ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

Corso di Laurea magistrale Specialized Translation (classe LM - 94)

Parallel Sentence Mining: A Semi-Automated Approach for the Creation of a *Comparallel* News Corpus in Greek and English

TESI DI LAUREA MAGISTRALE
in MACHINE TRANSLATION

Relatore

Prof. Federico Garcea

Presentata da

Elton Pistolia

Correlatrice

Prof. Silvia Bernardini

Sessione marzo 2024

Anno Accademico 2022/2023

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

Corso di Laurea magistrale Specialized Translation (classe LM - 94)

Parallel Sentence Mining: A Semi-Automated Approach for the Creation of a *Comparallel* News Corpus in Greek and English

TESI DI LAUREA MAGISTRALE
in MACHINE TRANSLATION

Relatore

Prof. Federico Garcea

Presentata da

Elton Pistolia

Correlatrice

Prof. Silvia Bernardini

Sessione marzo 2024

Anno Accademico 2022/2023

Acknowledgments

Today I sit down to write the acknowledgments for my thesis, a journey that has been a challenge that proved to me that I can do anything if I want to. I find myself overwhelmed with gratitude for all the individuals who have been pillars of support, guidance, and inspiration throughout these 2 and a half years. So, on this page, I want to express my gratitude to all of you.

To Professor Garcea and Professor Bernardini, whose guidance and patience have been indispensable for the completion of this project. Together with Professors Barrón-Cedeño, Ferraresi, and Miličević Petrović, they have been an inspiration for me.

To my colleagues and close friends of the TraTec group, you have been a source of strength, and I am glad I got to meet you.

To everyone I met at the residence Sassi Masini in Forlì, your friendship and company have made it feel like a home away from home.

To my friends back in Trani, despite the physical distance, your constant company and support have been a reminder that true friendship persists.

Finally, to my family who are always by my side and believe in everything I do, even though we are miles apart.

This thesis is not just a reflection of my work but a mosaic of the contributions, encouragement, and faith of each one of you.

Thank you.

Abstract

English

This thesis explores the opportunities presented by the large amount of multilingual comparable data in digital news platforms, focusing on the implications for multilingual news production and Translation Studies. With the rise of online news consumption, as evidenced by the preference for digital platforms over print in Europe over the last few years, there is a growing need for research in news translation. This study addresses the complexity of extracting parallel sentences from bilingual comparable news corpora of Greek and English, aiming to enhance understanding and methodologies within Translation Studies (TS) and Computational Linguistics (CL).

The research investigates the efficacy of cosine similarity measures applied to sentence and word embeddings for identifying parallel (translated) sentences across languages based on semantic similarity, with a focus on the peculiarities of journalistic language and the challenges of aligning sentences that involve not just direct translation but also cultural and contextual adaptation. Through a comprehensive workflow that includes data collection, algorithm implementation, and performance evaluation, this thesis attempts to answer three critical research questions regarding the automatic extraction of pairs of translated sentences and their classification into four categories, namely, translated, partial translation, non-translation, and unrelated, reflecting their translation relationship.

The findings confirm that cosine similarity in combination with sentence and word embeddings can effectively identify semantically similar sentences across bilingual news corpora. Moreover, they enable the categorization of sentence pairs into three categories, i.e., parallel, ambiguous, and unrelated, with further refinement into partial translations or non-translations for ambiguous pairs.

This thesis contributes to the fields of Translation Studies and Computational Linguistics by providing a novel approach to streamline parallel sentence extraction from news articles to enable the study of news translation.

Abstract

Italiano

Questa tesi esplora le opportunità presentate dalla mole di dati comparabili multilingue nelle piattaforme di notizie digitali, concentrandosi sulle implicazioni per la produzione di notizie multilingue e per gli studi sulla traduzione. Con l'aumento del consumo di notizie online, come dimostra la preferenza per le piattaforme digitali rispetto alla stampa in Europa negli ultimi anni, c'è una maggiore necessità di ricerca sulla traduzione delle notizie. Questo studio affronta la complessità dell'estrazione di frasi parallele da corpora di notizie bilingui comparabili di greco e inglese, con l'obiettivo di migliorare la comprensione e le metodologie nell'ambito degli studi sulla traduzione e della linguistica computazionale.

La ricerca studia l'efficacia della somiglianza del coseno (cosine similarity) applicata all'embedding di frasi e parole per identificare frasi parallele tra le due lingue sulla base della somiglianza semantica, con un'attenzione particolare alle peculiarità del linguaggio giornalistico e alle sfide dell'allineamento di frasi che coinvolgono non solo la traduzione diretta, ma anche l'adattamento culturale e contestuale. Attraverso un workflow che comprende la raccolta dei dati, l'implementazione dell'algoritmo e la valutazione delle prestazioni, questa tesi cerca di rispondere a tre domande di ricerca riguardanti l'estrazione automatica di coppie di frasi tradotte e la loro classificazione in quattro categorie, ovvero translated, partial translation, non-translation e unrelated, che riflettono la loro relazione di traduzione.

I risultati confermano che cosine similarity, in combinazione con sentence e word embeddings, possono identificare efficacemente frasi semanticamente simili in corpora di notizie bilingue. Inoltre, consentono di classificare le coppie di frasi in tre categorie: *parallel*, *ambiguous* e *unrelated*, con un ulteriore affinamento in *partial translations* o *non-translations* per le coppie ambigue. Questa tesi contribuisce agli studi sulla traduzione e alla linguistica computazionale, fornendo un approccio innovativo per automatizzare l'estrazione di frasi parallele da articoli giornalistici e consentire lo studio della traduzione di notizie.

Contents

FIGURES

1 Introduction

The shift from traditional print to digital platforms has significantly increased the consumption of online news. As illustrated in Figure 1.1, most countries nowadays prefer accessing news online, which includes through social media. With the Internet's global reach, news agencies are now disseminating news in various languages to cater to a diverse audience. This surge in multilingual online news production offers researchers in Translation studies a valuable chance to investigate the nuances of news translation on a broader scale by accessing large multilingual comparable corpora, i.e., news articles published online. The digital nature of these data further facilitates the process, offering a new valuable source of comparable data that can be used for the subsequent extraction of parallel data. These are fundamental for many applications in the interdisciplinary fields of Translation studies (TS) and Computational Linguistics (CL).

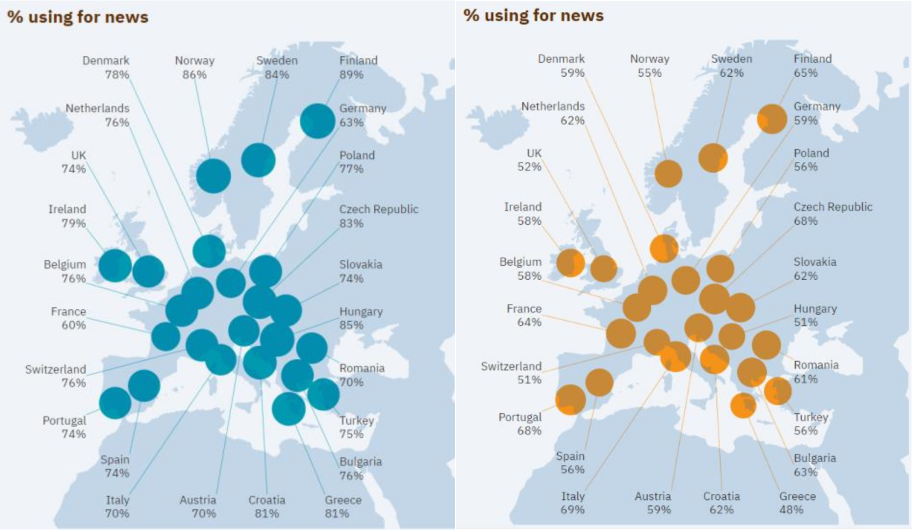


Figure 1.1: Percentage of people using online (blue) vs. print (orange) media for news in Europe 2023[1]

[1] The figures were produced by the interactive view of the Digital News Report 2023 available at [Interactive | Reuters Institute for the Study of Journalism (ox.ac.uk)](#).

Traditionally, TS exploit parallel corpora to study translation techniques. However, the process of news production involves various levels of transformations of the source text, from direct translation to re-writing, challenging the concept of a source and target text which is fundamental to parallel corpora and to translation theory in general. On this premise, there is a clear need for novel corpus approaches to creating bilingual resource for the study of news translation and multilingual news localization in general.

From a computational perspective, extracting parallel sentences from comparable news corpora is a non-trivial task. In general terms, the complexity of accurately aligning sentences across different languages arises from the need to not only directly translate words and phrases, but also adapt to cultural and contextual differences. Specifically for the field of News Translation an added complexity is presented through the peculiarities of journalistic language and localization (which includes, among others, translation, editing and rewriting) approaches, who do not follow a global standard and depend on multiple variables such as language pair, the purpose behind the agency's localization process and its target audience. Algorithms must match underlying meanings and nuances that are often specific to each language pair.

1.1 Research Questions and Hypotheses

Automated text classification systems that categorize sentence pairs based on their translation equivalence must be designed to identify different news translation techniques such as additions or omissions, especially in the domain of news text where these techniques are widely present. This study seeks to devise an automated way to extract parallel sentences from bilingual comparable news corpora of Greek and English. To formalise this investigation, the following research questions were defined:

1st Research Question: How effective are cosine similarity measures applied to sentence and word embeddings at automatically extracting similar sentences from pairs of news articles written in Greek and English, assuming that one article is derived from the other or that they have a common source?

2nd Research Question: After extracting sentences as outlined in RQ1, how accurately can these sentences be classified into categories (parallel, ambiguous, unrelated) that reflect the degree of similarity in terms of the translation relationship?

3rd Research Question: Among sentences classified as ambiguous based on sentence embeddings, how reliably can the system disambiguate ambiguous pairs and categorize them into partial translations or parallel segments?

To answer these research questions, this study attempts to prove the following hypotheses:

1st Research Hypothesis: The application of cosine similarity measures to sentence and word embeddings enables the effective and accurate identification of semantically similar sentences across bilingual news corpora.

2nd Research Hypothesis: Utilizing cosine similarity thresholds for the classification of sentences, based on their semantic similarity, offers a viable method for the initial categorization of sentence pairs into broad predefined categories such as parallel, ambiguous, and unrelated.

3rd Research Hypothesis: Utilizing cosine similarity thresholds on word embeddings in combination with manually crafted features, provides accurate categorization of sentences classified as ambiguous into the predefined labels partial translation and non-translation.

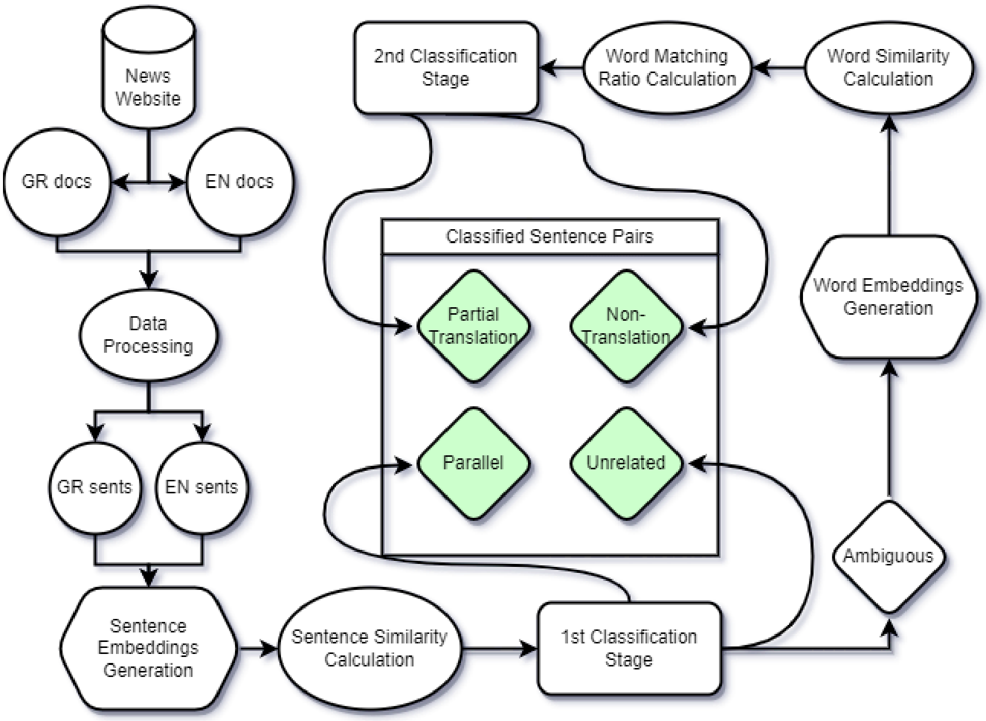

News Website
GR docs
EN docs
Data Processing
GR sents
EN sents
Sentence Embeddings Generation
Sentence Similarity Calculation
1st Classification Stage
Ambiguous
Word Embeddings Generation
Word Similarity Calculation
Word Matching Ratio Calculation
2nd Classification Stage
Classified Sentence Pairs
Partial Translation
Non-Translation
Parallel
Unrelated

Figure 1.2: Schematic Overview of the proposed System

To test these hypotheses, a comprehensive workflow was designed. The process starts with the data collection phase to construct the input data and leads in the generation of the output data, as shown in Figure 1.1. To effectively address the research questions, a multifaceted evaluation strategy was employed. This involved manually annotating output data and applying evaluation metrics to quantify the system's performance. In addition, several graphical representations were produced to visually interpret the results.

This approach allows for a comprehensive evaluation of the system's ability to identify and classify semantically similar sentences in bilingual comparable corpora. The study also highlights the intricacies of determining the semantic similarity of two sentences and of classifying them into categories denoting the translation relation between them based on that similarity.

1.2 Thesis Outline

The present thesis has 3 main Chapters (excluding the introduction and conclusion):

First the theoretical framework of the research is defined in Chapter 2, offering the theoretical basis that supports the proposed approach and informs the decisions made throughout the workflow. The theoretical concepts detailed in this chapter include foundational Natural Language Processing (NLP) concepts and more advanced Cross-Lingual Information Retrieval (CLIR) concepts, accompanied by studies proposing contemporary approaches. Finally, the chapter discusses news translation and its significance in the context of this project.

The methodology proposed in this study is outlined in Chapter 3. This part of the study provides practical details on the implementation of the proposed algorithm, starting with data collection and describing all the processing steps illustrated in Figure 1.2. The chapter ties together the theoretical concepts explored in this study and offers details on the experiments carried out during this research.

The results of the experiments are presented, analysed, and discussed in Chapter 4. This chapter includes several graphical representations of the results and a detailed analysis of the system's performance, looking at the evaluation metrics and conducting an error analysis to better understand its abilities and limitations. The chapter ends with a discussion focusing on the limitations and potential solutions.

Chapter 5 summarises the key findings and contributions of the thesis, highlighting its implications for the relevant field. It also outlines promising directions for further research, identifying gaps and opportunities that have emerged from the study's findings.

2 Theoretical Framework

The present study draws on several disciplines, including Corpus Linguistics, NLP, and News Translation. The following sections provide the theoretical concepts that underlie this study and are necessary for a better understanding of the project's practical implications. Section 2.1, outlines a general presentation of corpus typology, describing traditional approaches and exploring a novel approach. Sections 2.2 & 2.3 introduce relevant notions from the fields of Computational Linguistics, namely Text Similarity and Parallel Sentence Extraction. In closing, section 2.4 covers the theory of News Translation and underscores the need for new corpus approaches to fully exploit news corpora for Translation and Journalistic studies.

2.1 Corpus Typology

Corpora have been at the core of linguistic research for decades. They represent the main tools for Corpus Linguistics research and play a vital role in Translation studies and Computational Linguistics. Corpora have numerous applications in Machine Translation, Lexicography, Language learning, and more. For example, Neural Machine Translation engines such as Google's NMT system (Wu *et al.*, 2016) use large bilingual corpora to understand and replicate translation choices. There are more specific use cases for corpora in NLP, some of which are discussed in the following sections.

Before exploring specific applications, it is important to establish a baseline understanding of corpora and to clarify the typology of these linguistic resources. It is crucial to differentiate between the various types of corpora, as each possesses distinct characteristics that determine its suitability for specific research purposes. This foundational knowledge is essential for comprehending the full scope of this study.

2.1.1 Parallel and Comparable Corpora

Corpora have been used extensively in the disciplines mentioned above; however, a precise common terminology is still not agreed upon among researchers in different fields. Translation studies and Contrastive studies differ in their terminology, indicating distinct types of corpora with the same terms. Granger (2010) proposed a general typology outlined in the diagram in Figure to unify the terminology.

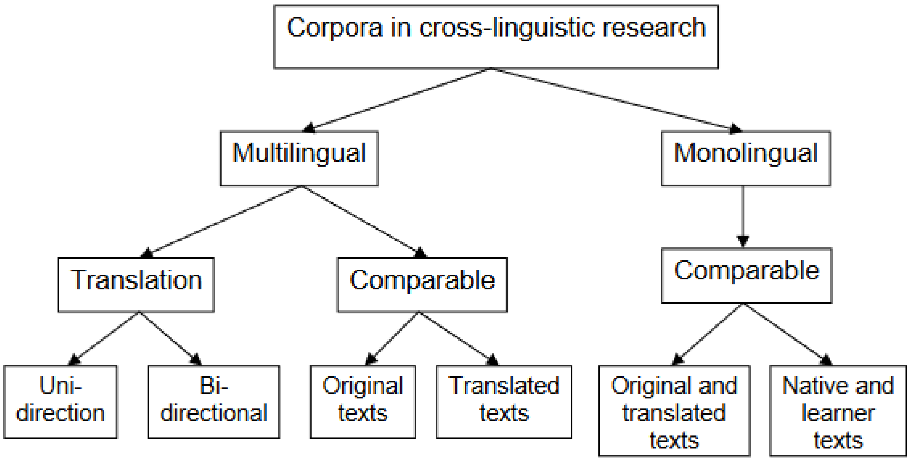


Figure 2.1: Corpora in cross-linguistic research (from Granger 2010)

The diagram defines corpora based on the distinction between multilingual and monolingual corpora. Monolingual corpora are comparable corpora that can be used to study translation features by examining original and translated text within the same language. Multilingual corpora can either be translation corpora or comparable corpora. Translation corpora contain texts in the source language and their translations in the target language, thus allowing the study of translation universals, i.e., shifts resulting from the process of translation that are characteristic and exist only in translated text , as well as comparisons between two or more languages. Multilingual comparable corpora instead can be used to compare languages by examining comparable original texts in different languages, or they can be used to study the nature of translated text by examining comparable corpora of translated texts (Granger, 2010).

The diagram does not include parallel corpora, since the term is generally considered very ambiguous in the literature, as it has been used to describe translation corpora and comparable corpora. This confusion is mainly because of inconsistencies around the criteria used to define a corpus. In this study, parallel corpora consist of source texts and their translations.

Parallel corpora can be characterised based on different features, such as the number of languages included and the translation direction. They can be unidirectional, bidirectional, or multidirectional (McEnery & Xiao, 2007). Directionality is a key concept in this thesis project, as it is affected by the domain of the texts in a parallel corpus. In the domain of news, the directionality is sometimes lost due to the nature of news translation. Another key feature is segment alignment. Parallel corpora are accurately aligned at phrase, sentence, or paragraph level. Alignment accuracy evaluation is an important task in computational linguistics as is the case in this study.

Recently, parallel corpora have become essential in NLP for tasks such as sentiment analysis, text summarization, and Information Extraction (IE). Their use has also extended to CLIR, which enables the development of more sophisticated search engines which can understand and retrieve information from different languages.

Comparable corpora are defined as such in the sense that they usually contain texts in the same genres and domains, sampling period, and representativeness (McEnery and Xiao, 2007). These features are not relevant for parallel corpora where the key features are the translation relationship between source and target text, and the directionality of that relationship. Comparable corpora are not aligned at any level and there is no translation relationship or directionality between the subcorpora in a comparable corpus.

In Contrastive studies, comparable corpora are essential for exploring the specificities of different languages and the variation between them (Granger, 2010). These corpora are crucial for extracting bilingual lexicons, which are the foundation of glossaries. Furthermore, parallel sentences can be extracted from comparable corpora. This is a key process in the development of machine translation systems and the compilation of language resources for educational purposes.

2.1.2 Comparallel Corpora

In the age of globalization, the vast amount of multilingual text data available online necessitates innovative approaches to linguistic studies. Recent advancements in technology have facilitated multilingual localisation. As a result, the web is full of texts that have been either directly translated or "transcreated" for diverse audiences, presenting a challenge to traditional parallel and comparable corpus methodologies.

Comparable news corpora may contain articles that are related to each other, in the sense that one derives from the other. This relation of derivability, when observed at a sentence level, can be the result of direct translation or partially translated, thus entailing a parallel relation between two sentences. In cases of partial translations resulting, e.g., from omissions or additions, the element of directionality of the translation is lost and the source text is unknown. Thus, simply classifying these as parallel texts would be misleading. Gaspari (2015) suggests that such collections of texts are better studied through a "comparallel" corpus approach, adapting to the intricacies of web-based, transformed content.

The concept of comparallel corpora has been discusses by a handful of scholars. Comparallel corpora represent a new approach to creating and studying corpora, combining comparable and parallel corpus approaches to address the intricacies of multilingual text analysis. Bernardini et al. (2010) coins the term "comparallel" and highlights the practical and didactic interests in transforming Wikipedia into a comparallel corpus. Wikipedia is seen as a large collection of monolingual corpora which can be used to extract smaller multilingual comparable corpora and use them to extract parallel segments.

The main feature of comparallel corpora is the way comparability is considered. Comparable corpora allow one to compare sentences without assumptions of directionality while still considering the sentences parallel or translations each other. Davier et al. (2018) stress the challenges in classifying texts within parallel or comparable corpora due to difficulties in identifying clear source-target relationships. They advocate for multilingual comparable corpora (or *comparallel*) corpora as a solution, emphasizing their utility in examining multilingual and translational characteristics of content, especially within news flows where the origins of translations are often obscured.

A good implicit definition of comparallel corpora is given by Gaspari (2015) who suggests that:

> despite the ostensible parallelism, the collection of bilingual texts under analysis should not be viewed as a parallel corpus at least, not in the traditional sense that one can confidently set apart the source and target languages/texts involved. (Gaspari, 2015, p. 330)

This underlines the necessity for a paradigm shift in how bilingual texts are approached, moving away from rigid classifications towards a more flexible understanding that reflects the varying processes of translation such as direct translation, paraphrasing, rewriting and summarization.

2.2 Text Similarity

In the field of NLP, text similarity is a fundamental concept that underlies a wide range of applications, from recommendation systems to plagiarism detection. At its core, text similarity seeks to quantify how closely two pieces of text are related in meaning or content. However, this is a multifaceted challenge due to the complexity and nuances of human language. This section explores the complex mechanisms that allow machines to capture and measure similarity between textual elements, focusing on the role of word and sentence embeddings and the different metrics that assess their similarity.

2.2.1 Embeddings

Word embeddings are a way to represent text in numerical values so that it can be analysed by computer algorithms. Models designed to generate word embeddings can be divided into two categories, that is *Count-based* and *Predictive* methods (Mandelbaum & Shalev, 2016).

Count-based models like the Positive point-wise mutual information (PPMI) matrix factorization method (Levy and Goldberg, 2014), leverage statistical properties of the corpus by counting how often words co-occur in certain contexts, often in a matrix that is then reduced in dimensionality to form word vectors. Predictive models, like Word2Vec (Mikolov *et al.*, 2013) or GloVe (Pennington *et al.,* 2014), use neural networks to predict words in context. Generally, they produce more accurate representations as they consider the predictive nature of words within a corpus, rather than just their co-occurrence statistics. Nowadays, state-of-the-art embedding models are based on the BERT (Bidirectional Encoder Representations from Transformers) framework (see Section 2.3).

In addition to creating embeddings for single words, researchers have also developed models to generate sentence embeddings. Sentence embeddings carry semantic and contextual information for sentences in a corpus.

Considering the importance of sentence embedding in several NLP tasks, there have been many sentence embedding models through the years. Lately, sentence embedding models are also based on the BERT framework.

2.2.2 Cosine Similarity

Text similarity metrics are fundamental for comparing the semantic similarity between words, sentences, or documents. Among all metrics, *cosine similarity* has emerged as a fundamental tool used in combination with word and sentence embeddings. Cosine similarity is defined as the cosine of the angle between the vectors, which is the dot product of the vectors divided by the product of their lengths.

For text embeddings, where documents are represented as vectors in a high-dimensional space, cosine similarity assesses the degree of semantic similarity based on the direction of the vectors rather than their length. This characteristic makes it highly effective for comparing documents of varying lengths and for applications requiring the evaluation of the semantic proximity of texts, as it is less sensitive to the differences in document size.

2.2.3 Text Classification Evaluation Metrics

Evaluation metrics for text classification are an important aspect for every study using text classification as they provide information that allows for the correct assessment of the performance of the classifier. The most common classification metrics are Accuracy, Precision, Recall and F1-score. Each one offers specific insight on itself; however, they are usually calculated as a group since they are interconnected.

Accuracy is the simplest metric to measure performance, and it is defined as the ratio of the correct predictions to the total number of predictions. Accuracy is a good metric for balanced datasets, but it might be misleading when there is a class imbalance. It provides an overall idea of how effective the classifier is, but it does not consider class imbalance.

Precision measures the proportion of True Positive (TP) results in relation to all positive results (the positive class is the one that classifiers are designed to predict), including both TP and False Positives (FP). It answers the question: "Of all the items labelled as positive, how many are actually positive?"

Recall measures the proportion of TP identified in relation to all actual positives in the data. It answers the question: "Of all the actual positives in the data, how many did the classifier correctly identify?"

Precision and recall are interconnected in such that improving one generally leads to a reduction in the other. For some applications precision is important while for others recall is more critical. The F1-score is defined as the harmonic mean of precision and recall. This means that a higher F1-score represents a better performance of the classifier. The classifier will only get a high F1-score if both precision and recall are high.

In the context of imbalanced datasets or when multiple classes are involved in the classification task the micro average and weighted versions of these metrics provide a more accurate evaluation. For precision and recall, micro averaging calculates these metrics based on the overall true predictions (True Positives+True Negatives) and total predictions, respectively, across all classes. This method gives equal weight to each sample, making it a better choice compared to macro averaging for imbalanced datasets as it accounts for the frequency of each class. Weighted averaging on the other hand calculates precision and recall for each class but averages them, weighing by the number of true instances in each class. This means that in measuring performance, the results from larger classes have more weight than those from smaller ones. This makes the metric more useful because it shows how well the classifier works on datasets where some classes have more samples than others, which is representative of situations in real-word applications.

2.3 Parallel Sentence Extraction Approaches

The vast amount of linguistic data available on the Web, including large collections of multilingual text, is a valuable resource for creating corpora. Among these, Wikipedia has been used extensively because of its coverage across various languages and domains, making it an ideal source for comparable multilingual texts. Researchers have also leveraged Wikipedia's extensive variety of languages to extract comparable corpora (Barrón-Cedeño *et al.*, 2015; Wołk *et al.*,2015). Other common sources of comparable multilingual data are institutional (Liu *et al*., 2018) and news agency (Clough *et al*., 2002; Sharjeel *et al.*, 2023) websites.

The extraction of parallel text from documents in different languages is a process known as parallel sentence extraction or bitext mining and is a key element of NLP and CLIR research. It is particularly valuable because of the limited availability of parallel online resources. Specifically, it is used to address the lack of parallel data for low-resource languages and for pairs of commonly used languages, such as Chinese and Portuguese, which lack bilingual parallel corpora (Liu, *et al.*, 2018).

The extraction of parallel sentences can be seen as an alignment task that can be based on various approaches to identifying similarities between sentence pairs. Several methods can be used to achieve alignment between two units of text (usually sentences or paragraphs), from simple ones based on sentence length and lexical similarity to more complex algorithms that incorporate syntactic and semantic analysis. In recent studies, the most common alignment method in parallel text mining is the use of contextual embeddings such as those provided by BERT or multilingual BERT (mBERT).

BERT is a Deep Learning model based on the Transformers' architecture which in turn is based on Neural Networks. This model can detect patterns in textual data by analysing each word in the context of all other words in a sentence in both directions. This allows the model to capture the meaning of a word based on the surrounding context, leading to good performance in many Natural Language Processing (NLP) tasks. BERT is easy to finetune for specific applications by “simply adding task specific inputs and outputs into BERT and finetun[ing] all the parameters end-to-end" (Devlin *et al.*, 2019)

Multilingual sentence embeddings are used as the basis for extracting parallel sentences from large comparable corpora. Multilingual embeddings were typically used in combination with cosine similarity thresholds to discriminate between parallel and non-parallel sentences (Guo *et al.*, 2018; Schwenk, 2018). After 2018, researchers found success in the bitext mining task by using multilingual embedding in combination with more flexible thresholding techniques. Artexte and Schwenk developed the LASER model, by creating a margin-based scoring approach to solve the scalability problem of hard-to-define cosine similarity thresholds. The LASER model performs well at extracting exact translations but is less performant for assessing the similarity of two sentences (Reimers & Gurevych, 2020). The LaBSE (Language-agnostic BERT Sentence Embedding) model presented by (Feng *et al.*, 2022), is based on dual-encoder models, and uses an additive-margin before their scoring function that allows the model to distinguish better between translations and nearby non-translations. LaBSE performs better at the Semantic Textual Similarity (STS) task, but struggles at confidently identifying exact translations.

2.4 News Translation

The digital age has transformed the way news is disseminated, with online platforms becoming the main medium for the global exchange of information. This shift towards digital news outlets not only facilitates instant access to global news, but also provides a unique opportunity for linguistic and Translation studies. The wide availability of online news in different languages provides a rich source of comparable textual data. This wealth of information allows researchers and practitioners to analyse, compare and translate texts across languages, thus increasing our understanding of linguistic nuances, cultural contexts, and the challenges of translating news content.

News Translation in the context of news agencies entails many different tasks of text transformation between two languages. Bielsa (Bielsa, 2007, p. 142) writes:

> […] we will simply refer to news translation to point to this particular combination between editing and translating, and more specifically to the form that translation takes when it has become integrated in news production within the journalistic field.

News translation is characterised by strategic adaptations, including translation shifts such as cutting, rewriting, and restructuring content to fit the new linguistic and cultural context. The work of Davier (Davier, 2022) illustrates this process, noting how elements of the source language are transformed or omitted to serve the needs and expectations of the target audience.

The process of news production in news agencies is not clearly defined, but many researchers have identified a set of key principles of news production in news agencies. Bielsa (2007) cites speed and hierarchy as the two main principles in news production. This means that news articles have to be produced in time and that some news are more important than others. This is highlighted by the fact that not all news is translated. Regarding production in a different language than the native language of production, Davier (Davier, 2022) identifies three main guiding principles: "accuracy of information [...] prioritized over faithfulness to a source text; speed of production; and strong adaptation to the target readership." These principles complicate the relationship between source and target text making it more difficult to determine.

An additional result of the series of transformations applied to the source text is the weakening of the source and target relationship. Davier (Davier, 2019, p. 73) writes:

> Even in the instances where a source text seems to exist, unexpected problems arise: the exact translation direction is not always clear; a written text can be presented as a source, although journalists based their translation on the oral version of a speech; and there can be pseudo-translations for political reasons (Holland 2006).

This highlights that the source text might not be the authentic source of the information and that news production includes transformations that erase authorship from the source text. In fact, generally news articles derive from the work of different journalists, editors and translators, whose initials are attached to news articles; however, this information is usually not consistently available to the public. The erasure of authorship poses methodological challenges when it comes to news translation research (Davier and Doorslaer, 2018). This concept ties back to comparallel corpora (see Section 2.1.2) which offer a way to study news translation through corpora.

The exploration of comparallel news corpora within Translation studies offers unique insights into the nuances of news translation and editorial strategies. The goal is not merely to trace the journey of news from a source to its translated versions but to delve deeper into the processes that shape this journey, highlighting the editorial decisions and translation practices that influence the final presentation of news articles.

The distinction between parallel and comparable corpora is also highlighted by Davier and Doorslaer (2018) who claim that it may become less relevant in the context of news translation research. The concept of "comparallel corpora" offers a framework for including texts where the source is not definitively known. This approach acknowledges the complex realities of news translation, where texts are often not direct translations but are influenced by multiple sources and editorial interventions.

Moreover, the focus on comparative analysis is usually on text fragments rather than entire texts due to the prevalence of omissions, additions, and the practice of drawing from multiple sources in journalism(2002, p. 1679) (Zanettin, 2021). This reality underscores the need for methodologies that can accommodate the patchwork and fragmentary nature of news translation, allowing for the study of how texts evolve and deviate as they cross languages.

Clough et al. (2002, p. 1679) pose the following question:

> "Given two texts is it possible to determine, within acceptable levels of probability, whether one text is derived from the other?"

This question points to the broader challenge of tracing influence, adaptation, and transformation in the flow of news content, which requires both technological and conceptual innovations to navigate the complexities of global news narratives.

2.5 Summing up

This chapter has outlined the theoretical concepts that underpin the methodology used in this study. The exploration of text similarity, parallel sentence extraction and news translation has allowed us to delve into the interactions between computational linguistics and Translation studies. This exploration has not only highlighted the technical advances in the field, but also the challenges and opportunities that arise when dealing with multilingual text analysis and translation.

Advanced computational models, particularly those leveraging contextual embeddings, offer sophisticated means of understanding and comparing texts beyond mere surface-level similarities. This technology enables a more granular analysis of text. The section on parallel sentence extraction shed light on the critical role of alignment techniques in mining bilingual texts from large corpora.

The exploration of news translation highlighted the unique challenges and strategies involved in this specific domain of translation. The chapter has underscored the importance of innovative research methodologies that can navigate the complexities of working with news datasets and addressing the intricacies of source-target text relationships.

The concept of comparallel corpora offers a flexible framework for studying translation phenomena in contexts where direct comparisons between source and target texts are not straightforward. Building upon the theoretical framework, the ensuing methodology chapter introduces the implementation of the system designed specifically to facilitate the extraction of sentence pairs suitable for inclusion in a comparallel corpus. This system leverages computational techniques and components to identify and classify sentence pairs from bilingual comparable news articles, thereby potentially constructing a comparallel corpus with data that captures the intricacies of news translation.

3 Methodology

Building upon the theoretical framework laid out in the previous chapters, this chapter aims to detail the comprehensive methodology employed in this study for addressing the research questions, specifically focusing on the extraction of parallel sentence pairs from a bilingual comparable corpus of news articles. The articles used are published on the Greek National News Agency's website (Athens-Macedonian News Agency)[2]. The chapter is organized into three sections and details the technical strategies employed to facilitate cross-lingual comparison and analysis within the corpus.

Section 3.1 discusses the steps involved in data acquisition and processing, including the selection criteria and rationale behind the implemented processing steps. A thorough explanation of the method for extracting parallel sentences is provided in Section 3.2, encompassing information on the NLP components that constitute the designed classification system. Lastly, Section 3.3, delves into the implementation of the classification system.

3.1 Data Acquisition and Processing

3.1.1 Comprehensive Data Collection Strategy

The research process started with the development of a comprehensive news dataset. The data collection task followed a search protocol, the aim of which is to provide an easy-to-follow workflow, offer transparency, and facilitate research reproducibility.

The search specifically targeted articles related to the topic of migration published from January 2020 through October 2023 by the Athens-Macedonian News Agency (AMNA), which is Greece's National News Agency. The goal of

[2] The agency produces news that is published online on the agency's website (https://www.amna.gr/en) in Greek and English.

the search was to extract candidate related article pairs. This span was selected because of the social, cultural, and geopolitical implications associated with migration during this period. Specifically, the period aligns with significant political developments in Greece, notably the appointment of Kyriakos Mitsotakis as the Prime Minister in July 2019, who had to deal with a variety of events regarding migration (*Migration in Greece: Recent Developments in 2019*, 2020).

3.1.2 Targeted Keyword Search and Query Optimization

The AMNA website provides limited options for users to search and select tailored news articles. To refine the search results, specific keywords were used to query the website search engine. The approach of querying search engines with keywords to extract documents is a foundational technique in the field of Information Retrieval (IR). It leverages specific vocabulary or document metadata to navigate large data repositories. Specifically, in CLIR tasks, bilingual wordlists have been utilized to enable the detection of candidate parallel document pairs from large sources, such as the World Wide Web (Chen *et al.*, 2004). Bilingual wordlists serve not only as a direct translation tool but also as a means to capture semantic equivalence across languages, enhancing the precision of retrieved documents.

In this study, the search queries are based on a small list of keywords in the domain of migration. The list contains words that are expected to represent the domain and be present in most of the articles about migration. Considering the source of the data and the specificity of the domain, this approach efficiently yields many relevant articles.

Domain-specific keywords were used to limit the results and try to avoid very general articles that could be considered as noise. The search query construction was, thus, based on the following lists of keywords:

***en:** migration, migrant, immigration, immigrant, illegal, asylum, rescued*

***gr:** μεταναστευτικό, μετανάστες, παράνομοι, άσυλο, διασωθέντες*

This list functioned as a basis for other related search queries including derivatives of these terms. To ensure that all the derivatives were captured, the asterisk (*) symbol was used, which generally represents a wildcard in search engine queries. As a result, the query sub-word terms used were “migra*” and “immigra*” for English, while for Greek the equivalent “μεταναστ*” was used. In cases where more general query terms are selected, one might decide to include synonyms of such terms to yield a higher number of articles where different words are used in the same context.

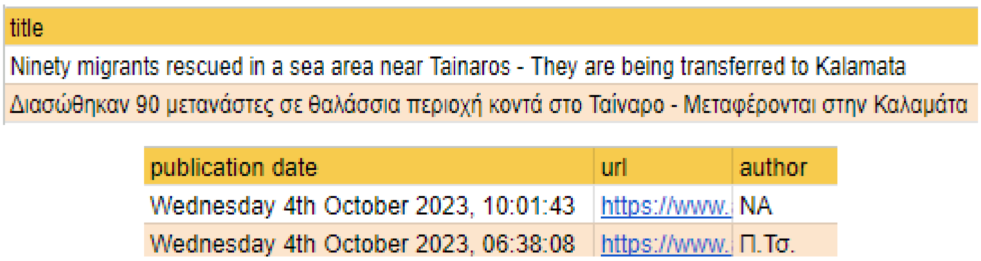
In this approach, synonyms were not included in order to maintain a targeted dataset that accurately reflects the topic of migration without skewing the search results with marginally related articles. However, the omission of synonyms may limit the scope of the search to some extent. This methodology was chosen on the assumption that the core terms used were sufficiently representative of the literature on migration, based on preliminary research. In future work, extending the search to include synonyms could yield a broader data set, potentially revealing additional facets of the migration discourse.

3.1.3 Data Screening and Quality Assurance

Depending on the context of the study, different screening processes can be implemented to discard documents that are considered noise and ensure the quality and validity of the collected data. This can include removing duplicates, and irrelevant articles or it can be based on more specific criteria that provide full control to the researcher on the characteristics of the data.

Both deduplication and length criteria were applied to ensure that the data consisted of contextually rich segments. These criteria also allow for consistency in cases of multiple researchers working on the same project. To monitor and streamline the screening process, an article metadata database was created.

The database was used to store key metadata, critical for identifying potentially related articles, such as headlines, publication dates, URLs, and authors (although author information was often limited to initials) while also allowing one to avoid duplicate article pairs. Such metadata comes with each article on the website and was collected during the search process. From a practical perspective, the database can also be used as a log for future work within the same research project.



title
Ninety migrants rescued in a sea area near Tainaros - They are being transferred to Kalamata
Διασώθηκαν 90 μετανάστες σε θαλάσσια περιοχή κοντά στο Ταίναρο - Μεταφέρονται στην Καλαμάτα

publication date	url	author
Wednesday 4th October 2023, 10:01:43	https://www.	NA
Wednesday 4th October 2023, 06:38:08	https://www.	Π.Τσ.

Figure 3.1: Example of an item entry in the metadata database

The metadata shown in Figure 3.1 were used for the initial screening and pairing process. This methodology was complemented by a manual assessment of document similarity, aimed at verifying the contextual relevance of the articles through the analysis of similar structural patterns and coverage of the same events. While this step ensures content relevance, one can also argue that in situations with limited human and time resources, such manual effort could be minimized or even omitted.

This approach yielded 43 articles, 2 of which were discarded based on article length criteria. For this study, the decision was made not to include short articles called “briefs” that consisted of less than three sentences. These articles seem not to hold much interest either in the context of parallel sentence extraction or for the study of journalism or journalistic translation.

3.2 System Design and Implementation

The methodology employed in this study draws inspiration from prior research in the field of multilingual parallel sentence extraction. More specifically, the approach to developing a semi-automated system for parallel sentence extraction primarily builds upon the work of Schwenk (2018), where cosine similarity was used with bilingual sentence embeddings. The implemented pipeline leverages standard NLP components to construct a system that relies on both sentence and word embeddings for its operation.

3.2.1 Data Loading

The system is specifically designed for a set of corpora in two languages. In the experiment conducted in the context of the present work, the collection of Greek and English news articles is loaded into the system as two separate monolingual corpora. The data are processed using a variety of processing techniques as detailed below.

Data processing is a critical step in NLP and Machine Learning (ML) projects. The performance of a model on a specific task heavily relies on the pre-processing and post-processing steps, which should be carefully constructed and clearly defined to ensure the validity of the approach. The data processing step used for this project consists of sentence segmentation, tokenization, short segment merging, and cleaning of elements that were considered noise and could negatively impact the performance of the system.

Firstly, the text of each document in the two collections is split into sentences using SpaCy models[3]. SpaCy offers language models of varying sizes for many languages. Specifically, the parser components of the *el_core_news_md* and

[3] Available at https://spacy.io/models.

en_core_web_md models were used to perform sentence segmentation of the two collections. Despite being trained on different types of data (news vs web), this difference should not be significant given the exact sources of the models found on each model's details table. Additionally, slight differences in training data types should not affect the performance of the components used in the proposed system.

This process resulted in 358 English and 808 Greek sentences. The discrepancy in the number of sentences can be justified by a general trend observed during the data collection process, where Greek articles were longer and included direct quotes with shorter sentences. The English articles usually used indirect speech transforming short quotes into single sentences. Moreover, structural differences between the articles in the two languages affected sentence segmentation. To normalize some structural differences, for example, lists of new policies or adopted measures, numbers, or special characters such as hyphens, were removed when found at the beginning of the segments as they interfered with the segmentation process.

To enhance the quality of the data after the cleaning step, a decision was made to merge segments based on character length. Different data merging strategies can be applied depending on the dataset characteristics and the analytical focus. In this study, sentence-level analysis is conducted to observe and examine the nuances of translation and other modifications inherent in the news production process across different languages.

Given this objective, the preprocessing included a specific strategy for handling short text segments. To ensure that the units of analysis contained sufficient information for meaningful observation, segments comprising fewer than 20 characters were merged with the preceding segment. Such segments were appended to the preceding segment rather than the following one, under the assumption that they are more likely to be a continuation of a previously introduced thought rather than the beginning of a new one. This assumption exploits the unique characteristics of journalistic text, where short passages - less than 20 characters - are atypical. Such brief snippets are more frequently present only in direct speech quotations, as individuals tend to produce short utterances when speaking. This method preserves sentence-based analysis but also enriches the segments, making them more informative and complete. Following this preprocessing step, the dataset was organized into a set of segments, totalling 734 for the Greek corpus and 351 for the English corpus.

The extracted Greek sentences were automatically translated into English to allow for monolingual comparison of the sentences using word embeddings. The choice of the specific translation engine is a variable that could influence the system's performance. However, in this study, such potential differences were not considered. The Google Cloud Translation API[4] was integrated into the pipeline and was used to translate all the sentences from Greek to English.

The final processing step was the tokenization of the translated and original English sentences. The extracted tokens were used to generate the word embedding representation of the articles to be compared. The SpaCy models were used for the tokenization step, removing all punctuation from the tokens, to reduce the dimensionality of the dataset. Considering that punctuation does not carry semantic weight in the context of word embeddings, this step also allowed for clearer graphical representations such as heatmaps of word embeddings similarity scores.

3.2.2 Generation of Sentence Embeddings

The system uses a sentence embedding model to generate the embeddings of the sentences that were previously extracted. These embeddings represent the semantic and contextual information of a sentence. This decision was informed

[4] Find the documentation at https://cloud.google.com/translate/docs/reference/rest/.

by conducting a preliminary experiment to verify the performance of sentence embedding models on the task of bitext-mining.

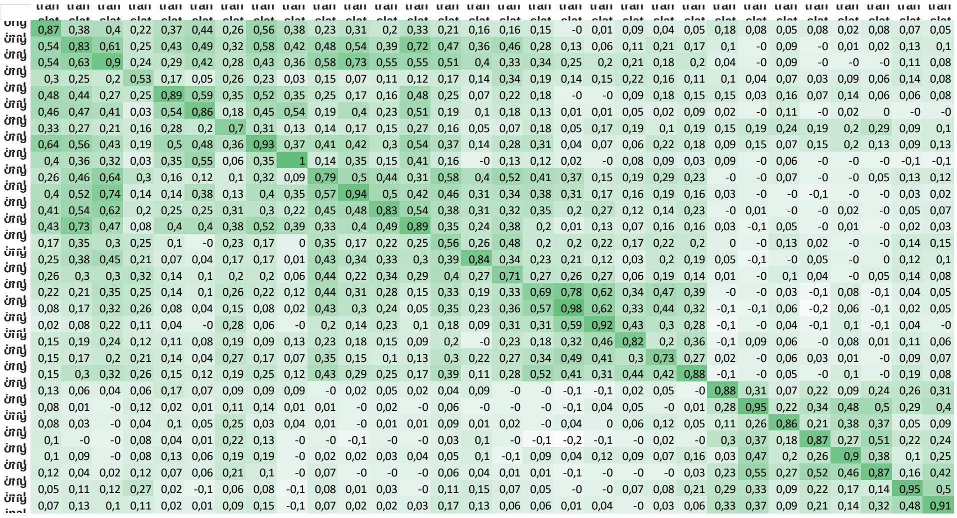
For that, the NTREX[5] (News Text References of English into X Languages) (Federmann *et al*, 2022) dataset was used. The dataset contains pairs of parallel sentence pairs in the domain of news. This enables observations on the performance of the embedding model when used to identify translated sentence pairs by comparing the generated embeddings with cosine similarity.

Specifically for this experiment, the English-Greek pairs were used to generate the embeddings using the *all-MiniLM-L6-v2* sentence embeddings model[6] which is based on the *SentenceTransformers* framework[7]. The generated embeddings were compared using the cosine similarity metric and the heatmap in Figure 3.2 was produced. This visual representation is essential for identifying parallel sentences and assigning a quantifiable score to each sentence pair. This score reflects the probability of the sentences being translations of each other. The heatmap serves as an indicator of how likely a pair of sentences is to be translated and illustrates the effectiveness of the sentence embedding model in distinguishing between likely translated and unrelated pairs of sentences.

[5] The dataset is available at https://github.com/MicrosoftTranslator/NTREX.

[6] The model is available at https://github.com/UKPLab/sentence-transformers or https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

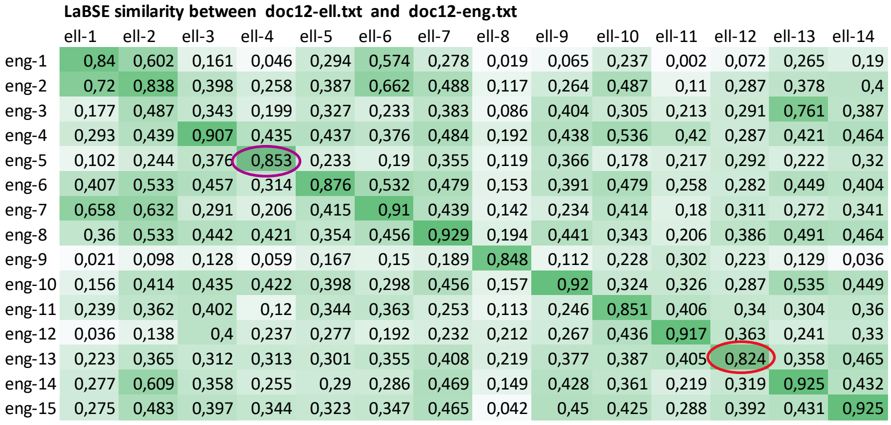
[7] The documentation is available at sbert.net.



	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat	tran slat
orig	0,87	0,38	0,4	0,22	0,37	0,44	0,26	0,56	0,38	0,23	0,31	0,2	0,33	0,21	0,16	0,16	0,15	-0	0,01	0,09	0,04	0,05	0,18	0,08	0,05	0,08	0,02	0,08	0,07	0,05
orig	0,54	0,83	0,61	0,25	0,43	0,49	0,32	0,58	0,42	0,48	0,54	0,39	0,72	0,47	0,36	0,46	0,28	0,13	0,06	0,11	0,21	0,17	0,1	-0	0,09	-0	0,01	0,02	0,13	0,1
orig	0,54	0,63	0,9	0,24	0,29	0,42	0,28	0,43	0,36	0,58	0,73	0,55	0,55	0,51	0,4	0,33	0,34	0,25	0,2	0,21	0,18	0,2	0,04	-0	0,09	-0	-0	-0	0,11	0,08
orig	0,3	0,25	0,2	0,53	0,17	0,05	0,26	0,23	0,03	0,15	0,07	0,11	0,12	0,17	0,14	0,34	0,19	0,14	0,15	0,22	0,16	0,11	0,1	0,04	0,07	0,03	0,09	0,06	0,14	0,08
orig	0,48	0,44	0,27	0,25	0,89	0,59	0,35	0,52	0,35	0,25	0,17	0,16	0,48	0,25	0,07	0,22	0,18	-0	-0	0,09	0,18	0,15	0,15	0,03	0,16	0,07	0,14	0,06	0,06	0,08
orig	0,46	0,47	0,41	0,03	0,54	0,86	0,18	0,45	0,54	0,19	0,4	0,23	0,51	0,19	0,1	0,18	0,13	0,01	0,01	0,05	0,02	0,09	0,02	-0	0,11	-0	0,02	0	-0	-0
orig	0,33	0,27	0,21	0,16	0,28	0,2	0,7	0,31	0,13	0,14	0,17	0,15	0,27	0,16	0,05	0,07	0,18	0,05	0,17	0,19	0,1	0,19	0,15	0,19	0,24	0,19	0,2	0,29	0,09	0,1
orig	0,64	0,56	0,43	0,19	0,5	0,48	0,36	0,93	0,37	0,41	0,42	0,3	0,54	0,37	0,14	0,28	0,31	0,04	0,07	0,06	0,22	0,18	0,09	0,15	0,07	0,15	0,2	0,13	0,09	0,13
orig	0,4	0,36	0,32	0,03	0,35	0,55	0,06	0,35	1	0,14	0,35	0,15	0,41	0,16	-0	0,13	0,12	0,02	-0	0,08	0,09	0,03	0,09	-0	0,06	-0	-0	-0	-0,1	-0,1
orig	0,26	0,46	0,64	0,3	0,16	0,12	0,1	0,32	0,09	0,79	0,5	0,44	0,31	0,58	0,4	0,52	0,41	0,37	0,15	0,19	0,29	0,23	-0	-0	0,07	-0	-0	0,05	0,13	0,12
orig	0,4	0,52	0,74	0,14	0,14	0,38	0,13	0,4	0,35	0,57	0,94	0,5	0,42	0,46	0,31	0,34	0,38	0,31	0,17	0,16	0,19	0,16	0,03	-0	-0	-0,1	-0	-0	0,03	0,02
orig	0,41	0,54	0,62	0,2	0,25	0,25	0,31	0,3	0,22	0,45	0,48	0,83	0,54	0,38	0,31	0,32	0,35	0,2	0,27	0,12	0,14	0,23	-0	0,01	-0	-0	0,02	-0	0,05	0,07
orig	0,43	0,73	0,47	0,08	0,4	0,4	0,38	0,52	0,39	0,33	0,4	0,49	0,89	0,35	0,24	0,38	0,2	0,01	0,13	0,07	0,16	0,16	0,03	-0,1	0,05	-0	0,01	-0	0,02	0,03
orig	0,17	0,35	0,3	0,25	0,1	-0	0,23	0,17	0	0,35	0,17	0,22	0,25	0,56	0,26	0,48	0,2	0,2	0,22	0,17	0,22	0,2	0	-0	0,13	0,02	-0	-0	0,14	0,15
orig	0,25	0,38	0,45	0,21	0,07	0,04	0,17	0,17	0,01	0,43	0,34	0,33	0,3	0,39	0,84	0,34	0,23	0,21	0,12	0,03	0,2	0,19	0,05	-0,1	-0	0,05	-0	0	0,12	0,1
orig	0,26	0,3	0,3	0,32	0,14	0,1	0,2	0,2	0,06	0,44	0,22	0,34	0,29	0,4	0,27	0,71	0,27	0,26	0,27	0,06	0,19	0,14	0,01	-0	0,1	0,04	-0	0,05	0,14	0,08
orig	0,22	0,21	0,35	0,25	0,14	0,1	0,26	0,22	0,12	0,44	0,31	0,28	0,15	0,33	0,19	0,33	0,69	0,78	0,62	0,34	0,47	0,39	-0	-0	0,03	-0,1	0,08	-0,1	0,04	0,05
orig	0,08	0,17	0,32	0,26	0,08	0,04	0,15	0,08	0,02	0,43	0,3	0,24	0,05	0,35	0,23	0,36	0,57	0,98	0,62	0,33	0,44	0,32	-0,1	-0,1	0,06	-0,2	0,06	-0,1	0,02	0,05
orig	0,02	0,08	0,22	0,11	0,04	-0	0,28	0,06	-0	0,2	0,14	0,23	0,1	0,18	0,09	0,31	0,31	0,59	0,92	0,43	0,3	0,28	-0,1	-0	0,04	-0,1	0,1	-0,1	0,04	-0
orig	0,15	0,19	0,24	0,12	0,11	0,08	0,19	0,09	0,13	0,23	0,18	0,15	0,09	0,2	-0	0,23	0,18	0,32	0,46	0,82	0,2	0,36	-0,1	0,09	0,06	-0	0,08	0,01	0,11	0,06
orig	0,15	0,17	0,2	0,21	0,14	0,04	0,27	0,17	0,07	0,35	0,15	0,1	0,13	0,3	0,22	0,27	0,34	0,49	0,41	0,3	0,73	0,27	0,02	-0	0,06	0,03	0,01	-0	0,09	0,07
orig	0,15	0,3	0,32	0,26	0,15	0,12	0,19	0,25	0,12	0,43	0,29	0,25	0,17	0,39	0,11	0,28	0,52	0,41	0,31	0,44	0,42	0,88	-0,1	-0	0,05	-0	0,1	-0	0,19	0,08
orig	0,13	0,06	0,04	0,06	0,17	0,07	0,09	0,09	0,09	-0	0,02	0,05	0,02	0,04	0,09	-0	-0	-0,1	-0,1	0,02	0,05	-0	0,88	0,31	0,07	0,22	0,09	0,24	0,26	0,31
orig	0,08	0,01	-0	0,12	0,02	0,01	0,11	0,14	0,01	0,01	-0	0,02	-0	0,06	-0	-0	-0	-0,1	0,04	0,05	-0	0,01	0,28	0,95	0,22	0,34	0,48	0,5	0,29	0,4
orig	0,08	0,03	-0	0,04	0,1	0,05	0,25	0,03	0,04	0,01	-0	0,01	0,01	0,09	0,01	0,02	-0	0,04	0	0,06	0,12	0,05	0,11	0,26	0,86	0,21	0,38	0,37	0,05	0,09
orig	0,1	-0	-0	0,08	0,04	0,01	0,22	0,13	-0	-0	-0,1	-0	-0	0,03	0,1	-0	-0,1	-0,2	-0,1	-0	0,02	-0	0,3	0,37	0,18	0,87	0,27	0,51	0,22	0,24
orig	0,1	0,09	-0	0,08	0,13	0,06	0,19	0,19	-0	0,02	0,02	0,03	0,04	0,05	0,1	-0,1	0,09	0,04	0,12	0,09	0,07	0,16	0,03	0,47	0,2	0,26	0,9	0,38	0,1	0,25
orig	0,12	0,04	0,02	0,12	0,07	0,06	0,21	0,1	-0	0,07	-0	-0	-0	0,06	0,04	0,01	0,01	-0,1	-0	-0	-0	0,03	0,23	0,55	0,27	0,52	0,46	0,87	0,16	0,42
orig	0,05	0,11	0,12	0,27	0,02	-0,1	0,06	0,08	-0,1	0,08	0,01	0,03	-0	0,11	0,15	0,07	0,05	-0	-0	0,07	0,08	0,21	0,29	0,33	0,09	0,22	0,17	0,14	0,95	0,5
orig	0,07	0,13	0,1	0,11	0,02	0,01	0,09	0,15	-0,1	0,07	0,02	0,02	0,03	0,17	0,13	0,06	0,06	0,01	0,04	-0	0,03	0,06	0,33	0,37	0,09	0,21	0,14	0,32	0,48	0,91

Figure 3.2: Heatmap of cosine similarity of the NTREX dataset using all-MiniLM-L6-v2

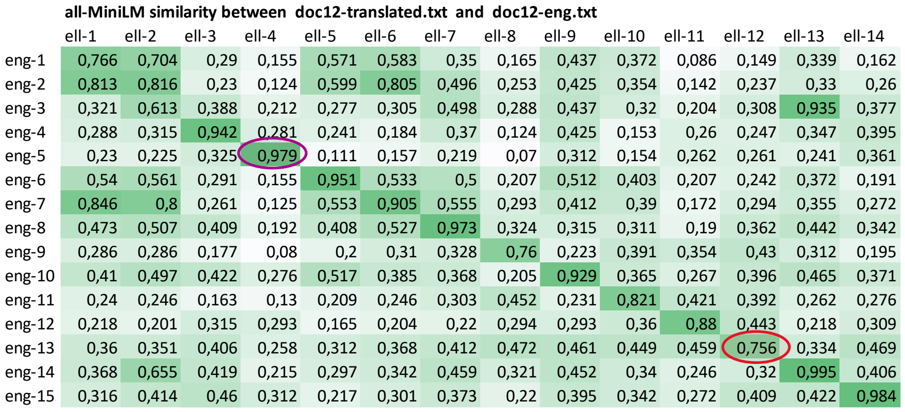
For cross-lingual tasks like bitext-mining, there are different models available, such as m-USE (Multilingual Universal Sentence Encoder), SBERT (Sentence-BERT) or LaBSE (Language-agnostic BERT Sentence Embeddings) (Feng *et al.*, 2022). In this study, the *all-MiniLM-L6-v2* and LaBSE models were considered and compared. The comparison was conducted on a sample of 5 articles from the collected article pairs; the results are detailed in Figure 3.3 and Figure 3.4 accordingly. The idea was to compare a SOTA sentence embedding model (see Section 3.2.2) to a smaller but efficient transformer model (see Footnote 6).



LaBSE similarity between doc12-ell.txt and doc12-eng.txt

	ell-1	ell-2	ell-3	ell-4	ell-5	ell-6	ell-7	ell-8	ell-9	ell-10	ell-11	ell-12	ell-13	ell-14
eng-1	0,84	0,602	0,161	0,046	0,294	0,574	0,278	0,019	0,065	0,237	0,002	0,072	0,265	0,19
eng-2	0,72	0,838	0,398	0,258	0,387	0,662	0,488	0,117	0,264	0,487	0,11	0,287	0,378	0,4
eng-3	0,177	0,487	0,343	0,199	0,327	0,233	0,383	0,086	0,404	0,305	0,213	0,291	0,761	0,387
eng-4	0,293	0,439	0,907	0,435	0,437	0,376	0,484	0,192	0,438	0,536	0,42	0,287	0,421	0,464
eng-5	0,102	0,244	0,376	0,853	0,233	0,19	0,355	0,119	0,366	0,178	0,217	0,292	0,222	0,32
eng-6	0,407	0,533	0,457	0,314	0,876	0,532	0,479	0,153	0,391	0,479	0,258	0,282	0,449	0,404
eng-7	0,658	0,632	0,291	0,206	0,415	0,91	0,439	0,142	0,234	0,414	0,18	0,311	0,272	0,341
eng-8	0,36	0,533	0,442	0,421	0,354	0,456	0,929	0,194	0,441	0,343	0,206	0,386	0,491	0,464
eng-9	0,021	0,098	0,128	0,059	0,167	0,15	0,189	0,848	0,112	0,228	0,302	0,223	0,129	0,036
eng-10	0,156	0,414	0,435	0,422	0,398	0,298	0,456	0,157	0,92	0,324	0,326	0,287	0,535	0,449
eng-11	0,239	0,362	0,402	0,12	0,344	0,363	0,253	0,113	0,246	0,851	0,406	0,34	0,304	0,36
eng-12	0,036	0,138	0,4	0,237	0,277	0,192	0,232	0,212	0,267	0,436	0,917	0,363	0,241	0,33
eng-13	0,223	0,365	0,312	0,313	0,301	0,355	0,408	0,219	0,377	0,387	0,405	0,824	0,358	0,465
eng-14	0,277	0,609	0,358	0,255	0,29	0,286	0,469	0,149	0,428	0,361	0,219	0,319	0,925	0,432
eng-15	0,275	0,483	0,397	0,344	0,323	0,347	0,465	0,042	0,45	0,425	0,288	0,392	0,431	0,925

Figure 3.3: Heatmap of LaBSE embeddings similarity example



all-MiniLM similarity between doc12-translated.txt and doc12-eng.txt

	ell-1	ell-2	ell-3	ell-4	ell-5	ell-6	ell-7	ell-8	ell-9	ell-10	ell-11	ell-12	ell-13	ell-14
eng-1	0,766	0,704	0,29	0,155	0,571	0,583	0,35	0,165	0,437	0,372	0,086	0,149	0,339	0,162
eng-2	0,813	0,816	0,23	0,124	0,599	0,805	0,496	0,253	0,425	0,354	0,142	0,237	0,33	0,26
eng-3	0,321	0,613	0,388	0,212	0,277	0,305	0,498	0,288	0,437	0,32	0,204	0,308	0,935	0,377
eng-4	0,288	0,315	0,942	0,281	0,241	0,184	0,37	0,124	0,425	0,153	0,26	0,247	0,347	0,395
eng-5	0,23	0,225	0,325	0,979	0,111	0,157	0,219	0,07	0,312	0,154	0,262	0,261	0,241	0,361
eng-6	0,54	0,561	0,291	0,155	0,951	0,533	0,5	0,207	0,512	0,403	0,207	0,242	0,372	0,191
eng-7	0,846	0,8	0,261	0,125	0,553	0,905	0,555	0,293	0,412	0,39	0,172	0,294	0,355	0,272
eng-8	0,473	0,507	0,409	0,192	0,408	0,527	0,973	0,324	0,315	0,311	0,19	0,362	0,442	0,342
eng-9	0,286	0,286	0,177	0,08	0,2	0,31	0,328	0,76	0,223	0,391	0,354	0,43	0,312	0,195
eng-10	0,41	0,497	0,422	0,276	0,517	0,385	0,368	0,205	0,929	0,365	0,267	0,396	0,465	0,371
eng-11	0,24	0,246	0,163	0,13	0,209	0,246	0,303	0,452	0,231	0,821	0,421	0,392	0,262	0,276
eng-12	0,218	0,201	0,315	0,293	0,165	0,204	0,22	0,294	0,293	0,36	0,88	0,443	0,218	0,309
eng-13	0,36	0,351	0,406	0,258	0,312	0,368	0,412	0,472	0,461	0,449	0,459	0,756	0,334	0,469
eng-14	0,368	0,655	0,419	0,215	0,297	0,342	0,459	0,321	0,452	0,34	0,246	0,32	0,995	0,406
eng-15	0,316	0,414	0,46	0,312	0,217	0,301	0,373	0,22	0,395	0,342	0,272	0,409	0,422	0,984

Figure 3.4: Heatmap of all-MiniLM embeddings similarity example

As illustrated in these examples, the two models exhibit comparable performances. The scores circled in the two heatmaps represent instances where the models differ significantly. Table 3.1 provides two examples illustrating these observations.

Table 3.1 Examples of similar sentences from AMNA corpus

ell-12	*Η Ευρωπαϊκή Ένωση οφείλει να κυνηγήσει όλους όσους εμπλέκονται στο κύκλωμα της παράνομης διακίνησης,* ***όπως οι κατασκευαστές των φθηνών φουσκωτών».***
trans-12	*The European Union must go after all those involved in the smuggling ring,* ***such as the manufacturers of cheap inflatables."***
eng-13	*The European Union must go after all those involved in the illegal trafficking circuit."*

ell-4	*Χθες βράδυ, μόλις έφτασα, συναντήθηκα με την ομάδα του Ευρωπαϊκού Λαϊκού Κόμματος, στην οποία ανήκει η ΝΔ, και αποτελείται από ομολόγους μου από την υπόλοιπη Ευρώπη και τον αντιπρόεδρο της Ευρωπαϊκής Επιτροπής, κ. Μαργαρίτη Σχοινά.*
trans-4	*Last night, as soon as I arrived, I met with the Group of the European People's Party, to which New Democracy belongs, and consists of my counterparts from the rest of Europe and the Vice-President of the European Commission, Margaritis Schinas.*
eng-5	*Last night, as soon as I arrived, I met with the group of the European People's Party, to which the New Democracy belongs, and which consists of my counterparts from the rest of Europe and the Vice-President of the European Commission, Margaritis Schinas.*

LaBSE considers the *ell-12* and *eng-13* (not exact translations) more similar than all-MiniLM does, assigning a cosine similarity score of 0,824 compared to 0,756. On the contrary, all-MiniLM considers *trans-4* and *eng-5* (exact translations) more similar than LaBSE does, assigning a cosine similarity score of 0,979 compared to 0,853. These results highlight LaBSE's ability to identify parallel sentences that are not exact translations of each other. While the performance of all-MiniLM was comparable to LaBSE, the latter was selected for the ensuing experiments as it aligns better with the scope aims of the research as detailed below.

The LaBSE model is indeed a sophisticated model derived from the BERT model (Devlin *et al.*, 2019) trained on a corpus encompassing 109 languages. LaBSE's dual-encoder architecture and its training in on translation pairs make it an excellent choice for extracting exact translations. However, its performance is compromised when the two sentences are not direct translations of each other.

LaBSE's ability to encode sentences across languages is central to its selection for this project. The model's state-of-the-art performance in cross-lingual sentence similarity and retrieval tasks aligns perfectly with the main objectives of this study, which involves extracting both parallel and semi-parallel sentences from bilingual news articles.

In this system, LaBSE embeddings are the primary input for the subsequent process of similarity assessment. Its multilingual vector space allows for the generation of the vector representation of the sentences in the bilingual dataset of news articles and for a comparison of the vectors using the cosine similarity metric to determine the similarity between sentences.

3.3 Classification System

The cosine similarity score is an effective measure for assessing the similarity between vectors and plays a crucial role here in comparing sentence embeddings. Specifically, the system measures the similarity between the embedding of a given source sentence in Greek and the embeddings of all target sentences in English, employing cosine similarity for this purpose.

The classification method utilized in this study employs cosine similarity thresholds to differentiate between various levels of sentence similarity. This method is straightforward and intuitive, allowing for a simple categorization of sentence pairs based on their similarity scores. Despite its simplicity, this approach may introduce a degree of ambiguity which is expressed through a mediocre similarity score.

A relevant application of a similar method is that of Feng et al. (2022). They explore parallel text mining from the CommonCrawl dataset, using a binary classification based on an arbitrary cosine similarity threshold of 0.6. In their methodology, sentence pairs scoring above this threshold are deemed parallel, while those below it are considered non-parallel.

The present system compares the vector of a given source (Greek) sentence to all the vectors of the target (English) sentences within a given document and calculates the cosine similarity. This approach is based on a two-stage classification scheme that includes three primary and two secondary classification labels, going beyond a binary classification. The first stage of the system implementation classifies sentence pairs into three categories based on the similarity of their embeddings: *Parallel*, *Ambiguous*, and *Unrelated*.

Following this approach, when the system assigns the “ambiguous” label to a sentence pair during the first stage, the process does not stop there. Instead, a more refined investigation is conducted in the second stage, which further scrutinizes the ambiguous cases. The ambiguous sentence pairs are classified as either “partial translation” or “non-translation”, providing a more specific relationship between the source and target sentences.

3.3.1 First-stage Classification: Assessing Sentence Embeddings Similarities

The introduction of the ambiguous category serves to mitigate the limitations inherent in a binary classification system by offering a middle ground for sentence pairs that do not fit as either strictly parallel or entirely unrelated. This approach, facilitated by the implementation of two thresholds instead of one, aims to significantly enhance the precision in assigning the two original binary labels (parallel and unrelated). This refinement is crucial in the context of analysing news production processes where the relationship between similar documents may not be immediately evident.

Understanding the process of news production becomes particularly important when the nature of the relationship between two documents is unknown. In scenarios where the source of an article is not transparent, it can be challenging to discern the exact nature of the relationship between two articles. In this study, the assumption is made that English articles published on the AMNA website, which have Greek counterparts, are likely derived from these Greek versions. Without clear insight into the production process, the nature of the English versions—whether they are direct translations, paraphrases, or summaries of the Greek originals—remains ambiguous.

This uncertainty extends to the resources available to the journalists crafting the English texts; access to the Greek "originals" or other sources could lead to English articles emerging through direct translation, paraphrasing, or summarization processes, as explored by Bernardini et al. (forthcoming).

Researchers have tried to categorize the relationship of articles in the framework of text reuse focusing on whether a document is derived from another document. In their work, Clough et al. (2002, p. 1680), provide three categories to describe the equivalence relationship between two news texts at a word level:

verbatim: text appearing word-for-word to express the same information;

rewrite: text paraphrased to create a different surface appearance, but express the same information and

new: text used to express information not appearing in agency copy (can include verbatim/rewritten text but being used in a different context).

Based on the journalistic text-reuse categories defined above, an attempt was made to adapt them to this study's purpose[8]. Specifically, the selected classification categories were defined as follows:

Parallel: The sentence pair consists of two sentences that are semantically and structurally similar, and hence likely to have resulted from translation processes.

Ambiguous: The sentence pair exhibits a certain degree of semantic similarity, yet this does not necessarily imply a direct translational relationship between the two sentences. The sentences are semantically close but not explicitly connected or related in context.

Unrelated: The two sentences are semantically and contextually dissimilar.

The initial classification system distinguishes between parallel, ambiguous, and unrelated sentence pairs based on predefined similarity thresholds. Parallel sentences require a similarity score above 0.8, whereas those scoring below 0.6 are deemed unrelated. Scores falling between these thresholds categorize a pair

[8] During the initial stages of the project, fifteen journalists and people associated with AMNA were contacted for insight into the agency's process regarding the production of articles in two languages. Only one person replied claiming that the agency employs journalists, correspondents, and translators for multilingual news production. Unfortunately, this information does not provide a clear image of the process in place.

as ambiguous. In the comparison phase, for each Greek sentence, only the top two matches are identified, evaluated for similarity, and subsequently classified into one of the three categories. The decision to extract more than one similar sentence is based on the fact that a news source text undergoes several editing processes, one of which consists in splitting a single sentence into multiple ones. Extracting the top two matches ensures that the model detects the majority of similar relevant sentences between two documents.

The thresholds were informed by the distribution of the cosine similarity scores which are detailed in Chapter 4. The cosine similarity score distribution (see Figure 4.1) indicates that most similarity scores fall between 0.4 and 0.6. This is understandable given the decision to include the second most similar sentence for each source sentence. While there is a clear distinction between unrelated and ambiguous sentences, the distinction becomes less clear between parallel and ambiguous ones.

3.3.2 Second-stage Classification: Assessing Word Embedding Similarities

There are different approaches one can take to disambiguate the relationship between two similar sentences. For this study, the decision was made to limit this second classification stage to a single parallel sentence extraction task. The ambiguous pairs can be (dis)similar in several ways.

In the context of bilingual news production, two sentences can be related to each other because one derives from the other. The process that resulted in the two similar sentences can be direct translation as is the case for parallel pairs, or it can be the case that journalistic choices reduce the similarity of two sentences without completely wiping out the similarity between them. This can be the effect of paraphrasing, rewriting or even patchwork (Davier, 2014). The latter technique consists in using information from diverse sources and combining them, thus, resulting in sentences that are similar up to a point, but then deviate. Another approach usually adopted by journalists is the omission or addition of information deemed (un)necessary for the new audience of the article.

To try and capture these cases, the “partial translation” and “non-translation” labels are created and assigned by the system comparing word embeddings using cosine similarity and specific classification rules. The first label identifies sentence pairs that, while not exact translations, show obvious signs that one is derived from the other, suggesting a connection beyond mere coincidence. These could include cases where the essence and key points are shared, albeit with differences in wording and structure, or cases where the target sentence is a “partial translation”, meaning that there has been an addition to or an omission from the source text, based on the actual source text or translation direction which is not assumed in this process. The second label is for sentence pairs that share semantic similarities but do not suggest one is directly derived from the other. This category includes sentences that might contain overlapping vocabulary but cover different events or instances of the same event.

This two-step approach allows for a more granular analysis of the content of the compared sentences offering more flexibility in defining the type and degree of similarity between two sentences. While the sentence embeddings comparison offers a single similarity score per pair, this approach offers a more flexible interpretation of the word embeddings similarity scores based on manually crafted rules.

The system leverages the Google News word2vec embeddings which is a pre-trained model developed by Tomas Mikolov et al. at Google (Mikolov *et al.*, 2013). This model is trained on approximately 100 billion words from a Google News dataset, resulting in a high-dimensional space of word vectors. Each vector has 300 features and contains linguistic information about a word, which can be useful in tasks like this one. In this research, this model was chosen for its domain specificity, closely matching the dataset utilized. Its computational efficiency and ease of use significantly contributed to its selection.

Word embeddings can be used to perform word-level comparisons and calculate the similarity between pairs of words. The word embeddings model used is trained on English data, thus allowing for comparison between English words only. This was addressed by using an MT engine, specifically Google Cloud Translator API, to translate Greek sentences into English to prepare the data for the Google News word2vec model which is trained only on English data and thus, generates accurate embeddings for specifically for English texts. To obtain word embeddings for the English sentence pairs, the Gensim[9] library is utilized.

During this classification phase, the system evaluates the word embeddings from one sentence against those of another. This process is specifically applied to sentence pairs initially deemed ambiguous, with the aim of clarifying their status by examining the similarity between individual word pairs.

The system measures the cosine similarity between vectors representing each word in the source sentence and those in the target sentence and extracts the pairs with the highest similarity scores. These scores provide insights into the cumulative extent of word similarity between two sentences. However, the mere presence of similar words does not necessarily mean that the sentences are partial translations.

To refine this assessment, the second classification involved sentence pairs that were classified as ambiguous by the first one. This classification is based on the cosine similarity between the words in the paired sentences. The system uses a predefined similarity score threshold to categorize word pairs as either

[9] The documentation for the Gensim library is available at https://radimrehurek.com/gensim/models/word2vec.html.

‘matching’ or ‘non-matching’. The system compares the two paired sentences in both directions, counts the matching words and calculates the matching ratio considering the higher ratio as the translation direction. A second threshold is employed to classify the sentence pairs as partial translation or non-translation based on the matching ratio. The process aims to determine if portions of one sentence are replicated in the other.

Two words are considered a match if their cosine similarity is above 0.5. The threshold was determined through empirical analysis and was chosen to account for synonyms. This threshold was informed by the distribution of word embeddings similarity displayed in Figure 3.5, where the central tendency of the distribution is around 0.5. To further evaluate the adequacy of the threshold, a manual analysis was conducted on randomly selected word embeddings heatmaps to check whether words with a similarity near the threshold were indeed similar. This is particularly important in the context of this study, where synonyms are common due to journalistic choices and to the machine translation step used in the word embeddings comparison.

The system labels a sentence pair as a partial translation or non-translation based on the ratio of matching words. The experiment utilizes a matching words ratio threshold of 0.6. This means that a sentence should have at least 60% of its words present in the other sentence. The threshold was informed by manual error analysis and aims to account for marginal cases of partial translations. The system prioritises identifying as many instances of partial translation as possible. This is known as higher recall, over ensuring every identified case is correct, known as higher precision.

In this study, we have opted for a strategy that favours the inclusion of as many partial translations as possible, aiming for higher recall. This approach supports the goal of the study to enrich corpora by including partial translations and construct a corpus where matching sentences are accurately aligned and readily accessible to researchers, making them more valuable for Translation Studies researchers. Despite the potential increase in False Positives, where sentences are incorrectly identified as partial translations, this is seen as a reasonable trade off within the scope of the research goals. Exposing researchers to some irrelevant data is worth it to ensure they do not miss out on relevant data.

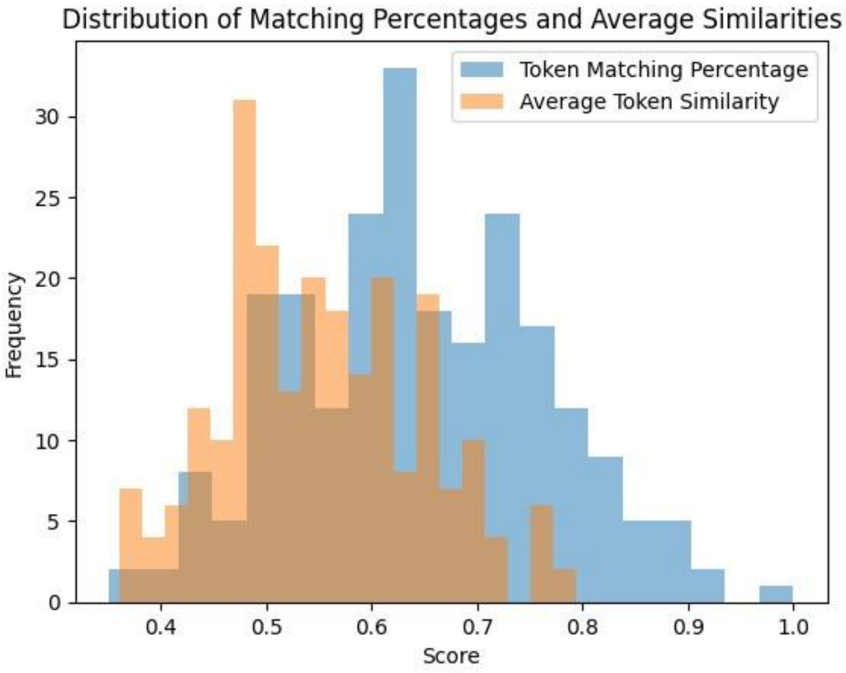


Figure 3.5: Distribution of Matching Words Ratio and Average Word Similarities

3.3.3 Performance Evaluation

The first evaluation round involved 148 sentence pairs, representing 10% of the total, which were categorized as parallel, ambiguous, and unrelated. For the second round, 47 pairs, or 20% of the ambiguous pairs, were examined to classify them into Partial Translations and Non-Translation. Both samples were extracted using the *train_test_split*[10] function to apply stratified sampling based on the automatic classification labels. Due to time constraints and a limited

[10] Documentation available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

number of available experts, the annotation was done solely by the author of this thesis project. As the annotator of this study, I adhered to specific guidelines that provided detailed definitions and examples for each label, ensuring accurate and consistent categorization of sentence pairs.

In summary, the two-stage classification introduced in this research represents a novel step in the analysis of sentence pair categorization, moving beyond the traditional binary threshold approach. It aims to do so by introducing an intermediate threshold range for sentence pairs that fall within a grey area. These are then further evaluated through two interconnected binary criteria: “matching words similarity” and “matching ratio.”

While this step may not completely clarify the exact type of relationship of all the extracted sentence pairs, it aims to disambiguate the initial similarity scores obtained via sentence embeddings. Moreover, this method increases the amount of parallel data and provides contextual insights, offering valuable resources for future research. The results of the implementation of the system and the evaluation of the performance on the task will be analysed in Chapter 4.

4 Results and Discussion

The system implementation presented in Chapter 3, yielded some interesting results which will be discussed and analysed in detail. The system design allows for several considerations to be made regarding the different components utilized and their implications for the results. This chapter presents a comprehensive analysis of the approach and methodology followed while it also ties back to the research questions and objectives of this thesis project.

Section 4.1 provides a summary of the system's data and performance results. In Section 4.2, an error analysis is conducted on a sample of the classification output data, with a focus on misclassified pairs. Lastly, Section 4.3 presents considerations on the performance and limitations of the components and strategies used.

4.1 Data Summary and Performance Metrics

For the experiments conducted in this thesis project, a corpus of 41 article pairs was used, each comprising a Greek and an English version. After initial preprocessing, which involved tokenization, normalisation and translation, a total of 734 Greek and 351 English sentences were prepared for further analysis.

4.1.1 First-stage Classification: Results

The system's algorithm was designed to extract and compare the top two matches for each Greek sentence, thereby resulting in a dataset of 1,468 sentence pairs for classification. These pairs were categorized as parallel, ambiguous, or unrelated following the proposed approach. This classification was based on a set of predefined rules leveraging the similarities between sentence embeddings.

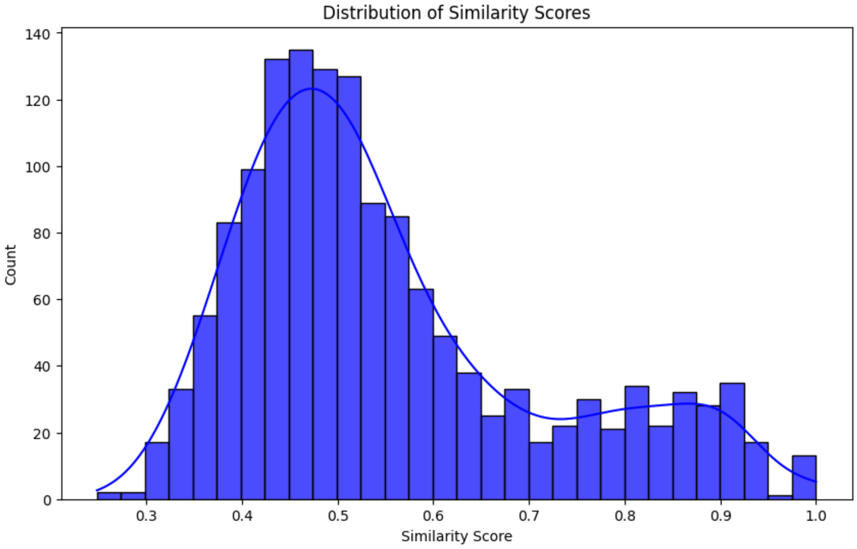


Figure 4.1 Distribution of Sentence Similarity Scores

The histogram presented in Figure 4.1 provides a visual representation of the distribution of the similarity scores produced by the sentence embedding comparison. The bell-shaped curve of this histogram suggests a right skewed distribution, where a substantial number of sentence pairs demonstrate moderate levels of similarity, primarily clustered around the 0.4 to 0.6 score range. In contrast, the distribution's tails indicate a lower frequency of sentence pairs for the extremes of the similarity scores. This variation is reflective of the dataset's inherent complexity with respect to the semantic similarity task, with most pairs falling within the ambiguous similarity area.

The central tendency represented by the peak of the histogram can serve as a reference for setting classification thresholds. An excessively high threshold for the unrelated category risks overlooking a substantial partition of moderately similar sentence pairs, potentially misclassifying them as ambiguous. On the other hand, a low threshold for the parallel category may result in an inaccurate labelling of distinctly dissimilar sentences as parallel.

Figure 4.2 complements the histogram by illustrating the distribution of sentence pairs post-classification. It offers a concise visualization of the prevalence of each label within the dataset, thereby facilitating an immediate grasp of the classification outcomes. This distribution plays a critical role in fine-tuning the classification thresholds; however, by itself, it is not highly informative.

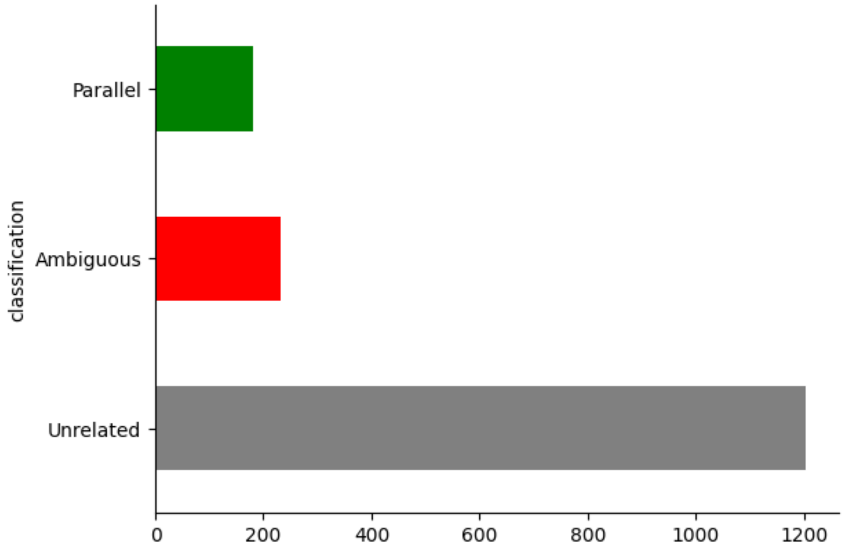


Figure 4.2: Label Distribution of 1st-stage Classification

The bar chart shows that most pairs are labelled as unrelated, indicating that their sentence similarity score was below 0.6. The specific distribution is reported in Table 4.1.

Table 4.1 Label distribution by rank

classification	rank1	rank2	total
parallel	168	14	182
ambiguous	139	94	233
unrelated	427	626	1053
all	734	734	1468

This outcome was anticipated due to the selection of the top two matches for each source sentence. Figure 4.3 illustrates this point, showing that very few parallel pairs are identified when only the second-best matching sentence is considered. Nevertheless, the 15 instances labelled as parallel in the *rank2* category still warrant further analysis.

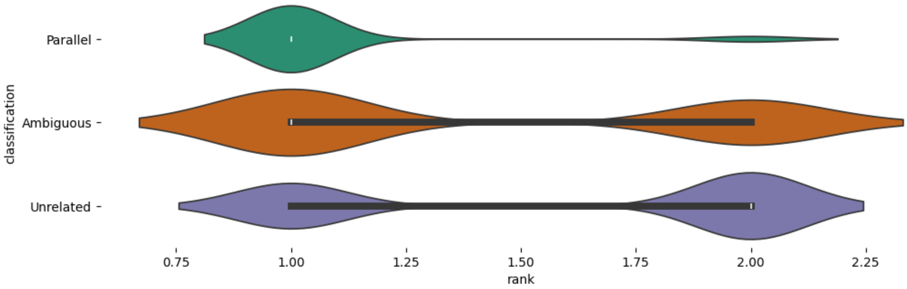


Figure 4.3 Violin Plot of Sentence Pair Rank Distributions by Classification Category

The evaluation of the initial automatic classification involved a detailed comparison with manually classified labels to establish the accuracy of our system. To do this, a subset of 148 sentence pairs, representing 10% of our dataset, was selected through stratified sampling to maintain the original distribution of labels. This subset was manually classified without prior knowledge of the automatic labels to ensure an unbiased comparison. Subsequently, the manually assigned labels served as ground truth to evaluate the correspondence and accuracy of the automatic classification system, allowing us to assess its performance through the metrics presented in Table 4.2. For this project, one annotator manually classified the data (see Section 4.1.1). Ideally, multiple annotators should perform manual evaluations, and the inter-annotator agreement would be measured to create the truth labels dataset. This was not possible for this thesis project due to lack of expert annotators.

Table 4.2 Evaluation metrics results

Metric	*Score*
Accuracy	0.8571
Macro Precision	0.7938
Macro Recall	0.7749
Macro F1 score	0.7817
Weighted Precision	0.8658
Weighted Recall	0.8571
Weighted F1 score	0.8603

The results in Table 4.2 show an overall high accuracy. However, there is a significant difference between the weighted metrics and the macro metrics. Weighted metrics consider the most frequent class as the most important one. In this dataset, the unrelated class is by far the most frequent label. This indicates that the unrelated class highly influences the weighted metrics, inflating the scores. On the other hand, the macro metrics consider all classes as equally important, thus providing a comprehensive and more accurate representation of the systems performance.

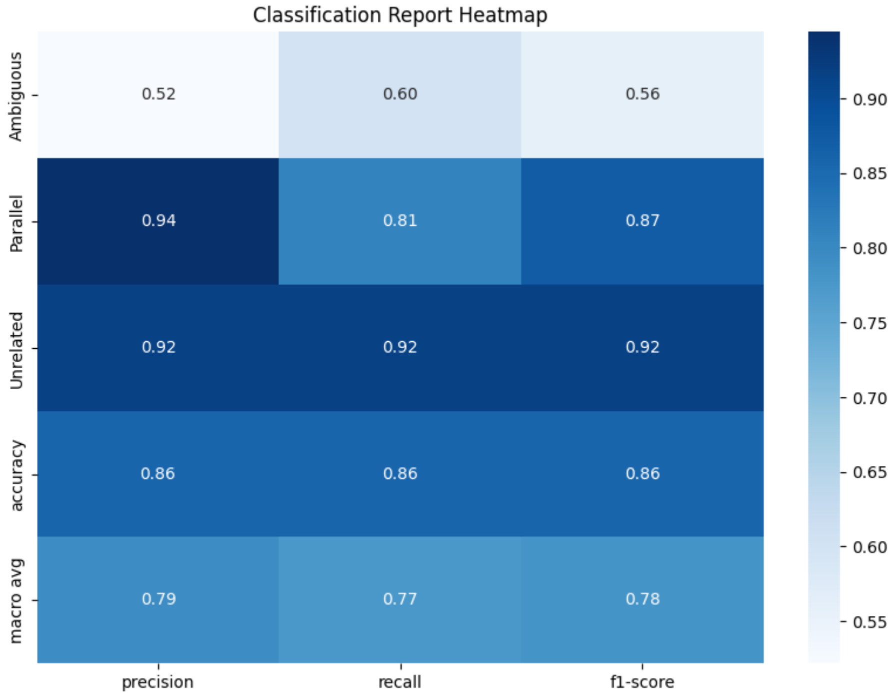


Figure 4.4 Classification Report Heatmap

The heatmap illustrated in Figure 4.4 provides the precise metrics for each category, allowing a nuanced examination of the system's performance. The results show that the system performs well in identifying unrelated and parallel sentences, but struggles with the ambiguous category, often misclassifying sentences that a human evaluator would identify as either parallel or unrelated. The issue is caused by the threshold settings employed during the classification process. Achieving a balance in accuracy between ambiguous sentences and more clearly defined parallel and unrelated categories is dependent on these threshold values.

The system's primary objective is to extract parallel sentences from comparable documents. The initial classification step achieves this goal with precision. The system's design enables a more in-depth analysis of ambiguous sentences, allowing for the extraction of parallel or quasi-parallel pairs through a two-stage classification process. It also enables the examination of pairs that may not appear similar initially but could be analysed further and included in the comparallel corpus as examples of related yet dissimilar sentences resulting from the journalistic production process.

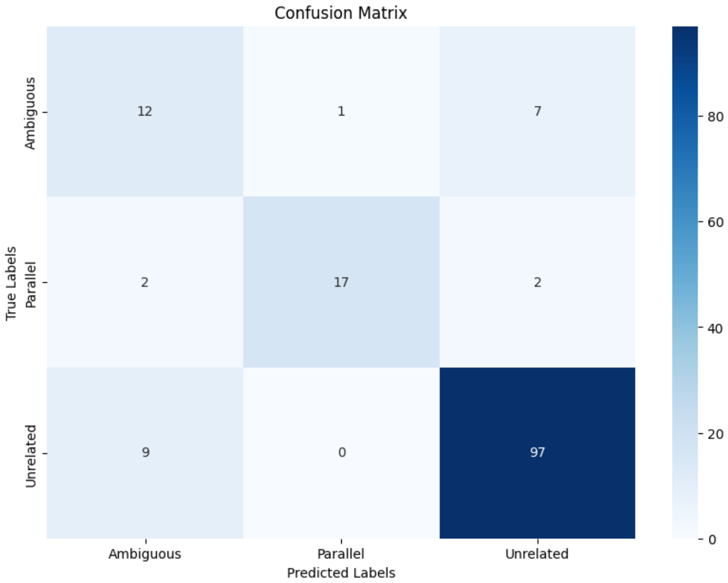


Figure 4.5 Confusion matrix of 1st-stage Classification

Finally, the confusion matrix in Figure 4.5 provides even more detailed information about the performance of the system. The matrix indicates that the system struggles to differentiate between unrelated and ambiguous sentence pairs, as evidenced by the 9 false negatives (FN) for the unrelated label and the 7 false positives for the ambiguous label. This difficulty may also reflect the challenges faced by human annotators in making these distinctions.

Due to the two-step nature of our automatic classification process and our approach's tolerance for false positives, we aim for inclusiveness in our classifications. Therefore, when uncertain, it may be more beneficial to classify a sentence pair as ‘ambiguous’ rather than ‘unrelated’. This recommendation is in line with the design of our system, which uses the word embedding step to further refine these preliminary classifications, optimising for a broader capture of potentially relevant sentence pairs.

4.1.2 Second-stage Classification: Results

The system's performance is closely linked to the thresholds for word cosine similarity and the ratio of matching words. Therefore, conducting various experiments with different thresholds is crucial to find an optimal balance that aligns with the classification objectives. In our task, the goal is to extract good instances of parallel sentences that could be included in a comparallel corpus to advance Translation and Journalistic studies. Figure 4.6 provides insight into the matching ratio distribution throughout the ambiguous sentence dataset.

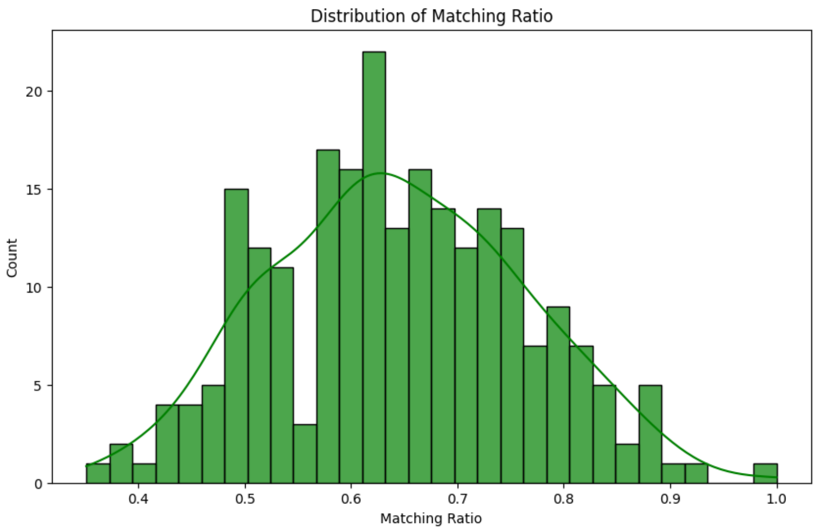


Figure 4.6 Distribution of Matching Words Ratio

The histogram indicates that the matching ratios are distributed normally, and the chosen thresholds strike a good balance between the two labels. The automatic classification was performed on 233 ambiguous sentences, and Figure 4.7 reports the resulting label distribution. The data shows that the partial translation label is more prevalent. This is expected because the label is ambiguous and includes sentence pairs with a cosine similarity score ranging from 0.55 to 0.75, as determined by sentence embeddings.

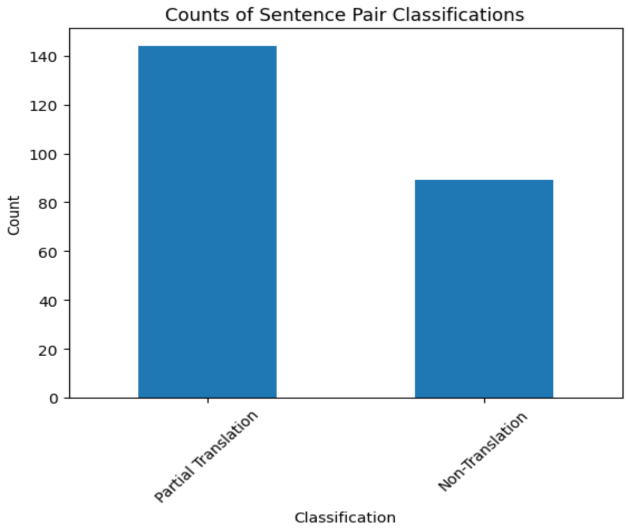


Figure 4.7 Label Distribution of 2nd-stage classification

To assess the system's performance on the second classification task, a sample of 20% (47) of the total number of pairs was extracted from the population of ambiguous sentences (233) using stratified sampling. A manual evaluation of the sample was conducted, which served as a reference dataset for the system's performance evaluation. The evaluation results report and heatmap are presented in Table 4.3 and Figure 4.8 respectively.

Table 4.3 Evaluation metrics for 2nd classification

Metric	Score
Accuracy	0.8085
Weighted Precision	0.8111
Weighted Recall	0.8085
Weighted F1 score	0.8064

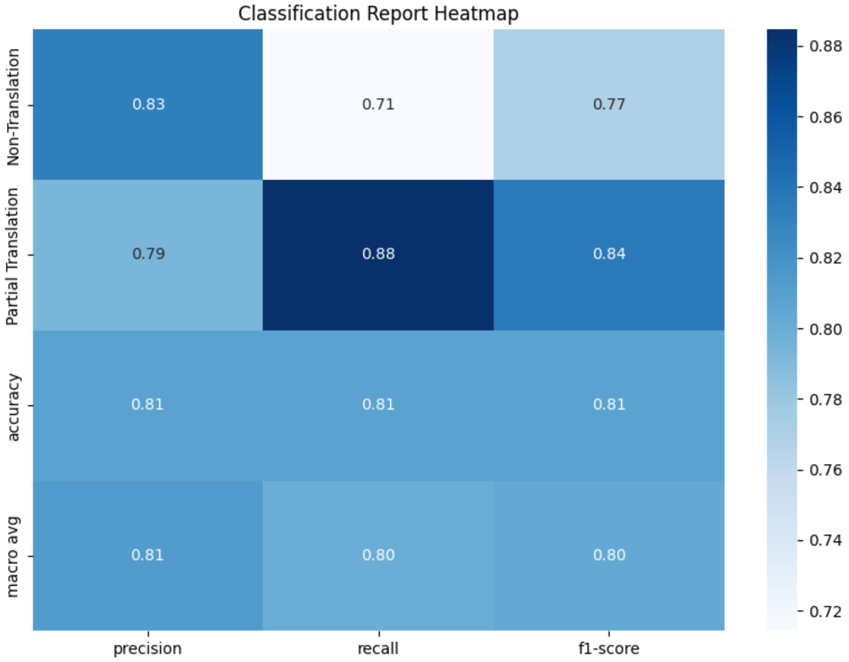


Figure 4.8 Evaluation Metrics Heatmap

The heatmap illustrates the performance of the classification system, displaying precision, recall, and F1-score, which are crucial metrics for evaluating its effectiveness. The system performs well for both classes, with a macro average of 0.81 for precision, 0.80 for recall, and 0.80 for F1-score.

To provide a more detailed evaluation of the system's performance, we created a confusion matrix based on the sample evaluation. This matrix shows where the system makes mistakes and provides additional information beyond the overall evaluation. Figure 4.9 displays the results, which indicate that the system occasionally misclassifies pairs as partial translations when they are not actually partial translations of each other.

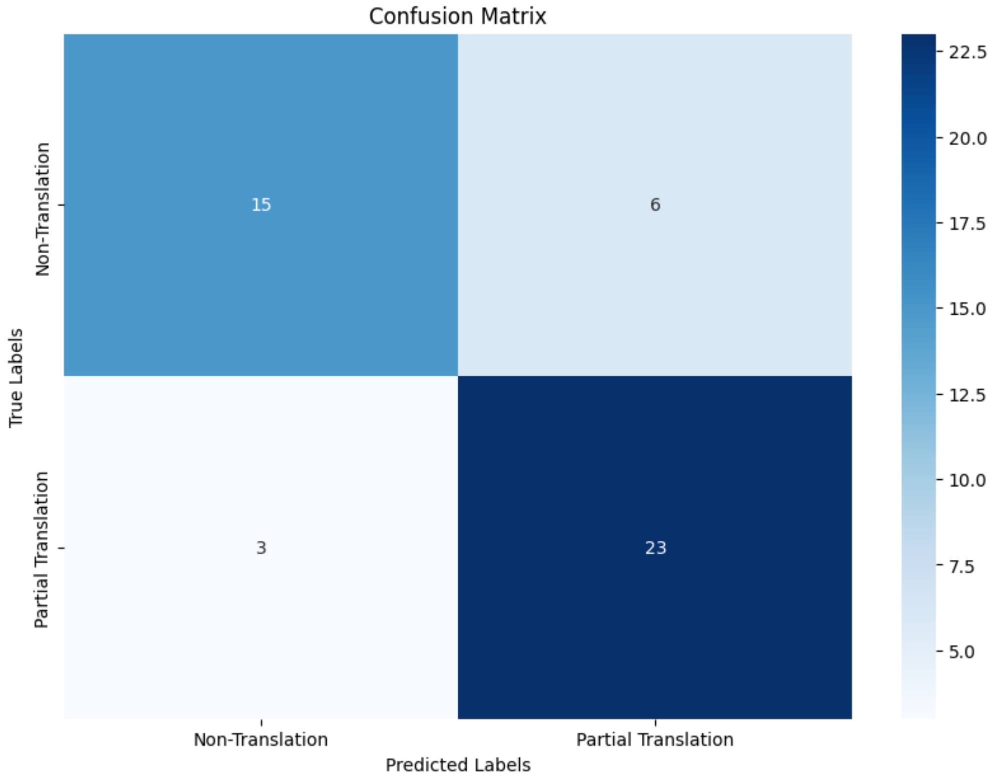


Figure 4.9 Confusion matrix for the 2nd-stage Classification

To enhance comprehension of the causes of these errors, an error analysis was conducted, focusing on misclassified pairs, and attempting to extract useful information to improve the system or understand its limitations.

4.2 Error Analysis

Although a systematic framework was used, misclassifications did occur. The following examination explores these cases intending to highlight both the effectiveness and the limitations of the methodology employed. The analysis is based on the classification results presented in Section 4.1.2 and is instrumental in revealing opportunities for refinement of the classification criteria, thereby improving accuracy, or identifying areas where further research is required. Given the system's high accuracy on the 1st classification task, the error analysis will primarily focus on the 2nd classification task. A manual analysis of 2 instances of the 1st classification and 5 misclassified instances for the 2nd classification is presented below to highlight the success and explore the potential limitations of the system.

To facilitate the comprehension of the examples, all sentences are in English. The Original Greek sentences are under the header "Greek" while the translated sentences from Greek into English for the purposes of the 2nd classification stage are marked as "Translated from Greek".

Example n.1:

1st stage misclassification

Greek	English
The body will be transported to the port of Kalamata.	**All of the yacht's passengers will be taken to Kalamata port**.

Match Ratio	Predicted Label	Truth Label
0,678	Ambiguous	Unrelated

This example of a misclassification by the system during the 1st classification stage highlights one of the shortcomings of the approach when dealing with short sentences. When the segment of the sentence that carries the contextual discrepancy is too short, it fails to provide sufficient information for the system to recognize the difference. In other words, when two short sentences describe the same event but differ for example in the subject, the two sentences are still not considered unrelated. While the similarity score assigned by the sentence embeddings comparison seems correct, the system needs more nuanced information to distinguish between ambiguous and unrelated in these cases.

Example #2:

1st stage misclassification

Greek	English
Once agreed, **Frontex** will ask **other** countries to **immediately** provide **border guards** and other relevant personnel, the statement concludes.	**Frontex** has already increased surveillance capacity at the Greek **borders** and is redeploying **officers** from **other** operations to provide **immediate** assistance.

Match Ratio	Predicted Label	Truth Label
0,611	Ambiguous	Unrelated

This second example illustrates a borderline situation where the sentences have some overlapping parts, yet their differences in meaning are substantial enough for the annotator to consider them distinct. Consequently, such an example would not be suitable for categorization as a Partial Translation in the second classification phase and would be excluded from the final corpus.

Example #3:

2nd stage correct classification

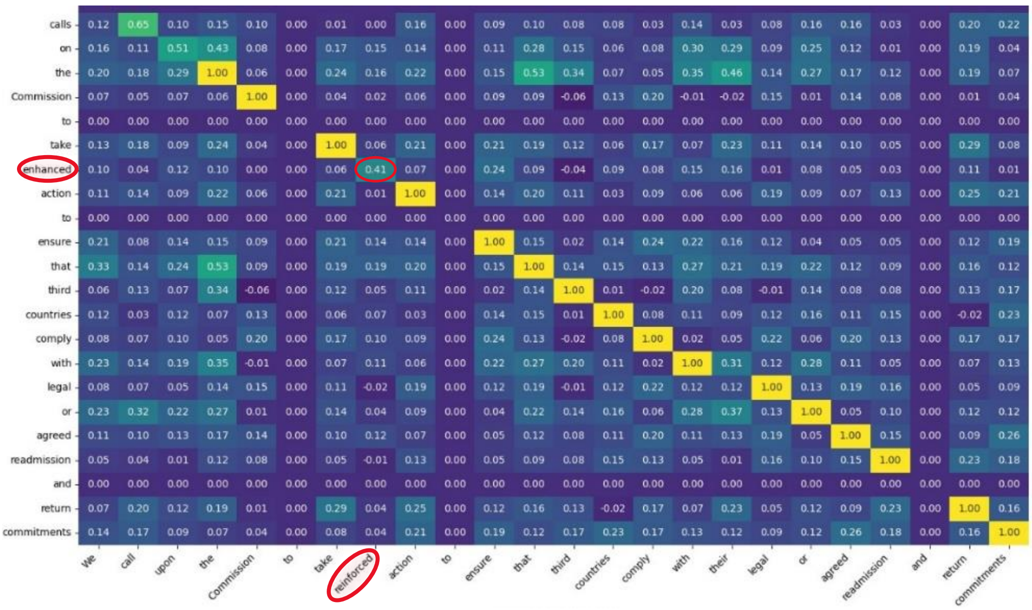
Translated from Greek	English
It highlights - inter alia - the importance of returns as a key pillar of EU migration policy, including voluntary assisted returns, and **calls on the Commission to take enhanced action to ensure that third countries comply with legal or agreed readmission and return commitments.**	We **call upon the Commission to take reinforced action to ensure that third countries comply with their legal or agreed readmission and return commitments.**

Match Ratio	Predicted Label	Truth Label
0,875	Partial Translation	Partial Translation

This first example from the 2nd classification stage is a True Positive, displaying the system's ability to detect parallel chunks of text in two sentences that are not exactly parallel. The English sentence is identified by the system as part of the translated sentence. This is a prime example of a partial translation which showcases a frequently used translation technique in journalistic text, i.e., addition (or omission).

This specific case is an excellent example as it accurately showcases a partial translation, but it also represents a case where probably the source text was probably not written in Greek but in English, further confirming the necessity of a comparallel approach for the study of news translation. The heatmap in Figure 4.10 depicts the similar part of the two sentences.

A specific word pair that is worth discussing is the match between “reinforced” and “enhanced”. This pair has a similarity of 0,41 which is low enough to not be considered a match. Both words are followed by the noun “action” and are clearly synonyms of each other, however the embeddings models did not detect this similarity, probably because it does not consider the surrounding context.



	we	call	upon	the	Commission	to	take	reinforced	action	to	ensure	that	third	countries	comply	with	their	legal	or	agreed	readmission	and	return	commitments
calls	0.12	0.65	0.10	0.15	0.10	0.00	0.01	0.00	0.16	0.00	0.09	0.10	0.08	0.08	0.03	0.14	0.03	0.08	0.16	0.16	0.03	0.00	0.20	0.22
on	0.16	0.11	0.51	0.43	0.08	0.00	0.17	0.15	0.14	0.00	0.11	0.28	0.15	0.06	0.08	0.30	0.29	0.09	0.25	0.12	0.01	0.00	0.19	0.04
the	0.20	0.18	0.29	1.00	0.06	0.00	0.24	0.16	0.22	0.00	0.15	0.53	0.34	0.07	0.05	0.35	0.46	0.14	0.27	0.17	0.12	0.00	0.19	0.07
Commission	0.07	0.05	0.07	0.06	1.00	0.00	0.04	0.02	0.06	0.00	0.09	0.09	-0.06	0.13	0.20	-0.01	-0.02	0.15	0.01	0.14	0.08	0.00	0.01	0.04
to	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
take	0.13	0.18	0.09	0.24	0.04	0.00	1.00	0.06	0.21	0.00	0.21	0.19	0.12	0.06	0.17	0.07	0.23	0.11	0.14	0.10	0.05	0.00	0.29	0.08
enhanced	0.10	0.04	0.12	0.10	0.00	0.00	0.06	0.41	0.07	0.00	0.24	0.09	-0.04	0.09	0.08	0.15	0.16	0.01	0.08	0.05	0.03	0.00	0.11	0.01
action	0.11	0.14	0.09	0.22	0.06	0.00	0.21	0.01	1.00	0.00	0.14	0.20	0.11	0.03	0.09	0.06	0.06	0.19	0.09	0.07	0.13	0.00	0.25	0.21
to	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ensure	0.21	0.08	0.14	0.15	0.09	0.00	0.21	0.14	0.14	0.00	1.00	0.15	0.02	0.14	0.24	0.22	0.16	0.12	0.04	0.05	0.05	0.00	0.12	0.19
that	0.33	0.14	0.24	0.53	0.09	0.00	0.19	0.19	0.20	0.00	0.15	1.00	0.14	0.15	0.13	0.27	0.21	0.19	0.22	0.12	0.09	0.00	0.16	0.12
third	0.06	0.13	0.07	0.34	-0.06	0.00	0.12	0.05	0.11	0.00	0.02	0.14	1.00	0.01	-0.02	0.20	0.08	-0.01	0.14	0.08	0.08	0.00	0.13	0.17
countries	0.12	0.03	0.12	0.07	0.13	0.00	0.06	0.07	0.03	0.00	0.14	0.15	0.01	1.00	0.08	0.11	0.09	0.12	0.16	0.11	0.15	0.00	-0.02	0.23
comply	0.08	0.07	0.10	0.05	0.20	0.00	0.17	0.10	0.09	0.00	0.24	0.13	-0.02	0.08	1.00	0.02	0.05	0.22	0.06	0.20	0.13	0.00	0.17	0.17
with	0.23	0.14	0.19	0.35	-0.01	0.00	0.07	0.11	0.06	0.00	0.22	0.27	0.20	0.11	0.02	1.00	0.31	0.12	0.28	0.11	0.05	0.00	0.07	0.13
legal	0.08	0.07	0.05	0.14	0.15	0.00	0.11	-0.02	0.19	0.00	0.12	0.19	-0.01	0.12	0.22	0.12	0.12	1.00	0.13	0.19	0.16	0.00	0.05	0.09
or	0.23	0.32	0.22	0.27	0.01	0.00	0.14	0.04	0.09	0.00	0.04	0.22	0.14	0.16	0.06	0.28	0.37	0.13	1.00	0.05	0.10	0.00	0.12	0.12
agreed	0.11	0.10	0.13	0.17	0.14	0.00	0.10	0.12	0.07	0.00	0.05	0.12	0.08	0.11	0.20	0.11	0.13	0.19	0.05	1.00	0.15	0.00	0.09	0.26
readmission	0.05	0.04	0.01	0.12	0.08	0.00	0.05	-0.01	0.13	0.00	0.05	0.09	0.08	0.15	0.13	0.05	0.01	0.16	0.10	0.15	1.00	0.00	0.23	0.18
and	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
return	0.07	0.20	0.12	0.19	0.01	0.00	0.29	0.04	0.25	0.00	0.12	0.16	0.13	-0.02	0.17	0.07	0.23	0.05	0.12	0.09	0.23	0.00	1.00	0.16
commitments	0.14	0.17	0.09	0.07	0.04	0.00	0.08	0.04	0.21	0.00	0.19	0.12	0.17	0.23	0.17	0.13	0.12	0.09	0.12	0.26	0.18	0.00	0.16	1.00

Figure 4.10 Part of heatmap of a TP instance

Example #4:

2nd stage misclassification

Translated from Greek	English
Mr. Stanzos asked in his speech to **protect the country's eastern borders** with **FRONTEX** operating **on the** opposite **Turkish coast**.	Eastern Samos Mayor Giorgos Stantzos said that European border patrol agency **Frontex** should **monitor the country's eastern** maritime **borders on the Turkish coastline**, and said they will not stand for turning the island into a large migration camp.

Match Ratio	Predicted Label	Truth Label
0,571	Non-Translation	Partial Translation

This is an example of a False Negative instance in the second classification, indicating that the system failed to identify a case of Partial Translation, as identified by a human annotator. Despite the linguistic similarities and the ratio being remarkably close to the threshold, this pair is not considered similar enough to be classified as a partial translation.

This is an interesting case because the English sentence is a highly edited version of the Greek sentence. This would be an interesting pair to include in a comparallel corpus as it showcases some common editing processes that take place during the news translation process. For example, there is an element of explicitation, which can be defined as "a shift in translation from what is implicit in the source text to what is explicit in the target text" (Murtisari, 2016). The target text here provides the information of Mr. Stanzo's role and full name (Giorgos Stanzos). In addition, the the full name of the FRONTEX organization is provided in the target text and the exact type of "border", i.e., "maritime", is provided in the target text.

Example #5:

2nd stage misclassification

Translated from Greek	English
The **origin** of the **residents** of **Moria** is 73% **Afghan**, 12% **Syrian** and 5% **Somali**.	The total number of minors in Moria had reached 6, According to figures released by the hotspot's administration, 36 percent of the total population in Moria are men, 29 percent women and 35 percent under **The** **nationalities** of the **asylum seekers** hosted in **Moria** are as follows: 73 pct **Afghans**, 12 pct **Syrians** and 5 pct **Somalis**.

Match Ratio	Predicted Label	Truth Label
0,600	Non-Translation	Partial Translation

This False Positive case is interesting because the ratio is just below the threshold. The translated sentence is present in the English text, but it is not labelled as a partial translation. This is mainly due to the numbers not being recognised as similar tokens between the two sentences, because of the limitations of the word embedding model. In addition, the difference between using the abbreviation “pct” instead of % affects the classifiers performance.

Additionally, the English text uses two paraphrases that, for the embedding model, are difficult to detect as similar. Although the human annotator can discern the contextual similarity between the terms *residents* and *asylum seekers*, the system is unable to do so. It is worth noting that the original similarity score within the pair, as determined by the sentence embeddings comparison, was 0.614, which is remarkably close to the threshold for being considered unrelated. This underscores the challenge that this pair presents for the system.

Example #6:

1st stage misclassification

Translated from Greek	English
Of those who arrived on Lesbos, 8,089 **people** were moved to structures inside Greece in 2017, 14,135 in 2018 and 13,406 in 18,747 applicants **live today**, January 2, 2020, **in the camp** inside the Reception and Identification Center and in the surrounding estates in Moria, **Lesvos asylum** of which 1,150 minors and unaccompanied according to their declaration or proven.	There are 21,441 **people currently residing** at **asylum** seekers' accommodations in the Mytilini area at Lesvos, including the **Moria** camp, while another 6,007 **live** in Chios and 7,519 in Samos islands.

Match Ratio	Predicted Label	Truth Label
0,613	Partial Translation	Non-Translation

This instance was incorrectly labelled as a Partial Translation for several reasons. Firstly, the calculation of the match ratio does not include the numbers, .which are not considered because they are not represented in the generated embeddings due to the limitations of the word embedding model regarding numeracy (Sundararaman *et al.*, 2020). Additionally, the word similarity threshold of 0.5 may count unrelated words as matches in some cases. This is evident in some tokens highlighted in Figure 4.11, such as the word *Mytilini* which has a similarity score of 0.54 with the word *Lesvos*. Although both are islands, they do not refer to the same place and should not be considered a match.

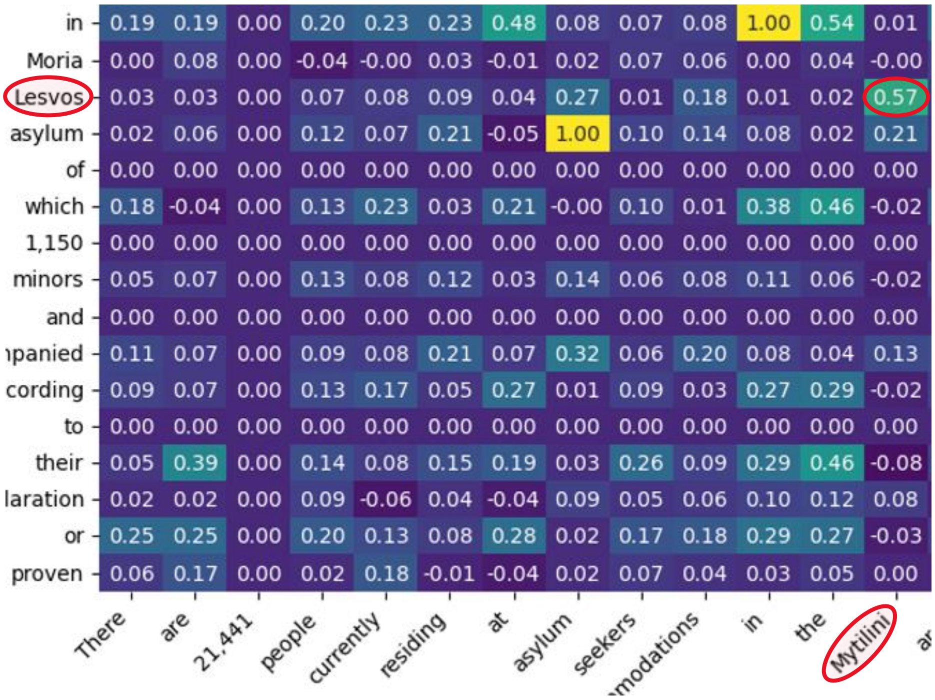


Figure 4.11 Part of heatmap of FP instance

Example #7:

1st stage misclassification

Translated from Greek	English
46 **migrants** who were on a boat off Samos were **rescued** - One woman dead Yesterday, **a Coast Guard offshore vessel** proceeded to pick up 47 **migrants**, including one dead woman, **who were on a boat in the sea area northwest of** Samos.	**Migrants rescued** in the sea area of Farmakonisi and Symi **A Coast Guard lifeboat rescued** 23 **migrants who were on a boat in the sea area southwest of** Farmakonisi.

Match Ratio	Predicted Label	Truth Label
0,759	Partial Translation	Non-Translation

This last example is a FP instance, meaning that the system incorrectly labelled it as a partial translation of two dissimilar sentences. Although the two sentences share many similar words, the English sentence describes a different event. The system considers the two sentences partial translations because of the type of words used. For instance, the words *southwest* and *northwest* have a cosine similarity of 0.91, while *Samos* and *Symi* have a score of 0.56. In both cases the words represent similar concepts (geographical directions and islands), thus, they are close in the vector space of word embeddings. This highlights the limitations of word embeddings regarding Named Entities (NEs).

These examples highlight the constraints of the system. Section 4.3 will provide a detailed analysis of these limitations, including their impact on the research outcomes. The discussion will also explore potential enhancements and adjustments to improve the system's performance. Additionally, it will address the broader impact of these limitations on the validity and applicability of the research findings.

4.3 Discussion

The results reported in Section 4.1 and the analysis conducted in Section 4.2 provide the information to be able to answer to the research questions defined in Section 1.1. To ensure a coherent discussion, it is first worth revisiting the research questions as they were originally posed:

> ***1st Research Question:*** *How effective are cosine similarity measures applied to sentence and word embeddings at automatically extracting similar sentences from pairs of news articles written in Greek and English assuming that one article is derived from the other or that they have a common source?*

Regarding the 1st RQ, the system uses cosine similarity effectively in combination with sentence and word embeddings to extract similar sentences. The cosine similarities of the sentence embeddings generated with LaBSE are an accurate representation of the actual similarity of the sentence pairs as suggested by a macro F1-score of 0.7817. The decision to compare sentence embeddings to assess the similarity of two sentences in combination with the design choice to consider two possible matches for each sentence enables the identification of similar sentences through a 1:2 alignment and thus, facilitates the identification of the most similar sentences in two comparable news articles in Greek and English.

> ***2nd Research Question:** After extracting sentences as outlined in RQ1, how accurately can these sentences be classified into categories (parallel, ambiguous, unrelated) that reflect the degree of similarity in terms of the translation relationship?*

In addressing the 2nd RQ, the system faces challenges in accurately classifying sentence pairs as ambiguous, as evidenced by low recall and precision scores of 0.60 and 0.52, respectively. Although it effectively classifies clear cases of parallel and unrelated sentence pairs (as demonstrated in the heatmap shown in Figure 4.4) utilizing the two thresholds of 0.55 and 0.75, the classification of ambiguous sentences underperforms, suggesting the need for further refinement of the criteria defining this category. This limitation is expected due to the system's dependence on a two-stage classification approach which aims at exploring in more detail the ambiguous pairs to extract partial translations.

> ***3rd Research Question:** Among sentences classified as ambiguous based on sentence embeddings, how reliably can the system disambiguate ambiguous pairs and categorize them into partial translations or parallel segments?*

Lastly, for the 3rd RQ, the 2nd classification stage performs equally well achieving an accuracy of 0.8085 and accurately distinguishing between partial translations and non-translations. Generally, the use of thresholds for text classifiers that use a rule-based classification approach poses an important limitation. While a fixed threshold may work for a specific dataset, it does not guarantee reliable performance when changes are applied to the dataset e.g., when more data is added or if the data is processed differently.

The proposed approach simplifies the process of setting thresholds for the parallel and unrelated categories, while also allowing for a focus on clarifying the ambiguous one. The ambiguous category utilizes word similarity and a ratio of matching words between the paired sentences to disambiguate pairs that were not clearly parallel or unrelated. By implementing these two features the system further classifies ambiguous pairs by categorizing them into two categories, i.e., partial translation and non-translation. This second categorization would not be possible, or it would be extremely arbitrary if the system was based only on sentence embeddings. This method aims to streamline the binary classification process by using simpler and more intuitive assumptions, thus reducing the arbitrary nature of the threshold used. Empirical analysis is still required to more adequately determine the exact values of the thresholds.

This study uses static thresholds which are inherently problematic in classification tasks. To address this issue, one potential area of research is the use of dynamic thresholding techniques. These techniques allow the system to learn the optimal threshold for different contexts or article types, potentially using machine learning techniques. An approach to mitigate the threshold issue was proposed by Artetxe and Schwenk (2019). The margin-based scoring approach they propose is defined as “the margin between the cosine of a given candidate and the average cosine of its k nearest neighbors in both directions”. This method addresses the inconsistencies of static thresholds by introducing a margin or a “buffer zone” around the cosine similarity scores. It adjusts the similarity measure based on a margin that accounts for the scale differences between vectors from different languages or domains.

The current implementation of the system can be used to extract parallel data for the creation of a comparallel Greek-English news corpus. However, the system still presents several limitations that have to be addressed to potentially improve its performance, allowing for more robust and scalable versions.

Improvements could be made by using a more suitable embedding model for word classification or applying better preprocessing to match the limitations of the current word2vec model used. Additionally, extra features such as positional information could be implemented to better assess word similarity and improve the matching ratio approach.

One of the main limitations of the word2vec model utilized in this study is the fact that it is a monolingual model. This limitation makes the MT engine a necessary component for the system adding variance and possibly alternating the original Greek sentence. Moreover, the numeracy problem of the model should be addressed in order to improve the performance of the system on sentences with a lot of numerical information which is quite common in news text. Some researchers have explored the ability of word embeddings models to capture numerical information. For example, Naik *et al.* (2019) argue that despite the widespread use of word embeddings in NLP, there is a significant gap in how these models handle numerical data. They highlight that numbers play a crucial role in language understanding, claiming that numerical information is essential for tasks where the meaning of a text can hinge on the precise interpretation of numerical values. Their study underscores the need for more sophisticated models that can accurately represent numerical information.

Finally, an area for improvement closely related to that of comparing numerical information between two sentences, is that of comparing Named Entities. The examples investigated in the Section 4.2, highlight that word embeddings alone cannot capture the similarities (or differences) between NEs. An approach to address this issue in the future would be to use a Named Entity Recognition (NER) model to extract NEs from the sentences and compare them to assess their similarity using NER alignment techniques.

5 Conclusions

This thesis project describes an automated approach to extracting parallel sentences from bilingual comparable news articles. The experiments detailed in the study focused on articles published by the Greek National News Agency (AMNA) and dealing specifically with the topic of migration. The approach, while demonstrated on a specific topic, is adaptable to other news subjects and language pairs, making it a valuable tool for researchers in both Translation studies and Journalism studies who are interested in the way multilingual news are produced.

In the Introduction (see Chapter 1) the research questions and hypotheses were defined focusing on the identification of parallel and partially translated sentences. The domain of news translation was argued to be a rich source for such data; however, due to the complex natures of journalistic text and specifically the unclear relationship between two comparable news articles, limited studies have focused on this task.

The thesis proceeds to lay out the theoretical framework that supports the ensuing methodology. Chapter 2 begins with exploring corpus typology, highlighting the possibilities offered by a novel 'comparallel' corpus approach, combining comparable and parallel perspectives. The focus then shifts to text similarity, covering essential elements such as text embeddings, cosine similarity, and evaluation metrics which are the main components of the classification process implemented here. The chapter concludes with presenting some of the main research contributions to the Parallel Sentence Extraction task and the News Translation field, which represent the foundation of the ensuing classification system.

In discussing system design and implementation, Chapter 3 provides a comprehensive overview of the data collection process, ensuring future reproducibility and highlighting the challenges and considerations of gathering data from news agency websites and other news outlets. The methodology section delves into the step-by-step process, going from data loading to the evaluation of the system, concluding with a detailed explanation of the evaluation process for assessing the system's performance.

Finaly, Chapter 4 presents the results of the experiments performed on the data collected from AMNA using the system proposed in this study. An error analysis is conducted on a sample of manually annotated instances showcasing the system's ability to extract parallel sentences and partial translations, while still illustrating some of the systems limitations that could be addressed in future research. Finally, this chapter addresses the research questions posed in Chapter 1, detailing how the findings align with the initial objectives and contribute to the field of Translation Studies.

The development of this system presents several implications beyond the technical ones explored in this study, including cross-linguistic analysis of news and the study of news translation techniques. These areas are linked by an understanding of how information crosses linguistic and cultural barriers to influence public perception. The system's output allows researchers to compare how news outlets report differently on the same events based on the culture and language of the target audience making it a great asset for Translation and Journalistic studies. For instance, researchers can examine how the portrayal of asylum seekers or migrants varies in English and Greek, revealing the agency's standards and potential bias, or they can study the phenomenon of explicitation (the process of making implicit information explicit) present in the English texts. This phenomenon provides evidence regarding the directionality of the translation and can be studied to gain insights into the agency's translation practices. From a Translation Studies perspective, the extracted parallel sentences can be used to translate texts containing cultural references and context-specific information.

In conclusion, this research enables the detailed study of how news are created and transferred across languages through translation. It highlights the importance of understanding multilingual text production, which includes the use of different translation strategies as a channel to consistently reach a broader audience and adapt to different realities.

References

Artetxe, M. and Schwenk, H. (2019) ‘Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings’, in A. Korhonen, D. Traum, and L. Màrquez (eds) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. *ACL 2019*, Florence, Italy: Association for Computational Linguistics, pp. 3197–3203. Available at: https://doi.org/10.18653/v1/P19-1309.

Barrón-Cedeño, A. *et al.* (2015) ‘A Factory of Comparable Corpora from Wikipedia’, in *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*. *BUCC 2015*, Beijing, China: Association for Computational Linguistics, pp. 3–13. Available at: https://doi.org/10.18653/v1/W15-3402.

Bernardini, S., Ferraresi, A., Garcea, F. and Rodriguez-Blanco, N. forthcoming. “Corpus approaches to news translation: We can do better than comparable!”. In Kajzer-Wietrzny, M. (ed). Capturing the complexity of language mediation with updated theories and enriched corpus designs. Special issue of Across Languages and Cultures.

Bielsa, E. (2007) ‘Translation in Global News Agencies’, *Target*, 19. Available at: https://doi.org/10.1075/target.19.1.08bie.

Chen, J., Chau, R. and Yeh, C.-H. (2004) ‘Discovering parallel text from the World Wide Web’, in *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32*. AUS: Australian Computer Society, Inc. (ACSW Frontiers ’04), pp. 157–161.

Clough, P., Gaizauskas, R. and Piao, S.L. (2002) ‘Building and annotating a corpus for the study of journalistic text reuse’, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02)*. *LREC 2002*, Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2002/pdf/218.pdf (Accessed: 30 July 2023).

Davier, L. (2014) ‘The paradoxical invisibility of translation in the highly multilingual context of news agencies’, *Global Media and*

Communication, 10(1), pp. 53–72. Available at: https://doi.org/10.1177/1742766513513196.

Davier, L. (2019) ‘The moving boundaries of news translation’, *Slovo*, 10, pp. 69–86. Available at: https://doi.org/10.5922/2225-5346-2019-1-5.

Davier, L. (2022) ‘Translating News’, in K. Malmkjær (ed.) *The Cambridge Handbook of Translation*. Cambridge: Cambridge University Press (Cambridge Handbooks in Language and Linguistics), pp. 401–420. Available at: https://doi.org/10.1017/9781108616119.021.

Davier, L. and Doorslaer, L.V. (2018) ‘Translation without a source text: methodological issues in news translation’, *Across Languages and Cultures*, 19(2), pp. 241–257. Available at: https://doi.org/10.1556/084.2018.19.2.6.

Devlin, J. *et al.* (2019) ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, in J. Burstein, C. Doran, and T. Solorio (eds) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. *NAACL-HLT 2019*, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. Available at: https://doi.org/10.18653/v1/N19-1423.

Federmann, C., Kocmi, T. and Xin, Y. (2022) ‘NTREX-128 – News Test References for MT Evaluation of 128 Languages’, in K. Ahuja et al. (eds) *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*. *SUMEval 2022*, Online: Association for Computational Linguistics, pp. 21–24. Available at: https://aclanthology.org/2022.sumeval-1.4 (Accessed: 1 November 2023).

Feng, F. *et al.* (2022) ‘Language-agnostic BERT Sentence Embedding’. arXiv. Available at: https://doi.org/10.48550/arXiv.2007.01852.

Gaspari, F. (2015) ‘Exploring Expo Milano 2015: a cross-linguistic comparison of food-related phraseology in translation using a comparallel corpus approach’, *The Translator*, 21(3), pp. 327–349. Available at: https://doi.org/10.1080/13556509.2015.1103099.

Granger, S. (2010) ‘Comparable and translation corpora in cross-linguistic research Design, analysis and applications’.

Guo, M. *et al.* (2018) ‘Effective Parallel Corpus Mining using Bilingual Sentence Embeddings’, in O. Bojar et al. (eds) *Proceedings of the Third Conference on Machine Translation: Research Papers. WMT 2018*, Brussels, Belgium: Association for Computational Linguistics, pp. 165–176. Available at: https://doi.org/10.18653/v1/W18-6317.

Levy, O. and Goldberg, Y. (2014) ‘Neural Word Embedding as Implicit Matrix Factorization’, in *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available at: https://papers.nips.cc/paper/2014/hash/feab05aa91085b7a8012516bc3533958-Abstract.html (Accessed: 28 February 2024).

Liu, S., Wang, L. and Liu, C.-H. (2018) ‘Chinese-Portuguese Machine Translation: A Study on Building Parallel Corpora from Comparable Texts’. arXiv. Available at: https://doi.org/10.48550/arXiv.1804.01768.

Mandelbaum, A. and Shalev, A. (2016) ‘Word Embeddings and Their Use In Sentence Classification Tasks’. arXiv. Available at: http://arxiv.org/abs/1610.08229 (Accessed: 23 February 2024).

McEnery, A. and Xiao, Z. (no date) ‘Parallel and comparable corpora: what are they up to?’

Migration in Greece: Recent Developments in 2019 (2020) *Sirius Project*. Available at: https://www.sirius-project.eu/news/migration-greece-recent-developments-2019 (Accessed: 17 February 2024).

Mikolov, T. *et al.* (2013) ‘Efficient Estimation of Word Representations in Vector Space’. arXiv. Available at: https://doi.org/10.48550/arXiv.1301.3781.

Murtisari, E. (2016) ‘Explicitation in Translation Studies: The journey of an elusive concept’, *Translation and Interpreting : the International Journal of Translation and Interpreting Research*, 8, pp. 64–81. Available at: https://doi.org/10.12807/ti.108202.2016.a05.

Naik, A. *et al.* (2019) ‘Exploring Numeracy in Word Embeddings’, in A. Korhonen, D. Traum, and L. Màrquez (eds) *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics. ACL 2019, Florence, Italy: Association for Computational Linguistics, pp. 3374–3380. Available at: https://doi.org/10.18653/v1/P19-1329.

Pennington, J., Socher, R. and Manning, C. (2014) 'GloVe: Global Vectors for Word Representation', in A. Moschitti, B. Pang, and W. Daelemans (eds) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). EMNLP 2014*, Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. Available at: https://doi.org/10.3115/v1/D14-1162.

Pym, A. (2008) 'On Toury's laws of how translators translate'. Available at: https://doi.org/10.1075/btl.75.24pym.

Pym, A. (2009) 'Exploring Translation Theories', *Anthony Pym* [Preprint]. Available at: https://doi.org/10.4324/9780203869291.

Reimers, N. and Gurevych, I. (2020) 'Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation'. arXiv. Available at: https://doi.org/10.48550/arXiv.2004.09813.

Schwenk, H. (2018) 'Filtering and Mining Parallel Data in a Joint Multilingual Space', in I. Gurevych and Y. Miyao (eds) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). ACL 2018*, Melbourne, Australia: Association for Computational Linguistics, pp. 228–234. Available at: https://doi.org/10.18653/v1/P18-2037.

Sharjeel, M. *et al.* (2023) 'Cross-lingual Text Reuse Detection at Document Level for English-Urdu Language Pair', *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6), p. 173:1-173:22. Available at: https://doi.org/10.1145/3592761.

Sundararaman, D. *et al.* (2020) 'Methods for Numeracy-Preserving Word Embeddings', in B. Webber et al. (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). EMNLP 2020*, Online: Association for Computational Linguistics, pp. 4742–4753. Available at: https://doi.org/10.18653/v1/2020.emnlp-main.384.

Wołk, K., Rejmund, E. and Marasek, K. (2015) ‘Harvesting comparable corpora and mining them for equivalent bilingual sentences using statistical classification and analogy- based heuristics’, in, pp. 433–441. Available at: https://doi.org/10.1007/978-3-319-25252-0_46.

Wu, Y. *et al.* (2016) ‘Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation’. arXiv. Available at: http://arxiv.org/abs/1609.08144 (Accessed: 19 January 2024).

Zanettin, F. (ed.) (2021) ‘Approaches to News Media Translation’, in *News Media Translation*. Cambridge: Cambridge University Press, pp. 86–132. Available at: https://doi.org/10.1017/9781108568364.004.

APPENDIX A

Search Protocol for AMNA Articles

Search Objective

This protocol describes the methods for finding articles belonging to the migration domain. Specifically, following the present protocol, we aim to find pairs of articles in Greek and English that write about the same topic.

The search is conducted on the website of the Greek news agency, namely, Athens-Macedonian News Agency (ANA-MPA).

The search will span from 2020 to October 2023 and articles on the migration situation in Greece will be the main topic. Articles on different aspects of the problem, e.g., EU discussions and decisions on the matter, will also be considered.

Keywords and Search terms

The search terms used will be reported in a list. The Greek and English search terms should be equivalent for better results. The first search terms will be more general, and then more specific search terms might be used for particular topics.

Articles Database

The retrieved article pairs are included in a database and are accompanied by multiple metadata, e.g., publication date, author(s), source, topic, URL, language, etc.

Detecting Similar Articles

In addition to the linguistic similarity of texts, some meta-textual information should be used to identify articles with a shared topic. Per the structure of the database, the meta-textual information considered is mainly publication date and author(s). However, additional patterns can be used, e.g., order of publication of the Greek and English articles, use of numbers and other anchors (e.g. proper nouns) in the title or main text, and images accompanying the articles. A list of such patterns will be compiled in addition to the search terms list.

Search Terms

1. **en:** migration, migrant, immigration, immigrant, illegal, asylum, rescued
2. **gr:** μεταναστευτικό, μετανάστες, παράνομοι, άσυλο, διασωθέντες

APPENDIX B

Annotation Guidelines for 1st Classification

Project Overview

The classification system to be evaluated aims at extracting parallel sentences from comparable corpora in the domain of news. This manual annotation will be used as the ground truth for the purpose of the evaluation of the classifier's performance.

Dataset Description

The data to be annotated include sentence pairs in Greek and English extracted from comparable news articles.

Annotation Task

The annotator must classify sentence pairs based on whether there is a translation relationship or not. The annotator must choose one of three provided classification labels: “Parallel” – “Ambiguous” – “Unrelated”.

Categories and Labels

Parallel: The sentence pair consists of two sentences that are semantically and structurally similar, and hence likely to have resulted from translation processes.

Example:

Greek: “Επίσης περιπολικό σκάφος του Λιμενικού που βρίσκονταν σε προγραμματισμένη περιπολία, εντόπισε σε παραλία της Σύμης 24 μετανάστες (18 άνδρες, 3 γυναίκες και 3 αγόρια).”
English: “Also, a patrol boat of the Coast Guard, which was on a scheduled patrol, spotted 24 migrants (18 men, 3 women and 3 boys) on a beach in Symi.”

Ambiguous: The sentence pair exhibits a certain degree of semantic similarity, yet this does not necessarily imply a direct translational relationship between the two sentences. The sentences are semantically close but not explicitly connected or related in context.

Example:

Greek: "Ιδιαίτερη αναφορά έγινε και στο ρόλο της Τουρκίας στη διαχείριση των μεταναστευτικών ροών στην Αν. Μεσόγειο, αλλά και ευρύτερα στην Ευρώπη."

English: ""Irregular migration is a critical problem not only for Greece, but also for Turkiye, as well as for the whole of Europe and the Balkans."

Unrelated: The two sentences are semantically and contextually dissimilar.

Example:

Greek: "Δύο παιδιά μεταφέρθηκαν στο νοσοκομείο της Λέσβου το ένα χωρίς τις αισθήσεις του."

English: "The other child was admitted to the hospital and is out of danger."

Consistency

When faced with uncertainty, especially in deciding whether a text is "ambiguous" or fits into "parallel" or "unrelated" categories, the annotator is encouraged to opt for "ambiguous". In such cases, it is better to mark a case as ambiguous for further review than to risk incorrect categorization.

APPENDIX C

Annotation Guidelines for 2nd Classification

Project Overview

The classification system to be evaluated aims at extracting partial translations from “ambiguous” sentence pairs. This manual annotation will be used as the ground truth for the purpose of the evaluation of the classifier's performance.

Dataset Description

The dataset consists of sentence pairs classified as “ambiguous” during the 1st classification stage. For this 2nd classification stage, the Greek sentences are translated into English to allow for a monolingual comparison. The task will help disambiguate “ambiguous” pairs enhancing the understanding of the spectrum of translational relationships beyond “parallel” and “unrelated”.

Annotation Task

Annotators are tasked with classifying sentence pairs according to the presence of a translational relationship. For this phase, the focus will be on identifying "Partial Translation" and "Non-Translation." Annotators must select one of the two provided labels for each sentence pair.

Categories and Labels

Partial Translation: The sentence pair shows evidence of translational equivalence for some parts of the sentences, indicating that a portion of the content has been translated, albeit not in its entirety. This category captures the essence of sentences that share partial semantic similarities due to translation but also contain elements that do not align directly.

Example:

Translated: “The borders at Evros are now much better guarded.”

English: “Mitsotakis stressed that Greece's policy on managing migration and refugees has changed: "The borders in Evros are now much better guarded and the sea borders are also better guarded."”

Non-Translation: The sentences in the pair are unrelated, showing no evidence of translational equivalence or semantic similarity. This category is for sentence pairs that do not share any content or meaning that would suggest a translation has occurred.

Example:

Translated: “Specifically, 43 refugees-immigrants were found on the coast of Lesvos in two different incidents, 33 in Chios and 46 in Samos.”

English: “Specifically, 34 migrants and refugees were located at Alexandroupolis, 7 at Chios, 109 at Samos and 35 at Kos.”

Consistency

In cases of uncertainty when differentiating between “Partial Translation” and “Non-Translation”, the annotator is encouraged to opt for "Partial Translation" if any semantic similarities or partial content overlaps are detected. This conservative approach ensures that potential translational elements are not overlooked, allowing for a more nuanced analysis of the data.