SCHOOL OF
ECONOMICS AND
MANAGEMENT

Master's Programme in Data Analytics and Business Economics

Semantic Similarity Analysis on English Translations of the Iliad

A study on the influence of different features on the semantic similarity scores based on document embeddings

Author:

Maria Bijkerk

Lund University School of Economics and Management

DABN01

Master's Thesis (15 credits ECTS)

May 2024

Supervisor:

Jakob Bergman

Abstract

Studying translations gives us more insight into cultures and languages. Machine Translation is an application area of the field Natural Language Processing (NLP), used to transfer information from one language to another. Creating these tools require a lot of data, including data about the semantic relationships of the texts, and for unspoken languages like Ancient Greek, there does not exist a lot of (digital) data. In this study, we explore 16 different English translations of the first book of the Iliad, an Ancient Greek epic seen as one of the most influential literary works on modern western literature. We use three different algorithms (GloVe, Word2Vec, and BERT) to create document embeddings for each translation. We then analyse how three features (publication year, genre, name versions) influence the cosine similarity scores between the documents. We also use hierarchical clustering to group the translations together without needed a pre-determined number of clusters, to see how the full document embeddings relate to each other. We find that the publication year does not have a significant influence on the similarity scores, but the genre and name versions do seem to have a significant influence.

Acknowledgements

I would like to thank my supervisor Jakob Bergman for guiding and supporting me through the process of writing this thesis. Thank you for providing me with useful feedback and ideas, and for supporting my slightly untraditional topic ideas.

I also want to thank our programme director Joakim Westerlund for his great amount of information before and during the thesis period. This made the process of finding a topic to writing the paper as smooth as possible.

Lastly, I want to thank my little sister Elien. Thank you for being as enthusiastic about Greek Mythology as me and talking me through my chaotic ideas even though you are multiple countries away and do not understand half of what I am talking about.

List of Figures

Figure 1: Number of lines in book 1 per translation grouped by prose (green) and verse (blue) 13
Figure 2: Most frequently used words through all documents 16
Figure 3: Scatter plot of the translations based on the frequency matrix 17
Figure 4: Similarity matrix GloVe 22
Figure 5: Similarity matrix Word2Vec 22
Figure 6: Similarity matrix BERT 22
Figure 7: Scatter plot of GloVe, Word2Vec, and BERT embeddings 23
Figure 8: Scatter plot of GloVe, Word2Vec, and BERT embeddings, grouped by century 25
Figure 9: Scatter plot of GloVe, Word2Vec, and BERT embeddings, grouped by name version 27
Figure 10: Scatter plot of GloVe, Word2Vec, and BERT embeddings, grouped by genre 28
Figure 11: Dendrograms for each model 31

List of Tables

Table 1: Author, publication year, genre, metre and number of lines of the 16 translations used in this study 12
Table 2: The first lines of each of the 16 translations used in this study 14
Table 3: Cluster information BERT cluster orange 30
Table 4: Cluster information BERT cluster green 30
Table 5: Cluster information BERT cluster red 30
Table 6: Term-frequency matrix of the 20 most frequently used words 38
Table 7: average pairwise and maximum distances for each century based on the GloVe similarity matrix 39
Table 8: average pairwise and maximum distances for each century based on the Word2Vec similarity matrix 39
Table 9: average pairwise and maximum distances for each century based on the BERT similarity matrix 39
Table 10: average pairwise and maximum distances for each genre based on the GloVe similarity matrix 40
Table 11: average pairwise and maximum distances for each genre based on the Word2Vec similarity matrix 40
Table 12: average pairwise and maximum distances for each genre based on the BERT similarity matrix 40
Table 13: Cluster information GloVe cluster orange 41
Table 14: Cluster information GloVe cluster green 41
Table 15: Cluster information GloVe cluster red 41
Table 16: Cluster information Word2Vec cluster orange 42
Table 17: Cluster information Word2Vec cluster green 42

Contents

Abstract 2
Acknowledgements 3
List of Figures 4
List of Tables 4
Introduction 6
Background 7
The Iliad 7
Translations 8
Machine translation 10
NLP and semantic textual similarity 10
Data 12
Data exploration 13
Methodology 15
Pre-processing 15
Term-document matrix 15
Models 18
Word embeddings 18
Word2Vec 18
GloVe 18
BERT 19
Multidimensional scaling 20
Hierarchical clustering 20
Results 21
Style 24
Hierarchical Clustering 29
Discussion 32
Conclusion 33
References 34
Appendix I 38
Appendix II 39
A. 39
B. 40
Appendix III 41
A. 41
B. 42

Introduction

Natural Language Processing (NLP) is an interdisciplinary field that combines Computer Science, statistics, and linguistics. NLP is concerned with automatically analysing and representing human languages (Chowdhary, 2020). This field encompasses many methods and tasks, like question answering (Devlin et al., 2018; Wang et al., 2018), speech recognition (Kamath et al., 2019; Murveit & Moore, 1990), and text summarization (Kiyani & Tas, 2017).

Developments in NLP have provided new ways of analysing texts. NLP has also helped developing techniques for translation analysis with the help of machine learning. Translation analysis is a field that has existed long before NLP, but with the help of computers we can discover patterns and relations in texts much faster, or even discover connections that we have not been able to see before.

Differences in translations give us more information about the history of language, cultures, and interpretations of important works throughout the years (Geng et al., 2015). In 1959, Roman Jakobson defined three types of translation: interlingual, defined as translating one language into another, intralingual, translating within the same language, e.g., translating old English into modern English, and intersemiotic, which he defined as the transfer of verbal signs into non-verbal signs, e.g., creating a movie from a book (Jakobson, 1959).

Another NLP task is machine translation. This task is concerned with teaching computers about human language to create models and programmes that can analyse and adapt languages. For these models to work well, a lot of training data is needed. Ancient languages are lacking in the amount of data available, so are falling behind in this field (Yousef et al., 2022).

The Iliad is seen as one of the pillars of Western Literature. Since it is written in Ancient Greek, a language that is not spoken anymore, translations are almost always used for both casual reading as well as research. The Iliad is an epic poem, written in a rhyme scheme that is not fully replicable in the English language. Next to this, Ancient Greek and English have very different linguistic features, preventing a direct translation of the work. Each English translation will have the interpretation and style of the translating author present in the work, making each translation at least a little bit different from each other.

The Iliad has been a topic of many quantitative analyses, but there is not a lot of studies using machine learning to analyse the texts. In this research we will analyse 16 different English translations of the first book of the Iliad to try to identify which features cause the texts to be similar using document embeddings. We will focus on features related to the style of the texts.

We will use three methods to get the embeddings: GloVe, Word2Vec, and BERT. By combining these techniques, we will get a well-rounded document embedding for each translation that can be used for semantic similarity analysis to identify similar patterns. Embeddings learned by methods as GloVe and Word2Vec have been used by several NLP applications, one of which is machine translation (Lau & Baldwin, 2016). NLP tools are important because they enable systematic analysis and comparison of translations, revealing linguistic patterns, cultural influences, and interpretive differences that enhance our understanding of both the source texts and their translations.

The aim of this study is to make a first attempt of using machine learning techniques to inspect translations between Ancient Greek and English to learn more about textual features of Ancient Greek in relation to English so we can develop well working machine translation models.

In the next sections we will go through previous research on the relevant topics, explore the data, and go through the analysis.

Background

The Iliad

The Iliad is argued to be one of the oldest substantial pieces of European literature (Dedović, 2018; Ilyas, 2022; Kim, 2023; Mendelsohn, 2011). The Ancient Greek epic poem, attributed Homer, consists of 24 books. One for each of the letters of the Greek alphabet. The epic is set at the end of the Trojan war and tells the story of the wrath of the hero Achilles. The books are written in dactylic hexameter, also known as heroic hexameter. It is a rhythmic scheme that was widely used in Greek and Latin poems. Hexameter means that each line of the poem consists of six metrical feet (counts). Each foot can be either a dactyl, consisting of one long syllable followed by two short syllables, or a spondee, consisting of two long syllables. There are many rules about the words that can be used in these feet and the compositions of dactyls and spondees in a line (Ingalls, 1970). English is a stress-timed language where rhythm is created by stressed and unstressed syllables (Low, 2006). Replicating the precise pattern of a text using a metre that creates rhythm from long and short syllables is thus a challenge.

A metre that is based on stressed and unstressed syllables is iambic pentameter, a rhythmic scheme consisting of five metrical feet per line. Each foot is an iamb. This has two syllables of which the first is unstressed and the second is stressed (Kiparsky, 2020). This metre allows for the natural flow of the English language, which is why it is popular metre that has been used by many famous English poets like Shakespeare and Alexander Pope. Pope is one of the authors whose translation of the Iliad we’re using in this study.

Translations of the Iliad have been a topic in translation studies for years. Exploring the translations is believed to lead to an important foundation of knowledge about modern literature and translation principles (Ilyas, 2022).

Pavlopoulos et al., (2022) found that a mechanical annotator can be created with a deep learning model for sentiment estimation that has a low error rate. They used the modern Greek translations of the Iliad for their study, which can be seen as the spoken language that is the most similar to Ancient Greek. Their results can thus be seen as a benchmark for sentiment analysis for other languages. In addition, mechanical annotators play a crucial role in machine translation. They can be used for example for quality control (Grosman et al., 2020) or translation evaluation tasks (Palladino et al., 2022).

Another study on the Iliad shows that translating Ancient Greek language into English is not as straightforward as other languages. Dedović (2018) explores the different Ancient Greek words that can all be translated into the English word ‘mind’ and their use in the Iliad and the Odyssey. He also examines how the use of these words vary in the Iliad compared to the Odyssey. The study finds that there is less mental language, words that have something to do with a mental action, in the Iliad than the Odyssey, but the Odyssey uses fewer different words. Dedović expects this to be the case because the Odyssey is written later than the Iliad and the language, which consisted of many different dialects, became more standardized.

Kim (2023) compared the prefaces of the Korean translations of the Iliad written by Lim Hak-Su. This writer has written three editions of the Iliad over the years. The prefaces show indications of the political situation in North Korea around each period of when the translations were created. This study shows that political systems and ideologies can also cause change in pieces of literature. One translation may give the reader a completely different idea or feeling based on the author’s biases.

Another study on translations of the Iliad compares several direct and indirect Persian translations (Palladino et al., 2022). This study aligns the translations on word-level to compare exactly how the language compares. Palladino et al. find that there is little consistency in the translations. Even the simplest words are translated in great variety.

Translations

The fact that the Iliad is written in a metre that is impossible to replicate in English makes translating this epic more difficult. There are several different thoughts and philosophies on the art of translating. While Voltaire says that translations weaken the meaning of the original text by reproducing each word (McNiff, 2015), (Bassnett, 2011) argues that translating is a ‘highly skilled and creative activity’.

Translating a text means not just translating each word, but also the meaning and feeling of the text. It follows that every translation is expected to be slightly different, based on the understanding of the original text by translating author and the goal that this author has regarding the translation. One might want to create a translation that lets the reader forget they are reading a translation (Arnold, 1905). The text should feel like it is an original piece of work originally written in this translating language. The focus here lies on creating the same story but make it flow in the translating language. On the other hand, the author might want to focus on keeping every peculiarity of the original text (Arnold, 1905). This way they stay as close to the original as possible. The text might not flow as well as with the other goal, but the original work will not be lost either. These two views are like Goethe's view on translations: one end of translating is to bring the foreign language over into the target language, and the other end is to adapt ourselves to the original language and its peculiarities (Weissbort & Eysteinsson, 2006, p.200).

Following these different views on translating, it seems that the translating author must make a choice before they start writing their translation. Do they want to be more faithful to the target language or to the original work? Neither of these options can be fulfilled entirely so each author must find the balance that fits with their vision.

This balance is especially important to find when working with poetry as these texts include a lot of figurative speech, expressions, and archaic words very specific to the language it is written in (Ehrmanntraut et al., 2022). Translating these texts word for word will most likely not give the translation the same meaning as the original text. Because the original version of the Iliad is written in a meter that is almost impossible to use in the English language, translating authors of this work have to make the choice to either find a rhythmic scheme close to the original that does work with English, try to use the original metre that makes the text very hard to understand for the reader, or not use any rhyme scheme at all and focus on the story rather than the poetry. We can clearly see these choices when looking at different translations of the Iliad as some are in prose while others are in verse with a diversity of rhyme schemes.

Even when authors roughly choose the same balance, the translated texts can still differ a lot. This is because translations contain the interpretation of original text by the translating author. In a text with a lot of expressions and figurative speech, the interpretations can be diverse. Most languages can also not be translated word for word into each other. Not all words may have a direct equivalent or even if they do, the undertone or association can be completely different (Turnbull, 1964). Some words may have several meanings or translations, so the translation completely depends on the interpretation of the translating author.

Each translation is a combination of the original work with the interpretation of the translating author, so analysing the different translations can help us understand the original work better. This way we can get closer to understanding the epic in the way Homer has meant it.

Machine translation

Cultural distinctions and differences in linguistic structures make it difficult for non-native speakers to get a clear understanding of Ancient Greek texts. Through the evolvement of translation studies, several analytical tools and methodologies have emerged that can be used to gain deeper insights into textual structures and features (Yang & Zhou, 2024). Natural Language Processing (NLP) is an example of these methods.

One of the tasks of NLP is machine translation. Machine translation is seen as one of the most difficult tasks in the NLP field (Wang et al., 2022).

Neural machine translation (NMT) models were first mentioned in by (Bahdanau et al., 2015; Sutskever et al., 2014). The goal of these models is to transfer a language into a dense semantic representation and then create a translation by using an attention mechanism (Wang et al., 2022). An advantage of these models is that they do not need any pre-defined rules and features made by the developer. NMT is an end-to-end framework and learns semantic information and relationships from the training corpora. So, the translation knowledge is gathered directly from the texts.

However, the models need to be evaluated and we can only ensure their validity when we have enough knowledge of how translations between the target languages work and what factors can influence the semantics of a text. In ancient languages, automatic translation models are still underdeveloped (Yousef et al., 2022). Especially with Ancient Greek and Latin, only a scarce amount of literature has been digitized. In addition, there is also a lack of aligned datasets or benchmarks for these languages and their translations. These things are needed to improve automatic translation models, either as training data or validation sets.

NLP and semantic textual similarity

Another task of NLP is calculating the similarity between texts. This is used for example in automatic Q&A applications (Wang et al., 2018), or in universities to grade assignments (Hearst, 2000). It is important to differentiate between semantic similarity and semantic relatedness (Budanitsky & Hirst, 2006; Kolb, 2009). Semantic similarity points to lexicon items that are close in meaning. According to (Geffet & Dagan, 2005) two words are semantically similar if you can substitute one for the other in context, like *rose* and *flower*. Semantic relatedness on the other hand holds between lexical items that are connected by any lexical association (Kolb, 2009). Words can be dissimilar but still related, like *flower* and *leaf*.

NLP applications also make a difference between semantic similarity and semantic relatedness measures (Cer et al., 2017). Applications focused on synonym detection require semantic similarity measures while applications focused on finding connections between frequency of words used rely on semantic relatedness measures (Sahlgren & Karlgren, 2008).

Another application where semantic similarity is used is information retrieval, where a similarity score is assigned to a query combined with a corpus (Ali et al., 2018).

Bär et al. (2011) noticed that the same term ‘text similarity’ is used for all these different kinds of applications. The question of how similar texts are is too broad for a simple answer. It completely depends on what properties of the text you are looking at. Texts can use different words and phrases and still be linguistically similar based on lexical and syntactic features (Delmonte, 2022).

Bär et al. (2011) put the different uses of text similarities in three dimensions: structure, style, and content. Structure refers to the order of the sections of a text, style indicates grammar, lexical complexities and mechanics, and content includes the facts and their connection to the text (Attali & Burstein, 2006; Bär et al., 2011).

When considering similarities in the different translations of the Iliad, we want to mainly focus on the dimension Style. As all texts considered here are translations of the same book, we can expect the content to be very similar, if not the same. The differences in this dimension are expected to come from how close the translation is to the original text. The style is expected to vary in the interpretation of the original text by the translating author. The same words and phrases can be translated differently based on the understanding of the original text. The choice between writing in verse or prose will most likely also have some influence on the style and structure dimensions. By adhering to a specific metre, the author will have to make choices to use words and phrases that fit into this scheme while authors that write prose have more freedom. As we consider the entire document embeddings and will not look at specific sections of the texts, we consider the choice of verse or prose as part of the style dimension and not relate it to the structure dimension.

Data

For this research we have gathered 16 different translations of the first book of the Iliad. Due to time and resource constraints, we were not able to use more books or more translations. The first book is where the story starts and sets the precedent for the rest of the story. The first book tells the story about the conflict between Achilles and Agamemnon. The book establishes where Achilles' anger comes from, and the other books are about the consequences of this anger. The first book also sets up the relationship between the gods and humans. We chose this book, because it is the base where the rest of the story build on. It introduces a lot of important characters and core concepts.

Among these 16 translations, 7 are written in prose and 9 in verse. The metres used for the verse translations are iambic pentameter, hexameter, or heptameter. These are rhyme schemes with 5, 6 and 7 metrical feet respectively. The texts are written over four different centuries, with the earliest translation being Chapman's version from 1611 and the latest version written by Kline in 2009.

In Table 1 we can see each translation with information about the author, year of publication, style, metre, and the number of lines of the first book of this translation. We see a clear favouritism towards using pentameters when writing this book in verse, as 7 out of 9 translations in verse are written in a 5-count metre. The number of lines used vary from 443 lines to 839 lines.

Table 1: Author, publication year, genre, metre and number of lines of the 16 translations used in this study

Author	Year	Prose/verse	Metre	No of lines
Chapman	1611	Verse	Iambic heptameter	613
Cowper	1838	Verse	Iambic pentameter	752
Edward, earl of Derby	1865	Verse	Iambic pentameter	717
Buckley	1873	Verse	Iambic pentameter	561
Butler	1898	Prose		504
Pope	1900	Verse	Iambic pentameter	839
Murray	1924	Prose		609
Rieu	1950	Prose		778
Lattimore	1951	Prose		616
Fitzgerald	1974	Verse	Iambic pentameter	754
Fagles	1990	Verse	Combined use of pentameters and hexameters	755
Lombardo	1997	Prose		644
Lang, Leaf & Meyers	2002	Prose		522
Johnston	2006	Verse	Combined use of pentameters and hexameters	688
Merrill	2007	Verse	Hexameters	611
Kline	2009	Prose		443

Data exploration

The dataset that we use in this study does not have a lot of different variables. The reason for this is that the information we are interested in the text itself is. We want to examine patterns within and between the texts, so there is not a lot of exogenous information that is relevant.

One metric that we do want to explore is the number of lines in each translation. In the original version of the Iliad, written in Ancient Greek, there are 606 lines. Translations generally have more lines because they want to clarify more. Especially in poetry, that includes a lot of figures of speech that may not be easily translated into another language, more explanation is necessary to make the work easier to read and understand for the audience.

In Figure 1 we see that approximately half of the translations have more lines than the original, 4 are very close to the number of lines in the original, and 4 have less lines.

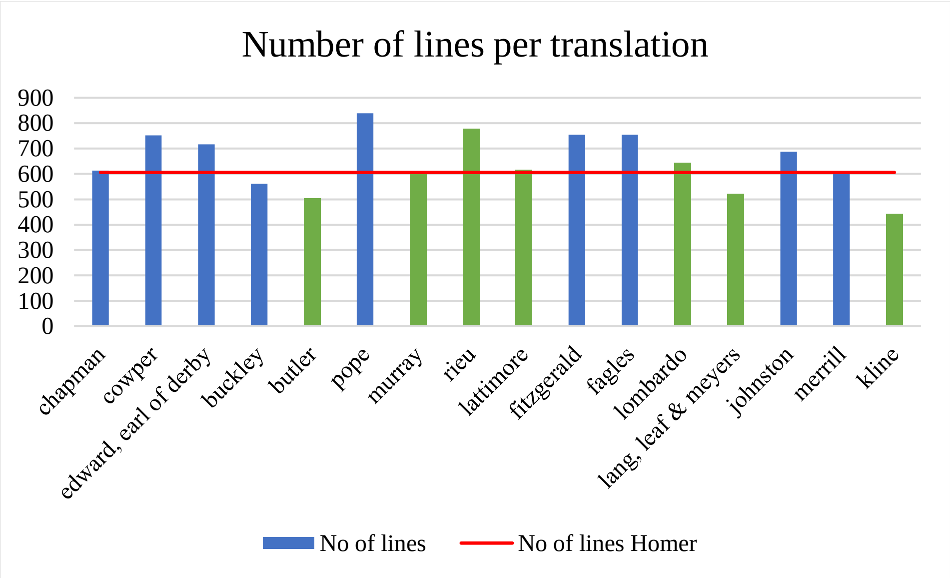

Figure 1: Number of lines in book 1 per translation grouped by prose (green) and verse (blue)

Interestingly, from Figure 1 we see that the translations written in verse (blue) generally have more lines than the translations written in prose (green). The average number of lines in prose translations is 588 while the average number of lines in verse translations is 699. This difference most likely comes from the fact that translations in verse need to adhere to the rhythmic scheme, which only allows for short sentences, while authors writing in prose can make much longer sentences.

To get an idea of how different each translation can be, we compare the very first lines of the book. Each book of the Iliad starts with a line or sentence that summarizes what the book is about. Book 1 starts with a fight between Achilles and Agamemnon and this is the story about the rage of Achilles.

Table 2: The first lines of each of the 16 translations used in this study

Author	Line
Chapman	Achilles’ baneful wrath resound, O Goddess, that impos’d
Cowper	Achilles sing, O Goddess! Peleus’ son;
Edward	Of Peleus’ son, Achilles, sing, O Muse,
Buckley	Sing, O goddess, the destructive wrath of Achilles, son of Peleus,
Butler	Sing, O goddess, the anger of Achilles son of Peleus, that
Pope	Achilles’ wrath, to Greece the direful spring
Murray	The wrath sing, goddess, of Peleus’ son, Achilles,
Rieu	The Wrath of Achilles is my theme, that fatal wrath which, in
Lattimore	SING, goddess, the anger of Peleus’ son Achilleus
Fitzgerald	Anger be now your song, immortal one,
Fagles	Rage-Goddess. sing the rage of Peleus’ son Achilles,
Lombardo	Sing, Goddess, Achilles’ rage,
Lang	Sing, goddess, the wrath of Achilles Peleus’ son, the ruinous wrath
Johnston	Sing, Goddess, sing the rage of Achilles, son
Merrill	Sing now, goddess, the wrath of Achilles the scion of Peleus,
Kline	Goddess, sing me the anger, of Achilles, Peleus’ son,

Table 2 shows the first line of the first book of each of the 16 translations of the Iliad in our dataset. Interestingly, we can see that all sentences approximately say the same, but in a slightly different way. Even for the way of saying Achilles is the son of Peleus, there are several different styles. In addition, some translations mention Achilles’ anger while others use the words wrath or rage. If we compare this using (Bär et al., 2011) three dimensions that were explained in the section *NLP and semantic textual similarity*, we can argue that the context of these lines is generally the same, but the style differs a lot. We cannot say much about structure, as this is just one line, and the structure dimension of this method is about how the full text is structured in specific sections.

In Table 2 we also see two versions of Achilles’ name: Achilles and Achilleus. There are multiple versions of many of the names used in the texts. The most remarkable difference is that some translations use the Roman versions of names, e.g., Jove for Zeus and Juno for Hera. Jupiter is not used for Zeus in any of the translations used in this study.

Methodology

In this section we explain the methods used on the data before we can do the analysis on semantic similarity of the documents.

Pre-processing

Before we can use the texts for our analysis, we must prepare the data. Since most texts come from different sources, there are different formats. Some texts contain figures, footnotes, or line numbers while others are just plain texts. Because we are only interested in analysing the text itself, not the entire document, we need a standardized format for all documents before doing any analysis. Each translation is manually transformed into a plain text file with non-relevant elements removed and only the translated text remaining.

These files are tokenized. This part of the process breaks the texts up into individual words or segments (tokens). The tokens are then all decapitalized and stop words and punctuation marks are removed. Stop words are frequently used words that offer no significance to the meaning of the texts (for example: ‘a’, ‘the’, ‘for’, ‘in’, etc.). Because many of the texts used here are written in old or middle English, both the modern version of these stop words as well as the old or middle English version have been included onto the list. For example, both ‘you’ and ‘thou’ are in the list.

The next step is deciding whether to add lemmatization and stemming. Lemmatization takes the words back to their root form and stemming removes suffixes, prefixes, and grammatical inflections. As we are interested in how the texts itself compare, we will not be doing this as this would remove too much of the author’s choice in words, so it would affect the style dimension that we want to explore.

Term-document matrix

With the pre-processed texts we can take the next step in our analysis process. We can now create a term-document matrix where each document is transformed into a vector of the number of instances of each token. All document vectors together make the matrix. Appendix I shows the term-document matrix for the 20 most frequently used terms in the corpus.

From this matrix we can get information about the frequencies of each token per document, so we can see the most frequently used tokens. In Figure 2 we can see the 20 most frequently used words in the entire corpus.

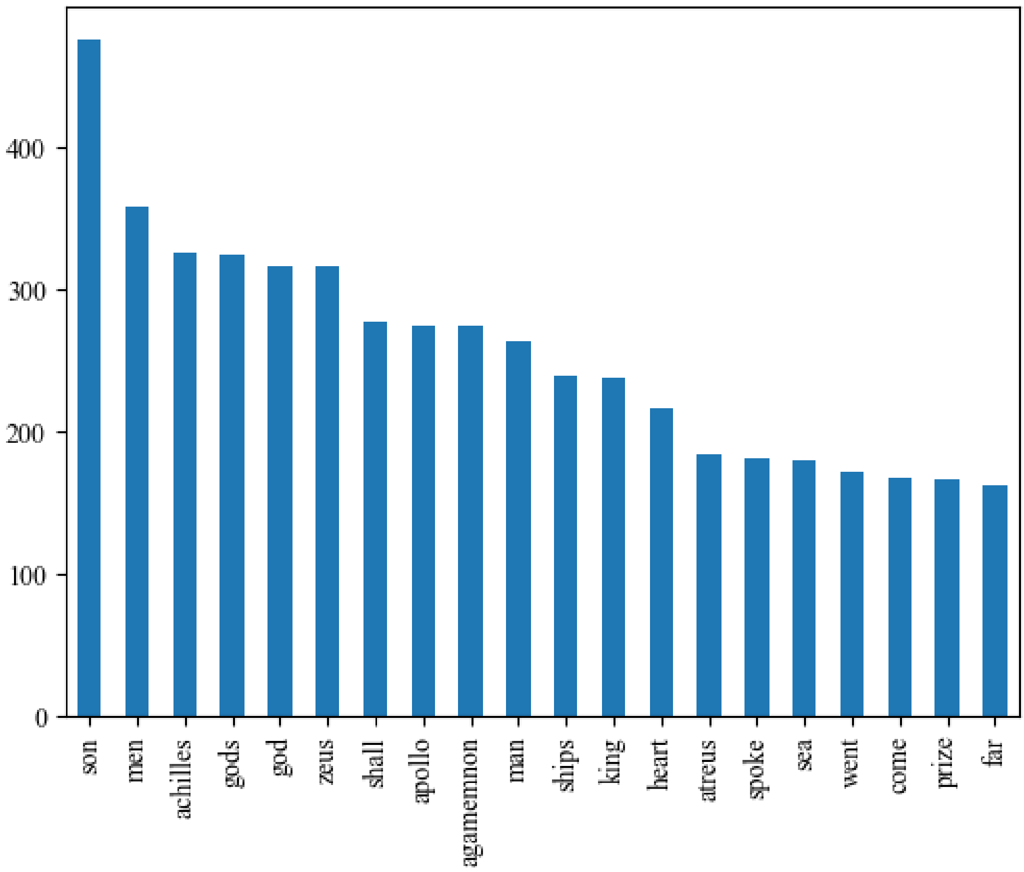

Figure 2: Most frequently used words through all documents

Core concepts are the backbones of the texts. The specific core concept words appear frequently in a text and are what give a piece of text a specific meaning or direction. These words can be difficult to translate accurately if two languages are not well aligned, like Ancient Greek and English. Because the words are frequently used, different translations of them can cause a big difference in semantic similarity measurements.

With this term-document matrix, the cosine similarity can be calculated between all documents. The vector representations of all documents have a high dimensionality, so we first need to scale them back using multidimensionality scaling (MDS). This method will be explained more in the section *Multidimensional Scaling*. Figure 3 shows a scatter plot created from the document vectors. The documents are grouped by century to look at whether the documents show a clear difference regarding the words that are used.

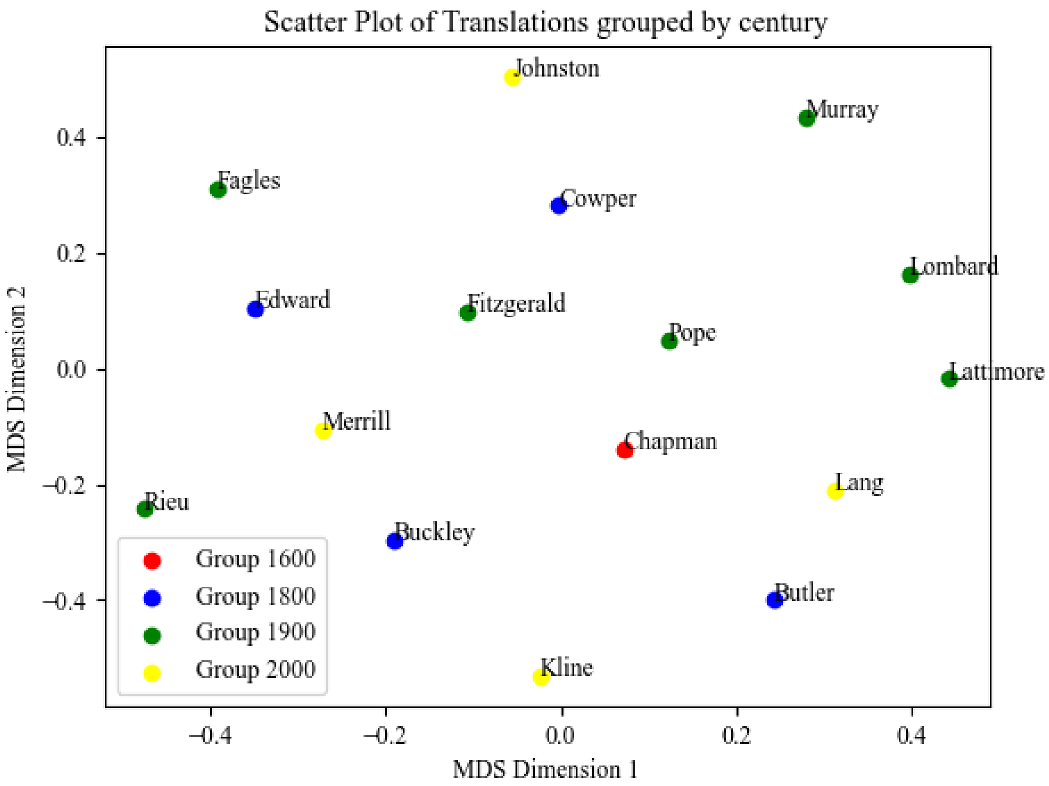

Figure 3: Scatter plot of the translations based on the frequency matrix

The dimensions, represented by the x-axis and y-axis, do not have a direct meaning related to the features of the data. Instead, the distances between the datapoints represent the similarities and dissimilarities as it was in the original high-dimensional data so in this case, the frequency vectors for each document.

The calculated similarity used in this figure is based only on the words used in the documents and the frequencies. There is no context considered with this method. In addition, the high dimensionality of the document vectors from the frequency matrix are a sparse representation of the documents, because all the words in the corpus are included, even if they do not appear in each document. This results in very noisy representations, even when scaling down using MDS.

To look deeper into the texts and to capture semantic relationships between the words, we will use embeddings for further analyses.

Models

Word embeddings

Word embeddings are vector representations of words in a continuous vector space (Zhang et al., 2018). These embeddings are often used as input features in NLP models (Collobert et al., 2011) . Word embeddings can capture similarities between words based on the context. The assumption here is that similar words are used on similar contexts. The vectors that represent word embeddings are first assigned randomly, and then trained with a model based on a certain objective. In this research we consider three different models to train word embeddings for each document: Word2Vec (Ma & Zhang, 2015), GloVe (Pennington et al., 2014; Sakketou & Ampazis, 2020), and BERT (Devlin et al., 2018; Tanaka et al., 2020). We use these three models, because they are among the most widely used pre-trained word embedding models (Chandrasekaran & Mago, 2022)

Word2Vec

Word2Vec is a model that learns word embeddings from raw textual data (Mikolov et al., 2013). It does this by training a neural network on a large corpus to make predictions. There are two main architectures in this model: Continuous-Bag-Of-Words (CBOW) and Skip-gram. CBOW uses the context to predict the target word while Skip-gram predicts the context using the target word (Ma & Zhang, 2015; Mikolov et al., 2013). For this study we will be using the Skip-gram architecture, because this has a better learning ability so is able to capture more semantic information (Mikolov et al., 2013). Since we do not have a very big corpus, we do not have to worry about computation speed (Ma & Zhang, 2015).

The model gives vector representations for each word. By taking the mean value of this vector for each word, we can create a document embedding. These documents embeddings can then be used to get the cosine distance between the documents. The results from this show the semantic similarity between the documents.

GloVe

Just like Word2Vec, GloVe (Global Vectors for Word Representation) is an algorithm that learns dense vector representations of words from raw textual data. However, in this algorithm, the embeddings for the words and the contexts are learned at the same time (Sakketou & Ampazis, 2020). The focus of this model lies on the co-occurrence of the words in the text. The model learns the embeddings in such a way that their dot product closely resembles the logarithm of the ratio of the probability of the co-occurrence of the words (Pennington et al., 2014).

The embeddings created by the GloVe algorithm can be used in the same way to compute document embeddings and the cosine distance between documents as the Word2Vec embeddings.

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model that only needs finetuning with an additional output layer to create a model fit for numerous NLP tasks (Devlin et al., 2018; Valli Mayil & Ratha Jeyalakshmi, 2023). As the name of the model indicates, BERT works bi-directional which is different than Word2Vec and GloVe. This means that it considers context both from left to right and from right to left.

The algorithm consists of two steps: pre-training and fine-tuning. During the first step, two tasks are performed at the same time. In the first task, the model uses a Masked Language Model (MLM) to predict the missing words in a masked sentence. This is done on a big, unlabelled dataset. The masking prevents the algorithm from seeing all the words during training (Devlin et al., 2018). Both the left-side context as well as the right-side context is considered, as it is a bidirectional algorithm. In this study, the base BERT uncased model is used. This model is pre-trained on BooksCorpus (Zhu et al., 2015) which has 800 million words, and English Wikipedia, which has 2500 million words. It has 12 hidden layers and 110 million parameters (Devlin et al., 2018).

The second task is Next Sentence Prediction (NSP). In this task, the model is a binary classifier that labels sentence pairs on whether the second sentence is predicted to come after the first (Devlin et al., 2018). The goal here is for the model to learn relationships between two sequences as this is an important part for numerous NLP tasks, like question answering.

The fine-tuning step is where the model updates the parameters on a smaller data set that is specific to the task that the model is meant to be used for. This is done by backpropagation through all the layers of the network. This task requires a large amount of labelled data and computational power. If either of these is not available, the alternative is to take the feature-based approach for this step. With the feature-based approach, the backpropagation is only going through the layers of the network that are added for the specific task (Devlin et al., 2018). Since we do not have a large amount of data for this study, we will take the feature-based approach instead of fine-tuning the model.

When a BERT model receives a document, it treats it as if it is one sentence. The algorithm puts a special token [CLS] in the first place of the input string. The embedding of this token is regarded as the embedding of the sentence. So, if the input is a document, the token [CLS] can be seen as the document embedding (Tanaka et al., 2020). For each of our translations, we use this token as our document embeddings.

One drawback of the original BERT model that we are using is that it can only take 512 tokens (Devlin et al., 2018). This limitation comes from the fact that the algorithm uses self-attention mechanisms to model the dependencies between the tokens in the input. These mechanisms are computationally very complex, and the longer the input sequence is, the more complex it becomes (Tanaka et al., 2020). The texts used in this study are longer than 512 tokens, so we have to cut them up into batches to use BERT with the full texts.

Multidimensional scaling

After getting the document embeddings from the models, we calculate the pairwise distance between each of the documents using cosine similarity, the cosine of the angle between the vectors. The cosine distance can then be calculated as 1 – cosine similarity (Senoussaoui et al., 2014).

The vector representations that we then get are still high-dimensional. To visualize and identify relationships between the translations, we use multidimensional scaling (MDS) to project these vectors onto a lower dimensional scale.

Lowering the dimension of vectors will always lead to some information loss (Li, 1991). Since this study is focused on the similarity between the documents, we want to keep as much information about the relative distance between the embeddings as possible. MDS is a dimensionality scaling method that preserves the pairwise distance by minimizing the differences between the distances in the original space and the lower dimensional space (Cox & Cox, 2000). The vector representations of the pairwise distances are fit to the model, which projects them on a pre-determined dimensional scale. Because we want to plot the distances, we project the vectors on a 2-dimensional scale. In the result, the dimensions have no direct meaning related to the data. The distance between the datapoints represent the similarities and dissimilarities that were captured by the models that created the word embeddings.

Hierarchical clustering

As mentioned before in the section *NLP and textual semantic similarity*, to compare texts, you need to know what exactly you are comparing. Text similarity is used for many different applications. In our study, we want to compare texts based on several features to find out which ones have a big influence on the calculated textual semantic similarity based on the document embeddings. Hierarchical clustering is a method that can be used to visualise document similarity. It provides a hierarchical view of document similarity (Zhao & Karypis, 2005). This method can be used with agglomerative algorithms, so it does not need a pre-set number of clusters. The number of clusters are based on the similarities of the documents. The agglomerative algorithm takes a bottom-up approach. It starts with assigning each document to its own cluster and then based on the similarities merges documents until they all belong to one cluster. This means that the first merges are between documents that are closest together, while the latest merges are between documents that are the least similar. The outcome is visualised in a dendrogram, making the result easy to interpret. The algorithm uses a linkage method to calculate the distance between two clusters (SciPy, 2024). The method we use here uses the average similarity value of the clusters.

Results

During the data exploration, we created document vectors and plotted these using MDS. From the plot, Figure 3, it seems that translations written in the 21st century are less similar to each other than translations written in the 20th century. However, there does not seem to be a clear difference in similarity between the centuries. All these time periods have translations scattered around. There are no clear clusters in this plot. As we previously mentioned, this method does not take context into account, only the words used and their frequency.

To take context into account, we created document embeddings with the three algorithms GloVe, Word2Vec, and BERT. For each algorithm, we created a similarity matrix that shows the pairwise similarity scores between each of the translations. Figures 4, 5, and 6 show the similarity matrices of the GloVe, Word2Vec, and BERT model respectively.

In these figures we see that they all see Chapman's translation as less similar to the rest compared to others. The GloVe and Word2Vec matrices also show Pope's translation to have a low similarity score compared to the rest, while BERT shows Pope having average similarity scores. However, both the Word2Vec and BERT models give Pope's translation a high similarity score with the translations of Chapman, Cowper, and Edward. Word2Vec also gives extremely high similarity scores to the translations by Buckley, Butler, Johnston, Kline, Lang, Lattimore, Merrill, Murray, and Rieu. So, this model finds that more than half of the translations are extremely similar.

To easily see how the distances between the translations compare to each other, we plot the document embeddings from the three models in scatter plots. Because the document embeddings are high-dimensional vectors, we first use MDS to project the vectors onto a 2-dimensional scale. Figure 7 shows the three scatter plots, created from the document embeddings from the GloVe, Word2Vec, and BERT models. The three plots look very different from each other. Important to note is the scale of the plots that are all different. The scale of the plot from the BERT model is 10 times bigger than that of the Word2Vec model, while the scale of the plot from the GloVe model is somewhere in between. This has no influence on comparing the distances within the plot itself, but when comparing them to the other plots it needs to be taken into account.

We see that all models plot Chapman's translation quite far away from most the other translations. This was expected based on the similarity matrices. Interestingly, while Word2Vec and BERT plot Pope's translation as one of the closest translations to Chapman's version, GloVe puts is as one of the least close translations.

	Buckley	Butler	Chapman	Cowper	Edward	Fagles	Fitzgerald	Johnston	Kline	Lang	Lattimore	Lombard	Merrill	Murray	Pope	Rieu
Buckley	1.000000	0.992938	0.980512	0.992836	0.987228	0.981177	0.984503	0.988850	0.992300	0.990519	0.989879	0.985682	0.994306	0.994018	0.984086	0.990126
Butler	0.992938	1.000000	0.982266	0.991043	0.988515	0.988507	0.991814	0.994200	0.995130	0.990498	0.993426	0.992310	0.993542	0.996687	0.980135	0.994664
Chapman	0.980512	0.982266	1.000000	0.989993	0.994207	0.979760	0.986511	0.982132	0.981950	0.981341	0.985908	0.982873	0.983210	0.983098	0.986604	0.987226
Cowper	0.992836	0.991043	0.989993	1.000000	0.993062	0.982546	0.986502	0.985512	0.989122	0.990871	0.987946	0.983479	0.990654	0.991086	0.993604	0.989212
Edward	0.987228	0.988515	0.994207	0.993062	1.000000	0.981129	0.985775	0.986852	0.989395	0.990223	0.987574	0.983389	0.987346	0.989815	0.990133	0.988893
Fagles	0.981177	0.988507	0.979760	0.982546	0.981129	1.000000	0.995505	0.993622	0.990902	0.980795	0.990864	0.993434	0.989421	0.990423	0.980526	0.993059
Fitzgerald	0.984503	0.991814	0.986511	0.986502	0.985775	0.995505	1.000000	0.993951	0.991063	0.982626	0.995095	0.996005	0.992530	0.992251	0.981372	0.996893
Johnston	0.988850	0.994200	0.982132	0.985512	0.986852	0.993622	0.993951	1.000000	0.997539	0.985747	0.993968	0.996270	0.993660	0.995745	0.978702	0.996524
Kline	0.992300	0.995130	0.981950	0.989122	0.989395	0.990902	0.991063	0.997539	1.000000	0.990145	0.993680	0.992215	0.994678	0.997645	0.983070	0.995807
Lang	0.990519	0.990498	0.981341	0.990871	0.990223	0.980795	0.982626	0.985747	0.990145	1.000000	0.987132	0.982161	0.987570	0.991369	0.985366	0.985852
Lattimore	0.989879	0.993426	0.985908	0.987946	0.987574	0.990864	0.995095	0.993968	0.993680	0.987132	1.000000	0.993160	0.996764	0.996021	0.981986	0.996014
Lombard	0.985682	0.992310	0.982873	0.983479	0.983389	0.993434	0.996005	0.996270	0.992215	0.982161	0.993160	1.000000	0.992272	0.992346	0.975809	0.995768
Merrill	0.994306	0.993542	0.983210	0.990654	0.987346	0.989421	0.992530	0.993660	0.994678	0.987570	0.996764	0.992272	1.000000	0.996874	0.983751	0.994515
Murray	0.994018	0.996687	0.983098	0.991086	0.989815	0.990423	0.992251	0.995745	0.997645	0.991369	0.996021	0.992346	0.996874	1.000000	0.983803	0.995164
Pope	0.984086	0.980135	0.986604	0.993604	0.990133	0.980526	0.981372	0.978702	0.983070	0.985366	0.981986	0.975809	0.983751	0.983803	1.000000	0.982179
Rieu	0.990126	0.994664	0.987226	0.989212	0.988893	0.993059	0.996893	0.996524	0.995807	0.985852	0.996014	0.995768	0.994515	0.995164	0.982179	1.000000

Figure 4: Similarity matrix GloVe

	Buckley	Butler	Chapman	Cowper	Edward	Fagles	Fitzgerald	Johnston	Kline	Lang	Lattimore	Lombard	Merrill	Murray	Pope	Rieu
Buckley	1.000000	0.999640	0.998613	0.999513	0.999325	0.999556	0.999421	0.999731	0.999782	0.999938	0.999749	0.999494	0.999911	0.999677	0.998372	0.999724
Butler	0.999640	1.000000	0.997561	0.999040	0.998455	0.999552	0.999174	0.999895	0.999595	0.999763	0.999840	0.999909	0.999820	0.999828	0.997377	0.999749
Chapman	0.998613	0.997561	1.000000	0.999586	0.999821	0.998770	0.999057	0.997641	0.997697	0.998639	0.997478	0.997779	0.998136	0.997229	0.999903	0.998225
Cowper	0.999513	0.999040	0.999586	1.000000	0.999850	0.999545	0.999511	0.998953	0.998871	0.999574	0.998844	0.999129	0.999289	0.998680	0.999551	0.999262
Edward	0.999325	0.998455	0.999821	0.999850	1.000000	0.999186	0.999320	0.998516	0.998646	0.999293	0.998418	0.998509	0.998937	0.998227	0.999726	0.998917
Fagles	0.999556	0.999552	0.998770	0.999545	0.999186	1.000000	0.999920	0.999625	0.999386	0.999785	0.999482	0.999752	0.999657	0.999404	0.998496	0.999855
Fitzgerald	0.999421	0.999174	0.999057	0.999511	0.999320	0.999920	1.000000	0.999370	0.999220	0.999636	0.999223	0.999432	0.999445	0.999114	0.998721	0.999719
Johnston	0.999731	0.999895	0.997641	0.998953	0.998516	0.999625	0.999370	1.000000	0.999838	0.999832	0.999964	0.999833	0.999909	0.999951	0.997322	0.999894
Kline	0.999782	0.999595	0.997697	0.998871	0.998646	0.999386	0.999220	0.999838	1.000000	0.999771	0.999886	0.999441	0.999884	0.999877	0.997304	0.999788
Lang	0.999938	0.999763	0.998639	0.999574	0.999293	0.999785	0.999636	0.999832	0.999771	1.000000	0.999800	0.999720	0.999944	0.999735	0.998396	0.999872
Lattimore	0.999749	0.999840	0.997478	0.998844	0.998418	0.999482	0.999223	0.999964	0.999886	0.999800	1.000000	0.999726	0.999923	0.999976	0.997149	0.999835
Lombard	0.999494	0.999909	0.997779	0.999129	0.998509	0.999752	0.999432	0.999833	0.999441	0.999720	0.999726	1.000000	0.999724	0.999678	0.997596	0.999816
Merrill	0.999911	0.999820	0.998136	0.999289	0.998937	0.999657	0.999445	0.999909	0.999884	0.999944	0.999923	0.999724	1.000000	0.999887	0.997843	0.999868
Murray	0.999677	0.999828	0.997229	0.998680	0.998227	0.999404	0.999114	0.999951	0.999877	0.999735	0.999976	0.999678	0.999887	1.000000	0.996858	0.999781
Pope	0.998372	0.997377	0.999903	0.999551	0.999726	0.998496	0.998721	0.997322	0.997304	0.998396	0.997149	0.997596	0.997843	0.996858	1.000000	0.997903
Rieu	0.999724	0.999749	0.998225	0.999262	0.998917	0.999855	0.999719	0.999894	0.999788	0.999872	0.999835	0.999816	0.999868	0.999781	0.997903	1.000000

Figure 5: Similarity matrix Word2Vec

	Buckley	Butler	Chapman	Cowper	Edward	Fagles	Fitzgerald	Johnston	Kline	Lang	Lattimore	Lombard	Merrill	Murray	Pope	Rieu
Buckley	1.000000	0.984322	0.947217	0.977995	0.956234	0.944747	0.962025	0.937349	0.978569	0.986747	0.977566	0.924504	0.976866	0.988889	0.956819	0.949679
Butler	0.984322	1.000000	0.950799	0.970836	0.958125	0.956947	0.976494	0.962881	0.989857	0.978321	0.986638	0.950037	0.977281	0.983489	0.957040	0.976270
Chapman	0.947217	0.950799	1.000000	0.981069	0.995669	0.915450	0.948467	0.934272	0.953726	0.955613	0.953954	0.910001	0.929276	0.926147	0.989931	0.943639
Cowper	0.977995	0.970836	0.981069	1.000000	0.986112	0.943599	0.964919	0.942745	0.972030	0.986271	0.974947	0.920587	0.964451	0.966384	0.988851	0.948195
Edward	0.956234	0.958125	0.995669	0.986112	1.000000	0.929831	0.955288	0.942152	0.963780	0.964150	0.960651	0.919381	0.938842	0.938804	0.994841	0.949391
Fagles	0.944747	0.956947	0.915450	0.943599	0.929831	1.000000	0.984680	0.978691	0.977081	0.945664	0.967426	0.980283	0.971765	0.945792	0.940163	0.972353
Fitzgerald	0.962025	0.976494	0.948467	0.964919	0.955288	0.984680	1.000000	0.981649	0.987016	0.964626	0.985256	0.976815	0.981683	0.960342	0.961028	0.985212
Johnston	0.937349	0.962881	0.934272	0.942745	0.942152	0.978691	0.981649	1.000000	0.980894	0.935012	0.965783	0.988318	0.956907	0.933363	0.945631	0.989099
Kline	0.978569	0.989857	0.953726	0.972030	0.963780	0.977081	0.987016	0.980894	1.000000	0.973525	0.986327	0.969613	0.978672	0.976725	0.966962	0.986406
Lang	0.986747	0.978321	0.955613	0.986271	0.964150	0.945664	0.964626	0.935012	0.973525	1.000000	0.983704	0.918598	0.980230	0.985972	0.965137	0.942018
Lattimore	0.977566	0.986638	0.953954	0.974947	0.960651	0.967426	0.985256	0.965783	0.986327	0.983704	1.000000	0.953089	0.986644	0.981887	0.963353	0.972003
Lombard	0.924504	0.950037	0.910001	0.920587	0.919381	0.980283	0.976815	0.988318	0.969613	0.918598	0.953089	1.000000	0.951108	0.921010	0.922931	0.983265
Merrill	0.976866	0.977281	0.929276	0.964451	0.938842	0.971765	0.981683	0.956907	0.978672	0.980230	0.986644	0.951108	1.000000	0.979882	0.944950	0.959731
Murray	0.988889	0.983489	0.926147	0.966384	0.938804	0.945792	0.960342	0.933363	0.976725	0.985972	0.981887	0.921010	0.979882	1.000000	0.941415	0.943045
Pope	0.956819	0.957040	0.989931	0.988851	0.994841	0.940163	0.961028	0.945631	0.966962	0.965137	0.963353	0.922931	0.944950	0.941415	1.000000	0.951096
Rieu	0.949679	0.976270	0.943639	0.948195	0.949391	0.972353	0.985212	0.989099	0.986406	0.942018	0.972003	0.983265	0.959731	0.943045	0.951096	1.000000

Figure 6: Similarity matrix BERT

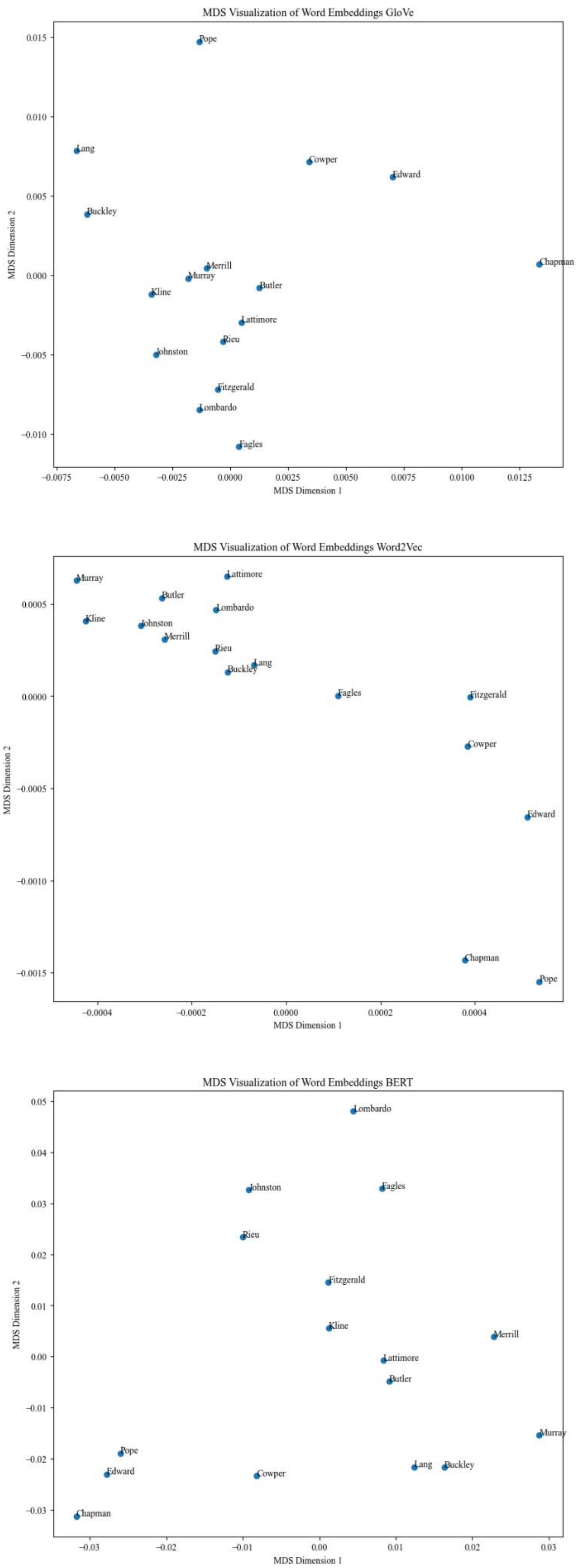

Figure 7: Scatter plot of GloVe, Word2Vec, and BERT embeddings

Figure 7 also shows the grouping we saw in the similarity matrix of Word2Vec where more than half of the translations are positioned very close together while the rest of the translations is spread out further. An interesting discovery is that we see the same grouping in the GloVe scatter plot. Even though it was less visible in this model's similarity matrix, the same translations are positioned quite close together in a group in the bottom half of the distance plot.

Another notable thing is that all three plots show the translations of Buckley and Lang to be quite similar. They are not necessarily grouped together, but in all three plots the distance between them is closer than their distance to any other translation.

What features these translations include that make them so similar is what we want to discover by looking specifically at the Style dimension.

Style

As explained before, the dimension Style refers to grammar, lexical complexities, and mechanics of the text (Bär et al., 2011). To explore this dimension, we will be looking at two features of the texts: year of publication, and the use of the Roman or Greek names.

From the document-frequency plot (Figure 3) we found that there is no clear separation between the publication years. Now that we take context into account using the document embeddings, we are interested to see if there is a difference. Figure 8 shows the document vectors plotted using MDS, grouped by the century of the publication year for each model.

In the plots we see that the groupings we found in the GloVe and Word2Vec plots before, mostly include translations from the 20^{th} and 21^{st} century. In the plot from the BERT model, we only see a slight pattern with the translations form the 19^{th} century, all positioned on the bottom half of the plot. The model seems to find the translations within this century more similar than translations within another century. However, the BERT model does not make a clear separation between different centuries.

In Appendix II.A. we see tables with the average and maximum pairwise distances per century for each of the models, calculated from the similarity matrices. From these tables we can see that all models find that the translations from the 20^{th} century are the least similar to each other. Both GloVe and BERT find that the average pairwise distances within the 19^{th} and 20^{th} century are very close to each other, meaning translations within the 19^{th} century are on average as similar to each other as translations within the 20^{th} are to each other. These two models also find that the maximum pairwise distances in these centuries are between the same translations. So even though the similarity values are measured on another scale, these models seem to give very similar results.

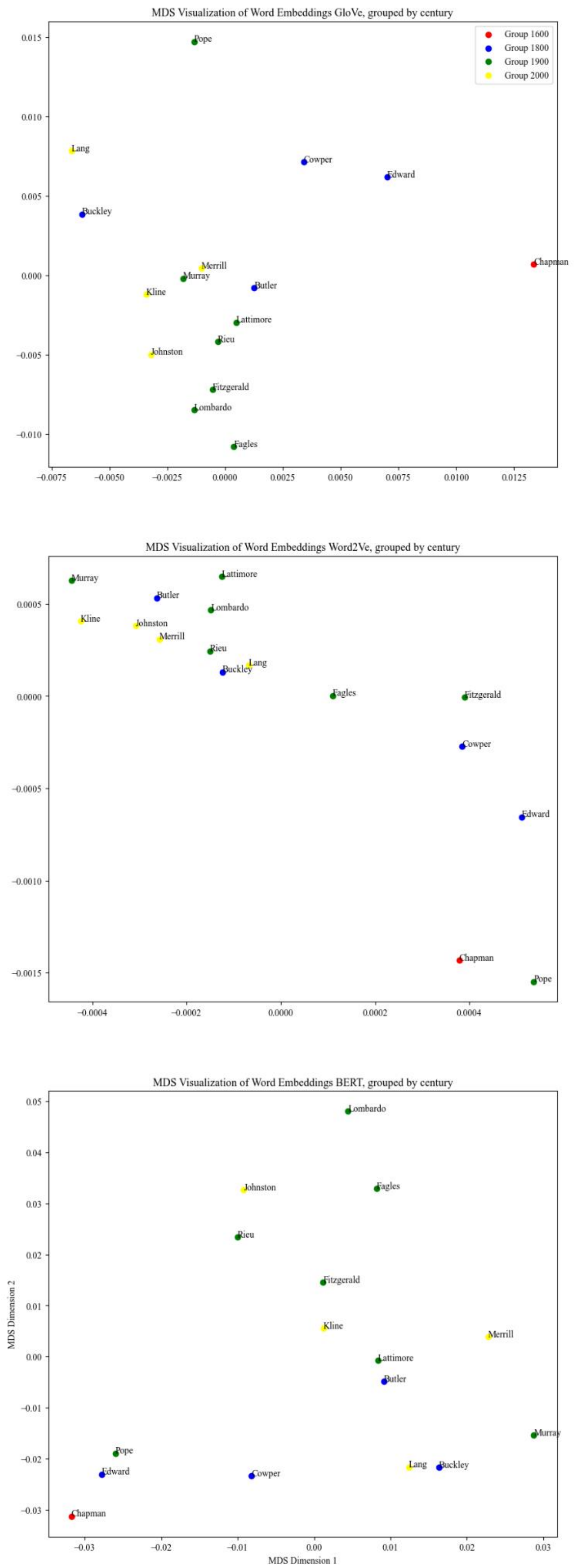

Figure 8: Scatter plot of GloVe, Word2Vec, and BERT embeddings, grouped by century

While exploring the data, we found that not all translations use the same names for the characters. The biggest difference was that some translations use the Roman names (i.e., Jove, Juno, Ulysses) while other translations use the Greek names (i.e., Zeus, Hera, Odysseus). We want to explore how the translations are grouped based on which names they use. In Figure 9 we see the how the Roman and Greek names are divided.

Even though there are only 6 translations that use the Roman names versus 10 that use the Greek version, it does seem like there is a division between which version of the names are used in all three plots. Although, the division seems clearer in the plots from the GloVe and BERT model than the one from the Word2Vec model.

The two big groups we see in the GloVe and Word2Vec plots are mostly consisting of translations using the Greek version of the names. However, the GloVe and Word2Vec plots do include 1 and 2 translations using the Roman version of the names in this group, respectively. In all the plots, it looks like the translations using the Greek version of the names are situated closed together than the translations using the Roman version, meaning these are seen as more similar. We can even see that the translations using the Roman version of the names are positioned more on the outside of the plots with a larger distance between them, leaving us to believe that these are not only seen as being less similar to the translations using the Greek names, but also that they are not even seen as very similar to each other.

The next feature we want to look at is the genre of the texts. Here the texts are either written in prose or verse. Prose is a text without a set rhythm while verse does have a rhythm. Texts written in verse usually have shorter sentences or lines, and sometimes rhyme but they do not have to. Blank verse is a form of unrhymed poetry. There are two translations in our dataset written in blank verse: Fitzgerald's version and Fagles' version.

Figure 10 shows the scatter plots of the translations, grouped by the genre. Other than the previous feature, these plots do not seem to give such a clear separation. Especially the two big groupings in the GloVe and Word2Vec plots contain both translations written in prose and verse. Interestingly, the translations written in verse have a much bigger distance between each other than the translations written in prose. This is quite unexpected, since this means that they are the least similar to each other. We would expect the distances between the prose translations to be much bigger than the distances between the verse translations, as writing in prose leaves much more freedom because authors do not have to keep to a specific rhythm. In the tables in Appendix II.B. we can see that all models indeed find that translations in verse have a smaller pairwise distance on average. However, this difference is much bigger with the GloVe and Word2Vec models, where the prose translations are two and three times further apart, respectively, than the verse translations. The BERT model only shows a slight difference.

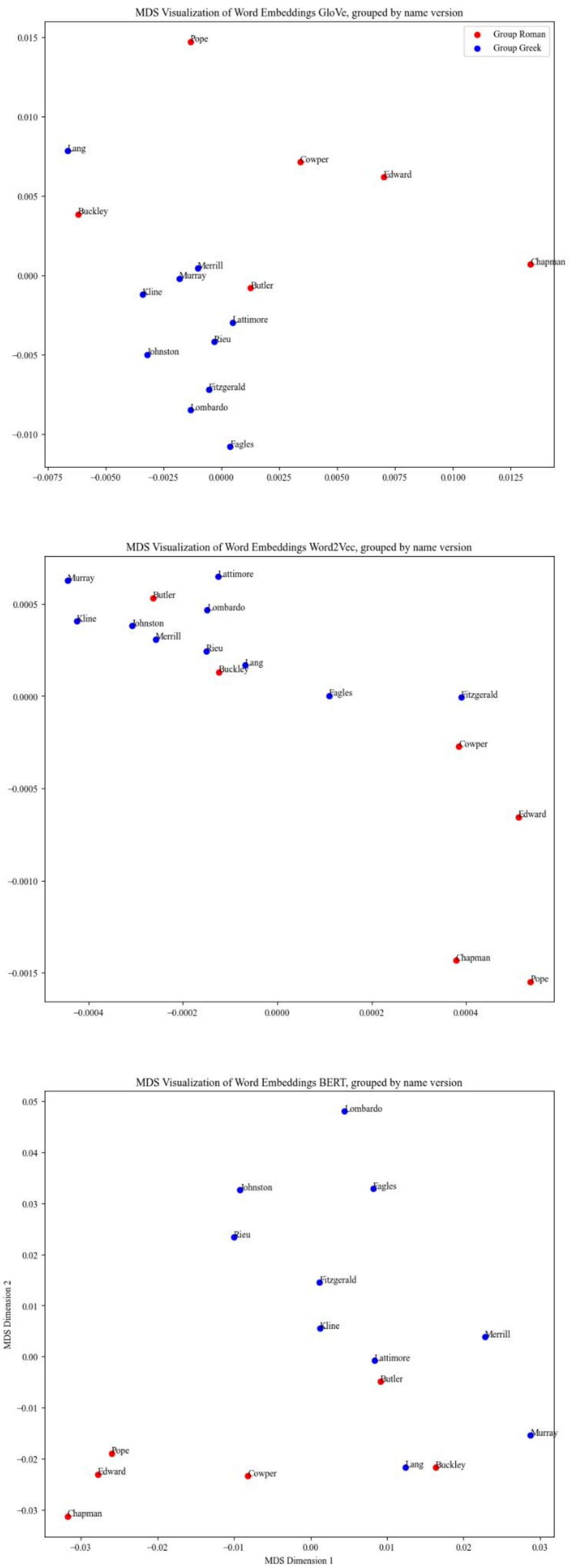

Figure 9: Scatter plot of GloVe, Word2Vec, and BERT embeddings, grouped by name version

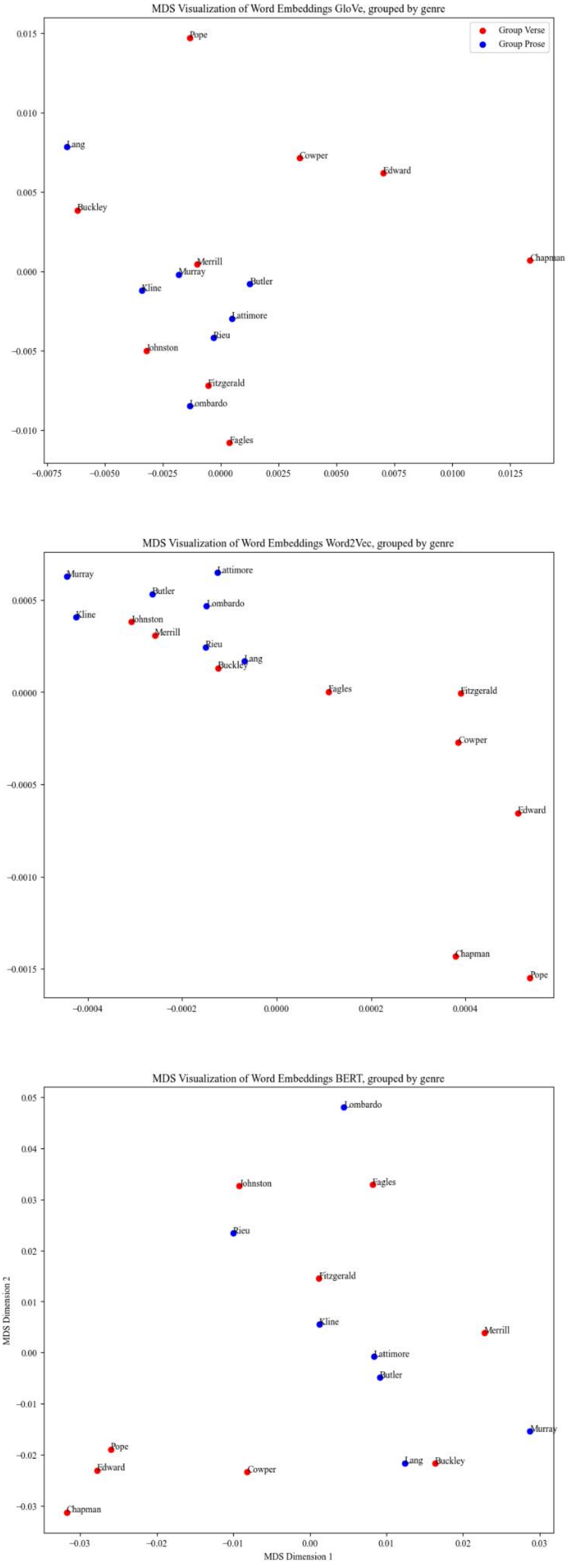

Figure 10: Scatter plot of GloVe, Word2Vec, and BERT embeddings, grouped by genre

Hierarchical Clustering

In the previous part we have looked at how the translations are grouped based on the individual features. We now want to see how the translations are grouped based on all the features together. It is possible that we think a feature has a big influence on the similarity score, while it is more related to another feature. In addition, there are more features and dimensions than what we have considered so there are most likely more influences on the similarity score than what we have found.

To group the documents based on all the features, we use the hierarchical clustering technique. With this method we can cluster the translations without a pre-determined number of clusters. Using a dendrogram we can then see how the translations are grouped together based on the similarity scores.

Figure 11 shows the dendrograms for each model. From this figure we see that all three models have created completely different dendrograms. The Word2Vec model found two clusters at the outer level, while the GloVe and BERT models both find three clusters. However, these three clusters contain very different translations.

Very clear are the two big groups we saw in both the distance matrices and the scatter plots of the GloVe and Word2Vec models. In Figure 11 we see them as the orange clusters of the dendrograms of these two models. They are constructed differently but end up with almost the same translations. The only difference is that the Word2Vec model includes the translations of Buckley and Lang in this cluster, while the GloVe model puts these in a cluster with Cowper's and Edward's translations. Another interesting finding from these two plots is that the Word2Vec model fuses Chapman's and Pope's translation into a cluster on a relatively low level (~0.001 from 0.006), while the GloVe model only fuses them at one of the latest levels (~0.023 from 0.035). The Word2Vec model even sees these two as the most similar within their cluster that also contains the translations by Buckley and Lang. The GloVe model puts them in their own cluster, showing that this model sees a relatively low similarity with the other translations. This was also visible in the scatter plot and the similarity matrix.

The biggest cluster in the dendrogram from the BERT model does include some of the same translations as the big groups in the other two dendrograms. The difference is that the translations by Lombardo, Fagles, Johnston and Rieu are put in their own cluster. However, in this dendrogram we see the translations by Cowper, Chapman, Edward, and Pope in their own cluster again.

Tables 3, 4, and 5 show the information from the translations for each cluster from the BERT dendrogram. The first thing that we immediately notice is that the publication year does not seem to be an important factor in making the clusters. All three clusters include translations from different centuries. We can also see that all the translations in Table 4 (green cluster) are written in verse with an iambic meter and use the Roman names. However, the translation by Buckley is not included in this.

Table 3: Cluster information BERT cluster orange

Author	Year	Prose/verse	Metre	Roman/Greek
Rieu	1950	Prose		Greek
Fagles	1990	Verse	Combined use of pentameters and hexameters	Greek
Lombardo	1997	Prose		Greek
Johnston	2006	Verse	Combined use of pentameters and hexameters	Greek

Table 4: Cluster information BERT cluster green

Author	Year	Prose/verse	Metre	Roman/Greek
Chapman	1611	Verse	Iambic heptameter	Roman
Cowper	1838	Verse	Iambic pentameter	Roman
Edward, earl of derby	1865	Verse	Iambic pentameter	Roman
Pope	1900	Verse	Iambic pentameter	Roman

Table 5: Cluster information BERT cluster red

Author	Year	Prose/verse	Metre	Roman/Greek
Merrill	2007	Verse	Hexameters	Greek
Murray	1924	Prose		Greek
Fitzgerald	1974	Verse	Iambic pentameter	Greek
Kline	2009	Prose		Greek
Lattimore	1951	Prose		Greek
Butler	1898	Prose		Roman
Lang, Leaf & Meyers	2002	Prose		Greek
Buckley	1873	Verse	Iambic pentameter	Roman

for some reason. So, there must be one or more other features that we have not looked at that make this translation seem different by the model. In all three dendrograms we also saw that the translations by Buckley and Lang were considered more similar than they are to the other translations. However, looking at the information in these tables, they seem to have nothing in common other than the fact that they are translations from the same text.

Appendix III A. and B. show the tables corresponding to the dendrograms of the GloVe and Word2Vec models respectively. The Word2Vec model dendrogram leaves us with the same question about the translation by Buckley as the BERT dendrogram does, since here again it is not grouped with the translations by Chapman, Pope, Cowper, and Edward. The tables from the GloVe dendrogram leave us with even more questions, as there does not seem to be a clear reason why the translations are clustered this way when looking at the features.

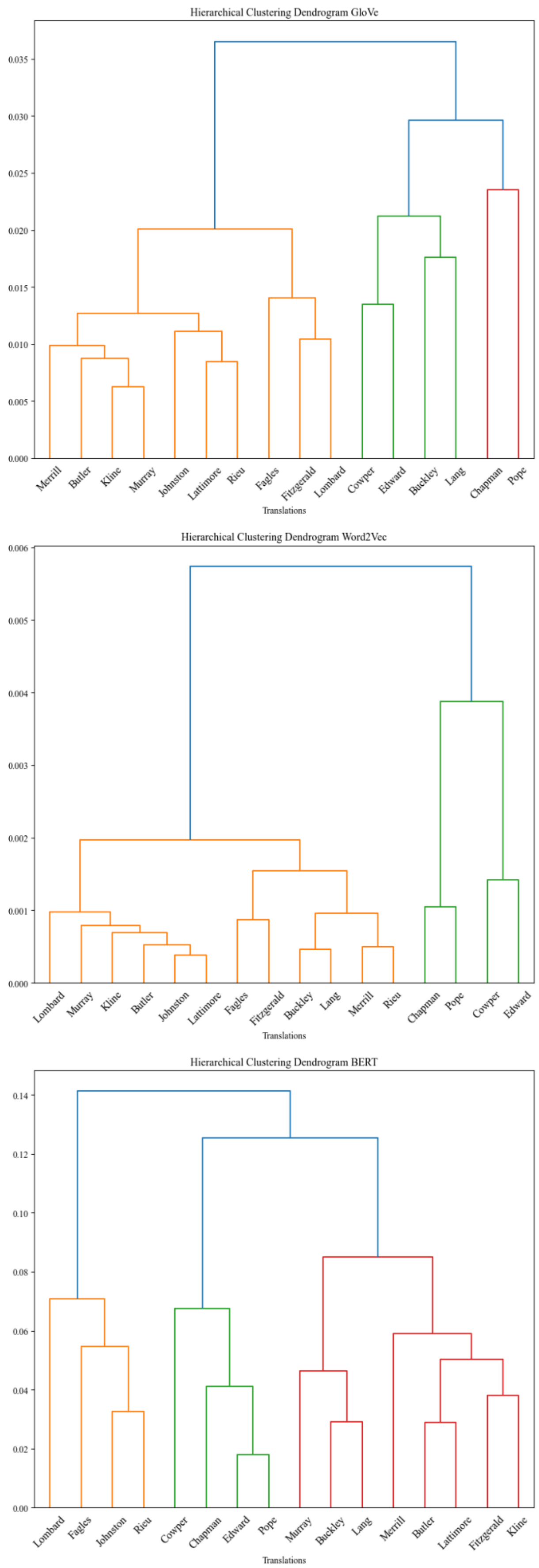

Figure 11: Dendrograms for each model

Discussion

Machine translation is an important NLP task that is used in several applications. One of the biggest fields of machine translation application is language and culture studies. But this application works both ways. With machine translations, we can learn about languages and cultures faster, but with more knowledge about languages and culture we can also improve our machine translation tools.

One of the gaps in the current machine translation tools is unspoken languages. From these languages there is not a lot of data, especially digitalized data, so training algorithms to create well working machine translation tools is hard. One of these languages is Ancient Greek.

This study contributes to the field of NLP by studying whether the features ‘publication year’, ‘name version (Roman/Greek)’, and ‘genre (prose/verse)’ have an influence on the similarity score between English translations of the first book of the Ancient Greek epic poem the Iliad. Studying these translations helps us understand the relation between the two languages. Knowing which features influence a translation will help us create better machine translation tools.

Through comparing document embeddings of each translation in our dataset, we have analysed how the different features seem to influence the similarity scores. The findings of this paper suggest the following:

1. The publication year of the translation seem to have little influence on the similarity score. There is no clear difference between the language used for these translations throughout the past three centuries.

2. Both the choice of using Roman or Greek names, and the choice of writing either Prose or Verse seem to have an influence on the similarity score, especially taken together. This suggests that names are an important part of this work. They are most likely related to core concepts within the epic. Translations written in verse with the use of the Roman names are seen as most similar by all three models used in this study. This also confirms the idea that more freedom in the writing process causes more variety in translations. Authors of translations written in verse are stuck with a rhythmic scheme so the vocabulary they can use is much smaller than authors that write the translation in prose.

However, there are features that we have not considered that also have a big influence on the similarity score. From the hierarchical clustering process, we found that the combination of names, genre and rhythmic scheme seem to have a big influence on the clustering, but there must be other features that have a bigger influence, because of how Buckley’s translation is clustered. In addition, the scatter plots showed that all 3 models find the translations written in prose more similar to each other than the translations written in verse, but the dendrograms from the Agglomerative clustering algorithm have grouped the translations written in verse more together than the translations written in prose. This gives a controversial result and leaves us with the question: do translations written in prose actually differ more from each other than translations written in verse or not? More research needs to be done into other features to get a more complete picture.

Obviously, the words that the authors use influence the similarity score a lot, since they are based on the document embeddings that are calculated from the word embeddings of each text. More research into the exact words used needs to be done to say more specific things about this. But this needs to be done by, or in collaboration with, a domain expert.

Using more data will also give a clearer picture of what way these features influence the similarity scores. The Iliad consists of 24 books, but due to time constraints this study was only able to include the first book in the dataset. There also exist more English translations than the 16 used in this study, but the limited availability of these resources and the time restraint did not make it possible to include more in this study.

Conclusion

Using machine learning techniques to analyse the variation in translations of the first book of the Iliad is a step in getting to a clear understanding of important features of translations of Ancient Greek works. Before one can write a translation, one must make several choices about certain linguistical features.

This study used 16 English translations of the first book of the Iliad to investigate the influence of the features ‘publication year’, ‘genre (prose/verse)’, and ‘name version (Roman/Greek)’ on the similarity score that was calculated from the document embeddings. The results make it clear that the choice of names, genre, and use of rhythmic scheme are two important features that have a definite influence on the resulting work. The publication year does not have a significant influence on the similarity scores.

References

Ali, A., Alfayez, F. & Alquhayz, H. (2018). SEMANTIC SIMILARITY MEASURES BETWEEN WORDS: A BRIEF SURVEY, *Sci. Int. (Lahore)*, Vol. 30

Arnold, M. (1905). On Translating Homer, London : Routledge

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With E-Rater® V.2, *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3

Bahdanau, D., Cho, K. & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate, in *Proceedings of the 3rd International Conference on Learning Representations*, 1 September 2015, Available Online: http://arxiv.org/abs/1409.0473

Bär, D., Zesch, T. & Gurevych, I. (2011). A Reflective View on Text Similarity, Available Online: www.ukp.tu-darmstadt.de

Bassnett, S. (2011). Reflections on Translation, Vol. 39, Multilingual Matters

Budanitsky, A. & Hirst, G. (2006). Evaluating WordNet-Based Measures of Lexical Semantic Relatedness

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-Lingual Focused Evaluation, in *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, 2017, pp.1–14

Chandrasekaran, D. & Mago, V. (2022). Evolution of Semantic Similarity—A Survey, *ACM Computing Surveys*, vol. 54, no. 2, pp.1–37

Chowdhary, K. R. (2020). Natural Language Processing, in *Fundamentals of Artificial Intelligence*, New Delhi: Springer India, pp.603–649

Collobert, R., Weston, J., Com, J., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch, *Journal of Machine Learning Research*, Vol. 12

Cox, T. & Cox, M. (2000). Multidimensional Scaling, Chapman and Hall/CRC

Dedović, B. (2018). 'Minds' in 'Homer': A Quantitative Psycholinguistic Comparison of the Iliad and Odyssey;

Delmonte, R. (2022). Measuring Similarity by Linguistic Features Rather than Frequency, Available Online: https://wacky.sslmit.unibo.it/

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, [e-journal], Available Online: http://arxiv.org/abs/1810.04805

Ehrmanntraut, A., Hagen, T., Jannidis, F., Konle, L., Kröncke, M. & Winko, S. (2022). Modeling and Measuring Short Text Similarities. On the Multi-Dimensional Differences between German Poetry of Realism and Modernism., *Journal of Computational Literary Studies*, vol. 1, no. 1

Geffet, M. & Dagan, I. (2005). The Distributional Inclusion Hypotheses and Lexical Entailment, in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, June 2005, Association for Computational Linguistics, pp.107–114

Geng, Z., Cheesman, T., Laramee, R. S., Flanagan, K. & Thiel, S. (2015). ShakerVis: Visual Analysis of Segment Variation of German Translations of Shakespeare’s Othello, *Information Visualization*, vol. 14, no. 4, pp.273–288

Grosman, J. S., Furtado, P. H. T., Rodrigues, A. M. B., Schardong, G. G., Barbosa, S. D. J. & Lopes, H. C. V. (2020). Eras: Improving the Quality Control in the Annotation Process for Natural Language Processing Tasks, *Information Systems*, vol. 93, p.101553

Hearst, M. A. (2000). The Debate on Automated Essay Grading, *IEEE Intelligent Systems and their Applications*, vol. 15, no. 5, pp.22–37

Ilyas, S. (2022). The Parameters of Poetry Translation: A Stylistic Analysis of the Linguistic and Literary Techniques Used in the Translations of the Odyssey and the Iliad, *The Journal of Humanities and Social Sciences*, [e-journal] vol. 30, no. 1, Available Online: https://ssrn.com/abstract=4276125

Ingalls, W. B. (1970). The Structure of the Homeric Hexameter: A Review, *Phoenix*, vol. 24, no. 1, p.1

Jakobson, R. (1959). ON LINGUISTIC ASPECTS OF TRANSLATION, in *On Translation*, pp.232–239

Kamath, U., Liu, J. & Whitaker, J. (2019). Deep Learning for NLP and Speech Recognition, Cham: Springer International Publishing

Kim, H. (2023). Reading Homer’s Iliad in North Korea: A Study on Lim Hak-Su’s Prefaces to His Three Translated Versions of the Iliad, *Korea Journal*, vol. 63, no. 4, pp.84–110

Kiparsky, P. (2020). Metered Verse, *Annual Review of Linguistics*, vol. 6, no. 1, pp.25–44

Kiyani, F. & Tas, O. (2017). A Survey Automatic Text Summarization, *Pressacademia*, vol. 5, no. 1, pp.205–213

Kolb, P. (2009). Experiments on the Difference between Semantic Similarity and Relatedness, Available Online: http://www.altavista.com

Lau, J. H. & Baldwin, T. (2016). An Empirical Evaluation of Doc2vec with Practical Insights into Document Embedding Generation, *Proceedings of the 1st Workshop on Representation Learning for NLP*, [e-journal] pp.78–86, Available Online: http://arxiv.org/abs/1607.05368

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, vol. 86, no. 414, pp.316–327

Low, E. L. (2006). A Review of Recent Research on Speech Rhythm: Some Insights for Language Acquisition, Language Disorders and Language Teaching, in *Spoken English, Tesol and Applied Linguistics*, London: Palgrave Macmillan UK, pp.99–125

Ma, L. & Zhang, Y. (2015). Using Word2Vec to Process Big Text Data, in *IEEE International Conference on Big Data*, 2015

McNiff, S. (2015). Imagination in Action : Secrets for Unleashing Creative Expression., 1st edn, Shambhala

Mendelsohn, D. (2011). Henrician Homer: English Verse Translations from the Iliad and Odyssey, 1531–1545, *The New Yorker*, Available Online:

https://www.newyorker.com/books/page-turner/englishing-the-iliad-grading-four-rival-translation [Accessed 22 March 2024]

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, in *International Conference on Learning Representations*, 16 January 2013, Available Online: http://arxiv.org/abs/1301.3781

Murveit, H. Y. & Moore, R. (1990). Integrating Natural Language Constraints into HNIM-Based Speech Recognition S11.2, in *International Conference on Acoustics, Speech, and Signal Processing*, 1990, IEEE

Palladino, C., Shamsian, F. & Yousef, T. (2022). Using Parallel Corpora to Evaluate Translations of Ancient Greek Literary Texts An Application of Text Alignment for Digital Philology Research, *Journal of Computational Literary Studies*, [e-journal] vol. 1, no. 1, Available Online: http://ugarit.ialigne

Pavlopoulos, J., Xenos, A. & Picca, D. (2022). Sentiment Analysis of Homeric Text: The 1st Book of Iliad, Available Online: https://github.com/ipavlopoulos/

Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October 2014, Association for Computational Linguistics, pp.1532–1543, Available Online: http://nlp.

Sahlgren, M. & Karlgren, J. (2008). Buzz Monitoring in Word Space, in *Lecture Notes in Computer Science* , Vol. 5376, pp.73–84

Sakketou, F. & Ampazis, N. (2020). A Constrained Optimization Algorithm for Learning GloVe Embeddings with Semantic Lexicons, *Knowledge-Based Systems*, [e-journal] vol. 195, p.105628, Available Online: https://doi.org/10.1016/j.knosys.

SciPy. (2024). Scipy.Cluster.Hierarchy.Linkage, *Docs.Scipy.Org*

Senoussaoui, M., Kenny, P., Stafylakis, T. & Dumouchel, P. (2014). A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp.217–227

Sutskever, I., Vinyals, O. & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks, in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014

Tanaka, H., Shinnou, H., Cao, R., Bai, J. & Ma, W. (2020). Document Classification by Word Embeddings of BERT, in *Communications in Computer and Information Science*, Vol. 1215 CCIS, 2020, Springer, pp.145–154

Turnbull, L. (1964). On Translating Greek Lyric Poetry, *Studies in English*, Vol. 5, Available Online: https://egrove.olemiss.edu/ms_studies_engAvailableat:https://egrove.olemiss.edu/ms_studies_eng/vol5/iss1/9

Valli Mayil, V. & Ratha Jeyalakshmi, T. (2023). Pretrained Sentence Embedding and Semantic Sentence Similarity Language Model for Text Classification in NLP, in *2023 3rd International Conference on Artificial Intelligence and Signal Processing, AISP 2023*, 2023, Institute of Electrical and Electronics Engineers Inc.

Wang, H., Wu, H., He, Z., Huang, L. & Church, K. W. (2022). Progress in Machine Translation, *Engineering*

Wang, W., Yan, M. & Wu, C. (2018). Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering, [e-journal], Available Online: http://arxiv.org/abs/1811.11934

Weissbort, D. & Eysteinsson, Á. (2006). Translation: Theory and Practice : A Historical Reader, Oxford University Press

Yang, L. & Zhou, G. (2024). Dissecting The Analects: An NLP-Based Exploration of Semantic Similarities and Differences across English Translations, *Humanities and Social Sciences Communications*, vol. 11, no. 1

Yousef, T., Palladino, C., Wright, D. J. & Berti, M. (2022). Automatic Translation Alignment for Ancient Greek and Latin, in *Language Resources and Evaluation Conference*, 2022, pp.101–107, Available Online: https://huggingface.

Zhang, L., Wang, S. & Liu, B. (2018). Deep Learning for Sentiment Analysis: A Survey, *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4

Zhao, Y. & Karypis, G. (2005). Hierarchical Clustering Algorithms for Document Datasets, *Data Mining and Knowledge Discovery*, vol. 10, pp.141–168

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. & Fidler, S. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books, [e-journal], Available Online: http://arxiv.org/abs/1506.06724

Appendix I

Word	Buckley	Butler	Chapman	Cowper	Edward	Fagles	Fitzgerald	Johnston	Kline	Lang	Lattimore	Lombard	Merrill	Murray	Pope	Rieu	total_count
son	44	38	25	25	27	24	16	35	38	33	44	23	22	46	12	23	475
men	18	13	28	9	16	31	26	34	25	22	29	18	21	24	10	34	358
achilles	21	22	13	22	21	28	0	34	23	21	0	29	21	21	23	27	326
gods	19	17	16	16	16	21	21	27	19	22	19	17	23	22	30	19	324
god	10	15	20	12	17	37	25	30	19	14	13	19	15	20	24	26	316
zeus	0	0	0	0	0	34	25	42	30	32	30	26	34	34	0	29	316
shall	21	35	19	37	15	2	11	0	8	27	22	0	1	25	38	16	277
apollo	17	19	8	17	10	23	15	27	20	17	21	23	17	16	7	18	275
agamemnon	17	23	4	15	11	24	23	28	20	15	15	27	15	14	1	22	274
man	15	15	12	6	8	30	24	23	16	15	23	13	19	17	7	21	264
ships	20	20	4	3	11	18	15	21	19	20	20	16	12	20	3	18	240
king	21	12	21	14	16	15	4	12	11	21	12	9	8	17	22	23	238
heart	10	9	15	7	7	23	10	14	12	25	25	7	13	30	4	6	217
atreus	19	16	4	10	8	8	3	13	11	13	20	8	22	22	3	4	184
spoke	18	8	0	0	9	2	3	13	14	0	34	1	29	36	10	4	181
sea	13	12	11	2	5	13	14	11	11	15	14	12	15	18	1	13	180
went	13	19	9	4	1	17	8	10	2	13	26	11	9	18	0	12	172
come	14	10	12	2	5	12	9	11	10	13	14	8	17	19	1	11	168
prize	14	10	0	9	7	13	9	8	13	2	13	12	14	14	16	12	166
far	18	4	14	6	9	14	6	8	14	20	8	4	15	10	11	2	163

Table 6: Term-frequency matrix of the 20 most frequently used words

Appendix II

A.

Table 7: average pairwise and maximum distances for each century based on the GloVe similarity matrix

Century	Average pairwise distance	Maximum pairwise distance	Number of translations
1800	0.0067	0.012 (Edward – Buckley)	4
1900	0.0082	0.024 (Pope – Lombardo)	7
2000	0.0063	0.014 (Lang – Johnston)	4

Table 8: average pairwise and maximum distances for each century based on the Word2Vec similarity matrix

Century	Average pairwise distance	Maximum pairwise distance	Number of translations
1800	0.0005	0.0015 (Edward – Butler)	4
1900	0.0007	0.0031 (Pope – Murray)	7
2000	0.0001	0.0002 (Lang – Kline)	4

Table 9: average pairwise and maximum distances for each century based on the BERT similarity matrix

Century	Average pairwise distance	Maximum pairwise distance	Number of translations
1800	0.021	0.043 (Edward – Buckley)	4
1900	0.033	0.078 (Lombardo – Murray)	7
2000	0.024	0.064 (Lang – Johnston)	4

B.

Table 10: average pairwise and maximum distances for each genre based on the GloVe similarity matrix

Verse /Prose	Average pairwise distance	Maximum pairwise distance	Number of translations
Verse	0.011	0.021	9
Prose	0.006	0.018	7

Table 11: average pairwise and maximum distances for each genre based on the Word2Vec similarity matrix

Verse /Prose	Average pairwise distance	Maximum pairwise distance	Number of translations
Verse	0.0007	0.0027	9
Prose	0.0002	0.0006	7

Table 12: average pairwise and maximum distances for each genre based on the BERT similarity matrix

Verse /Prose	Average pairwise distance	Maximum pairwise distance	Number of translations
Verse	0.036	0.085	9
Prose	0.027	0.081	7

Appendix III

A.

Table 13: Cluster information GloVe cluster orange

Author	Year	Prose/verse	Metre	Roman/Greek
Butler	1898	Prose		Roman
Murray	1924	Prose		Greek
Rieu	1950	Prose		Greek
Lattimore	1951	Prose		Greek
Fitzgerald	1974	Verse	Iambic pentameter	Greek
Fagles	1990	Verse	Combined use of pentameters and hexameters	Greek
Lombardo	1997	Prose		Greek
Johnston	2006	Verse	Combined use of pentameters and hexameters	Greek
Merrill	2007	Verse	Hexameters	Greek
Kline	2009	Prose		Greek

Table 14: Cluster information GloVe cluster green

Author	Year	Prose/verse	Metre	Roman/Greek
Cowper	1838	Verse	Iambic pentameter	Roman
Edward, earl of derby	1865	Verse	Iambic pentameter	Roman
Buckley	1873	Verse	Iambic pentameter	Roman
Lang, Leaf & Meyers	2002	Prose		Greek

Table 15: Cluster information GloVe cluster red

Author	Year	Prose/verse	Metre	Roman/Greek
Chapman	1611	Verse	Iambic heptameter	Roman
Pope	1900	Verse	Iambic pentameter	Roman

B.

Table 16: Cluster information Word2Vec cluster orange

Author	Year	Prose/verse	Metre	Roman/Greek
Buckley	1873	Verse	Iambic pentameter	Roman
Butler	1898	Prose		Roman
Murray	1924	Prose		Greek
Rieu	1950	Prose		Greek
Lattimore	1951	Prose		Greek
Fitzgerald	1974	Verse	Iambic pentameter	Greek
Fagles	1990	Verse	Combined use of pentameters and hexameters	Greek
Lombardo	1997	Prose		Greek
Lang, Leaf & Meyers	2002	Prose		Greek
Johnston	2006	Verse	Combined use of pentameters and hexameters	Greek
Merrill	2007	Verse	Hexameters	Greek
Kline	2009	Prose		Greek

Table 17: Cluster information Word2Vec cluster green

Author	Year	Prose/verse	Metre	Roman/Greek
Chapman	1611	Verse	Iambic heptameter	Roman
Cowper	1838	Verse	Iambic pentameter	Roman
Edward, earl of derby	1865	Verse	Iambic pentameter	Roman
Pope	1900	Verse	Iambic pentameter	Roman