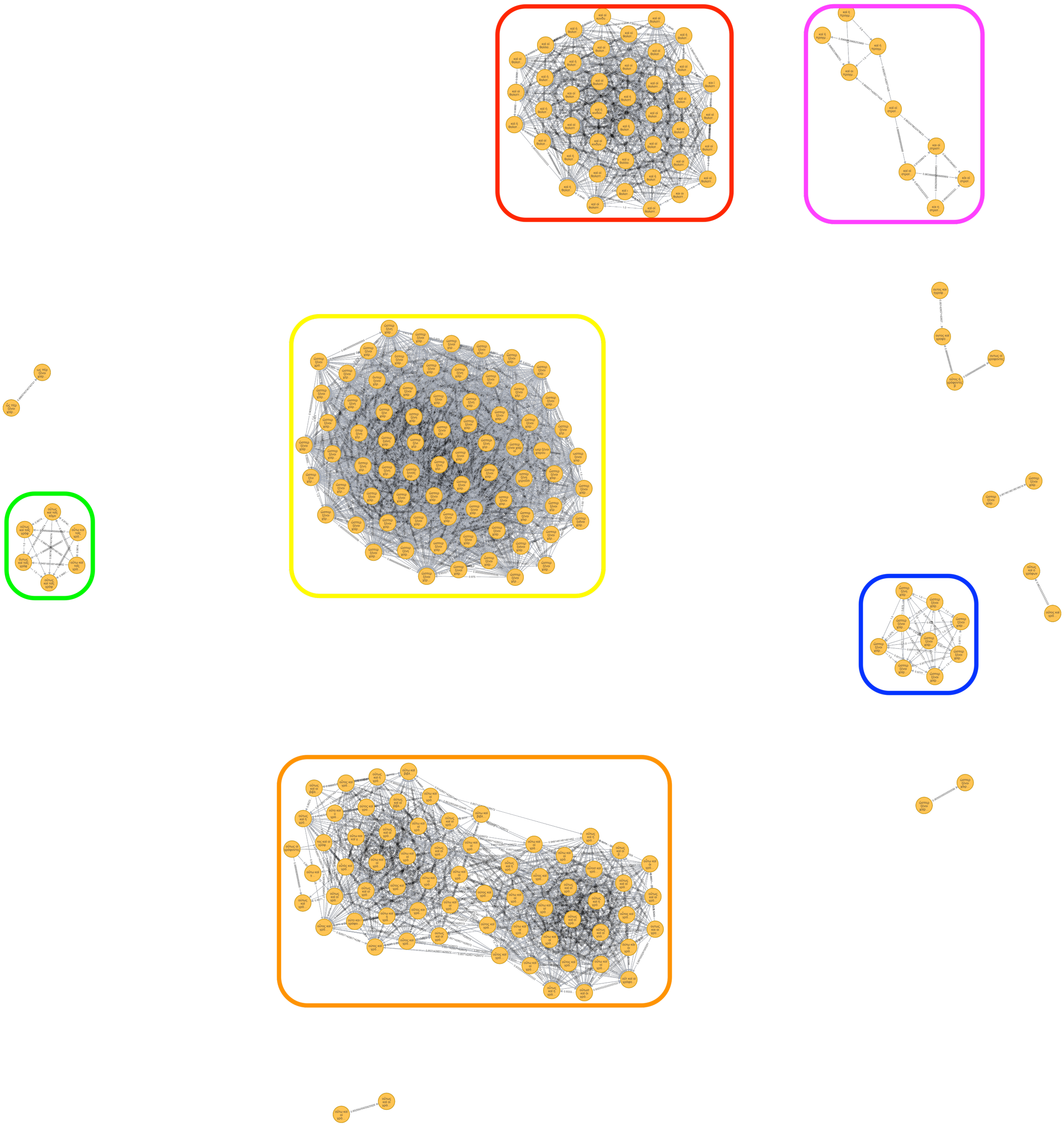

Figure 5: Orthographic similarity between verses

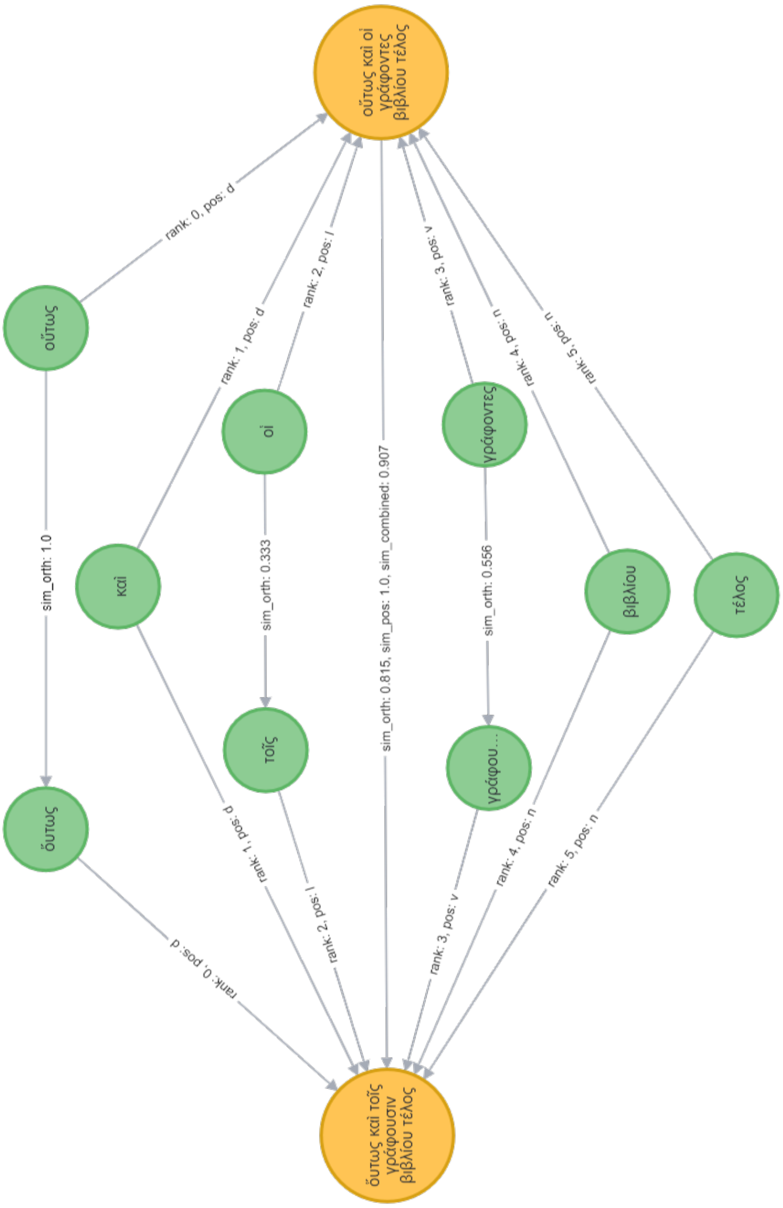

Figure 6: Detailed figure of verse variants of Example 1c: left with a dative construction, right with a nominative construction.

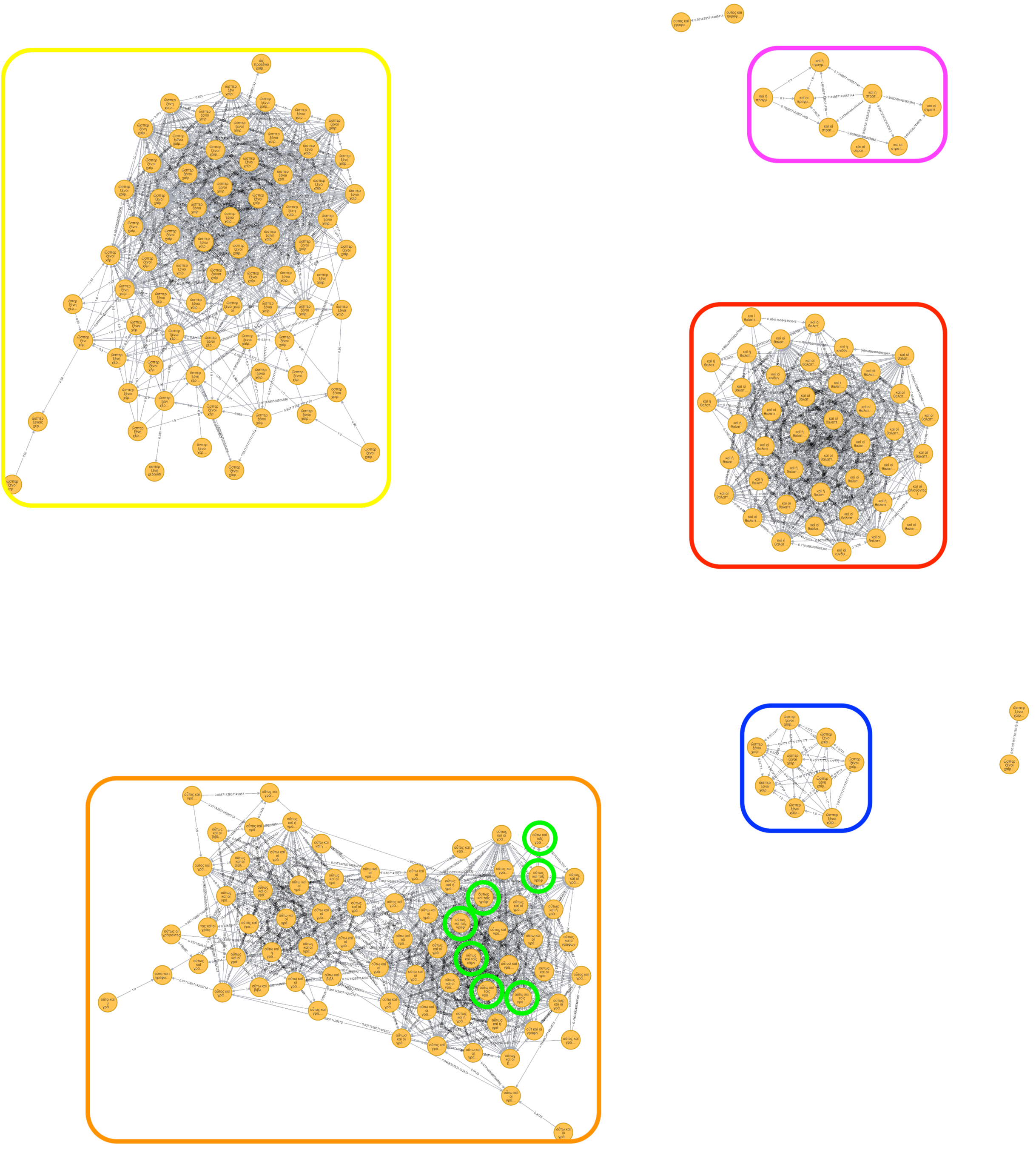

Figure 7: Bridge verses of orange group