Translation Alignment Applied to Historical Languages

Methods, Evaluation, Applications, and Visualization

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig

angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

Vorgelegt

von M. Sc. Tariq Yousef

geboren am 01.09.1987 in Homs (Syrien)

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Gerhard Heyer (Universität Leipzig)
2. Prof. Dr. Stefan Jänicke (University of Southern Denmark)
3. Prof. Dr. Alexander Mehler (Goethe-Universität Frankfurt)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 13.06.2023 mit dem Gesamtprädikat *magna cum laude*.

Abstract

Translation alignment is an essential task in Digital Humanities and Natural Language Processing, and it aims to link words/phrases in the source text with their translation equivalents in the translation. In addition to its importance in teaching and learning historical languages, translation alignment builds bridges between ancient and modern languages through which various linguistics annotations can be transferred. This thesis focuses on word-level translation alignment applied to historical languages in general and Ancient Greek and Latin in particular. As the title indicates, the thesis addresses four interdisciplinary aspects of translation alignment.

The starting point was developing Ugarit, an interactive annotation tool to perform manual alignment aiming to gather training data to train an automatic alignment model. This effort resulted in more than 190k accurate translation pairs that I used for supervised training later. Ugarit has been used by many researchers and scholars also in the classroom at several institutions for teaching and learning ancient languages, which resulted in a large, diverse crowd-sourced aligned parallel corpus allowing us to conduct experiments and qualitative analysis to detect recurring patterns in annotators' alignment practice and the generated translation pairs.

Further, I employed the recent advances in NLP and language modeling to develop an automatic alignment model for historical low-resourced languages, experimenting with various training objectives and proposing a training strategy for historical languages that combines supervised and unsupervised training with mono- and multilingual texts. Then, I integrated this alignment model into other development workflows to project cross-lingual annotations and induce bilingual dictionaries from parallel corpora.

Evaluation is essential to assess the quality of any model. To ensure employing the best practice, I reviewed the current evaluation procedure, defined its limitations, and proposed two new evaluation metrics. Moreover, I introduced a visual analytics framework to explore and inspect alignment gold

standard datasets and support quantitative and qualitative evaluation of translation alignment models. Besides, I designed and implemented visual analytics tools and reading environments for parallel texts and proposed various visualization approaches to support different alignment-related tasks employing the latest advances in information visualization and best practice.

Overall, this thesis presents a comprehensive study that includes manual and automatic alignment techniques, evaluation methods and visual analytics tools that aim to advance the field of translation alignment for historical languages.

Acknowledgements

I would like to express my sincere gratitude and appreciation to my PhD advisors, Prof. Gerhard Heyer and Prof. Stefan Jänicke. Your guidance, expertise, and support throughout my doctoral journey have been invaluable. Your dedication to my academic and personal growth has shaped me into the researcher I am today. I am also immensely grateful to Prof. Gregory Crane for guiding and supporting me during my first steps in academia.

I would also like to thank my colleagues Chiara Palladino and Farnoosh Shamsian. Our collaboration and teamwork have been instrumental in achieving our research goals. A special and heartfelt thank you goes to my parents, Nadima and Yousef. Your unwavering belief in my abilities and constant support have been the cornerstone of my success. Your sacrifices and encouragement have motivated me to overcome challenges and strive for excellence.

To my beloved wife Rana and our son William, thank you for your endless love, patience and understanding. Your support and belief in me has been my constant motivation. Your presence and encouragement made this journey worthwhile even in the most difficult times.

With deepest gratitude,

Tariq Yousef

Contents

List of Figures **vii**

List of Tables **xi**

1 Introduction **1**

1.1 Motivations and Challenges 2
1.2 Scope of the Thesis 4
1.3 Scientific Contribution 5
1.4 Thesis Structure 8

2 Manual Alignment with Ugarit **11**

2.1 Related Work 12
2.2 Development Process 14
2.3 Alignment Workflow 16
2.4 Visualization Techniques 19
2.4.1 Languages Graph 19
2.4.2 Aligned Texts 20
2.4.3 Translation Pairs & The Dynamic Lexicon 22
2.5 Ugarit in Practice 23
2.5.1 Research and Pedagogy 23
2.5.2 Translations Graph 26
2.5.3 Creating Training Data 28
2.5.4 Creating Alignment Gold Standard 28
2.5.5 Qualitative Analyses and Experiments 28
2.6 Future Work 42

3 Automatic Translation Alignment **43**
3.1 Overview of Automatic Alignment Models 43
3.1.1 Statistical Models 44
3.1.2 Neural Models 45
3.1.3 Hybrid Models 46
3.2 Automatic Translation Alignment Applied to Historical Languages . . . 47
3.2.1 Word Embedding 47
3.2.2 Alignment Workflow 48
3.2.3 Training Process and Objectives 53
3.2.4 Evaluation 56
3.2.5 Results 58
3.2.6 User Interface 61
3.2.7 Summary 61

4 Translation Alignment Evaluation **63**
4.1 Alignment Gold Standards 64
4.1.1 Alignment Gold Standard For Ancient Greek 66
4.2 Evaluation Metrics 69
4.2.1 AER Limitations 70
4.2.2 Alignment Coverage and Phrase Alignment Accuracy 72
4.3 Visual Evaluation of Translation Alignment Models 75
4.3.1 Implementation Details 75
4.3.2 User Interface 76
4.3.3 Visual Design 79
4.3.4 Usage Scenarios 83
4.4 Summary and Future Work 86

5 Applications of Translation Alignment **91**
5.1 Translation Alignment in Practice 92
5.2 Named-Entity Annotation Projection Applied to Ancient Greek 92
5.2.1 NER & Ancient Texts 93
5.2.2 Transformer-Based Models 93
5.2.3 Cross-lingual Annotation Projection 96
5.3 Bilingual Lexica Induction from Parallel Corpora 101
5.4 Evaluating Cross-Lingual Semantic Transfer 104
5.5 Future Work 108

6 Text Alignment Visualization **111**
6.1 Survey of Text Alignment Visualization 112
6.1.1 Text Alignment Visualizations 113
6.1.2 Design Implications 128
6.1.3 Future Challenges 129
6.2 Visualizing Different Alignment Levels 132
6.2.1 Ugarit Translation Alignment Editor 132
6.2.2 EvAlign 133
6.2.3 Parallel Corpus Visualizer 133
6.2.4 Summary Explorer 134
6.2.5 Miscellaneous Alignment-Related Visualizations 137

7 Conclusion **145**

References **149**

List of Figures

1.1 The thesis scope and scientific contributions. 5

2.1 User-Centered design process. 14
2.2 Overview of Ugarit's features and functionalities. 15
2.3 Ugarit bilingual alignment editor 17
2.4 Ugarit trilingual alignment editor 18
2.5 Languages graph. 19
2.6 Side-by-side visualization of bilingual aligned texts. 21
2.7 Visualization of translation pairs search results. 23
2.8 Translations cloud of the root word *city*. 27
2.9 Experiment 2, Case Study 1: instructors vs students. 36
2.10 Experiment 2, Case Study 2, Alignments of Hetrodotus's Histories (1.1) 38
2.11 Experiment 2, Case Study 2, Alignments of Odyssey (1.1-50) 39
2.12 Comparison criteria of Experiment 3. 41

3.1 Overview of automatic alignment models. 45
3.2 The alignment workflow. 48
3.3 An example of the alignment workflow using mBERT 49
3.4 An example of the alignment workflow using XLM-RoBERTa 50
3.5 A comparison among the alignment extraction approaches 59
3.6 Ugarit automatic alignment tool. 62

4.1 Example of alignment gold standard format 65
4.2 The correlation between AER and S/P Ratio 70
4.3 AER Limitations 71
4.4 A comparison among gold standard datasets 73
4.5 A comparison among 20 models regarding the alignment coverage 74
4.6 Comparison among alignment models regarding AER and PAC. 74

LIST OF FIGURES

4.7 Overview of EvAlign . . . 76
4.8 Sentence-level views . . . 77
4.9 Sentence-level table view . . . 78
4.10 The aggregated table view . . . 79
4.11 Single dataset-single model corpus-level view . . . 80
4.12 Sentence-level grid view . . . 81
4.13 Corpus-level pie chart view . . . 82
4.14 Examples from the EN-FR dataset . . . 84
4.15 The dataset card . . . 85
4.16 Sentence *127* from the DE-EN dataset . . . 86
4.17 Sentence *40* from the DE-EN dataset . . . 87
4.18 Sentence *101* from the RO-EN dataset . . . 88
4.19 Frequent alignment errors . . . 89

5.1 Named-Entity annotation projection pipeline. . . . 96
5.2 An example of the annotation projection using the proposed pipeline. . . . 97
5.3 A performance comparison between three STOA NER models . . . 98
5.4 Annotation projection example. . . . 99
5.5 The graph structure of the induced translation pairs. . . . 103
5.6 An example from graph clustering results. . . . 103
5.7 An example from graph clustering results with extended relations. . . . 104
5.8 The proposed pipeline for exploring cross-lingual semantic frameshifts. . . . 105
5.9 Examples of word aligned and semantic frame tagged sentences. . . . 106
5.10 Frameshifts of the semantic primes *See* and *Feel*. . . . 108

6.1 Visual representation of sequence-aligned heat maps. . . . 113
6.2 TexTile variation visualization at different text hierarchy levels (19). . . . 115
6.3 Visual representation of grid-based heat maps. . . . 116
6.4 Text re-use Grid and text re-use browser . . . 117
6.5 Visual representation of Aligned Barcodes. . . . 117
6.6 GuttenPlag Barcode Visualization. . . . 118
6.7 Visual representation of Text-oriented Heat Maps. . . . 118
6.8 EVT text-oriented heat map visualization. . . . 119
6.9 GuttenPlag text-oriented heat map visualization . . . 120
6.10 Visual representation of side-by-side views. . . . 120
6.11 LERA side-by-side view . . . 121
6.12 Iteal combines aligned barcodes with a side-by-side view (116). . . . 122
6.13 Ugarit side-by-side view . . . 122

6.14 Visual representation of Tabular Views. 123
6.15 CollateX tabular view visualization 124
6.16 Bible interlinear visualization. 124
6.17 Alpheios interlinear visualization. 124
6.18 Visual Representation of Text Variant Graphs 125
6.19 TRAViz variant graph. 126
6.20 Storyline-based text variant graph (7). 126
6.21 Parallel Corpus Visualizer aggregated search results 133
6.22 Summary Explorer heat map overview. 134
6.23 Summary Explorer entity coverage view. 135
6.24 Summary Explorer content coverage ciew. 136
6.25 Summary Explorer hallucinations aggregated view. 136
6.26 Summary Explorer Relation Coverage View. 137
6.27 Summary Explorer position bias view. 137
6.28 Summary Explorer positional bias aggregated view 138
6.29 Annotators agreement visualization view. 139
6.30 Divan Hafez sentence level view. 140
6.31 Divan Hafez poem-level view. 141
6.32 Information Flow. 141
6.33 Sentence-Level view showing the cross-lingual semantic transfer. 142
6.34 Semantic frames transfer aggregated view. 143

LIST OF FIGURES

List of Tables

1.1 A list of publications as first author. 9
1.2 A list of publications as co-author. 10

2.1 Top ten institutions from which UGARIT users come. 24
2.2 An overview of the training dataset. 29
2.3 Jensen-Shannon divergence and distance for three Gold Standard datasets 31
2.4 Alignment ratios of various subsets of translations from ancient languages into English. 32
2.5 Alignment ratios of various subsets of translations from Ancient Greek. 33
2.6 Experiment 2, Case Study 2, Users Group 3. A comparison of established translation pairs by the two users and their overlap and agreement. . . . 37

3.1 An overview of the conducted experiments. 54
3.2 Evaluation results on Ancient Greek-English gold standards. 57
3.3 Evaluation results on Ancient Greek-Portuguese gold standards. 57
3.4 Evaluation results on Ancient Greek-Latin gold standards. 58
3.5 Comparison between two different embeddings alternatives regarding AER - 60

4.1 An overview of the existing alignment gold standard datasets. 66
4.2 An overview of translation alignment gold standard datasets 67
4.3 An overview of the gold standards datasets and their IAA 69

5.1 An overview of the training dataset. 95
5.2 Training and evaluation results. 95
5.3 Manual evaluation of 100 randomly selected verses. 100
5.4 Overview of the processed parallel corpora. 102
5.5 Experiment Results 107

LIST OF TABLES

6.1 Surveyed Tools & Visualization Techniques 114
6.2 Applicability of Visualization Techniques to Corpus & Task-related Features 130
6.3 An overview of the implemented tools and the used visualization techniques. 132

7.1 A list of language models hosted on `https://huggingface.co`. 146
7.2 A list of online tools. 146

Chapter 1

Introduction

Translation is the procedure of understanding and interpreting a text in a *source* language and rendering it in another *target* language in a way that preserves its semantic and pragmatic meaning (108). Translation was the means of transferring knowledge and experiences between communities and civilizations. Texts accompanied by their translations are called *Parallel Texts*, and perhaps the greatest examples of this are the Ugarit inscriptions and the Rosetta Stone, which played a crucial role in deciphering Egyptian hieroglyphic scripts (195). Parallel corpora are valuable resources for cross-lingual information retrieval and machine translation systems. They have also gained significant importance in Digital Humanities and Corpus Linguistics as a way to understand relationships between languages and translation dynamics (24).

In basic terms, *Translation Alignment* is the process of finding translation equivalents between a source text and its translation. It can be performed manually and automatically at different text granularity levels. In this thesis, we deal with word-level alignment. Translation alignment, both manual and automatic, is a challenging task and its quality is subject to various factors, such as translation quality and style, text genre, and language pair.

The alignment of historical texts with their modern translations is essential to build bridges between the past and the present and facilitates studying and learning ancient languages and understanding history. It allows us to observe the complex interplay between cultures and linguistic structures of different languages. In addition to its cultural and pedagogical importance, translation alignment plays a key role in various natural language processing tasks such as statistical machine translation, cross-lingual annotation projection, and bilingual lexicon induction.

1.1 Motivations and Challenges

The title of the dissertation indicates that we are going to apply **translation alignment** to **historical languages** and address its different aspects. In the following, we explain our motivations and the challenges we tried to overcome:

Why Historical Languages? The development process in the fields of Natural Language Processing (NLP) and Digital Humanities (DH) follows a different pattern. While NLP models are evolving very rapidly, the transfer and application of these models in DH still faces many obstacles. The lack of adequate infrastructures and annotated data is the main obstacle to developing and adopting the recent advances in NLP to historical languages. It is no secret that working and developing models for modern languages are more attractive to young researchers due to the availability of the necessary tools, data, and means to start research directly. However, working with ancient languages and diving into their details is challenging but has its appeal. Our efforts in this thesis seek to bridge the gap between NLP and DH by employing the recent advances in NLP and Information Visualization to texts in historical languages. Since we can not cover all ancient languages in this thesis, we focused on the Ancient Greek language in the first place and Latin and classical Persian in the second place, noting that all the development pipelines and models presented can be applied to other ancient or modern languages.

Why Translation Alignment? Translation Alignment allows students, scholars, and literary researchers to explore texts in ancient languages without needing prior knowledge of those languages, especially when it is integrated with interactive reading environments of parallel texts. It facilitates and simplifies the learning process of historical languages. In addition to its importance in teaching and learning historical languages (190), the added value of accurate translation alignment lies in creating a bridge between ancient and modern languages through which much of the linguistics annotations can be transferred. Further, a vast amount of OCR-ed books and manuscripts containing ancient texts with parallel translations in modern or classical languages are available in unstructured or semi-structured formats. We believe that processing these books and aligning them at some granularity level (paragraph, sentence, word) would be beneficial for better understanding the rich history.

Challenges

Developing an automatic alignment model requires three main components, training data, selecting the appropriate alignment approach, and a systematic evaluation to assess the quality of the developed model.

Data: Word-level alignment models use parallel sentences for unsupervised training. Although parallel texts are available to some extent, we still face two main obstacles: 1) the accuracy of the available parallel corpora. i.e., the parallel sentences are not accurately aligned. 2) the amount of parallel text is insufficient to train accurate models, and if they are sufficient for some language pairs, such as Ancient Greek-English and Latin-English, there are no or minimal amount of parallel texts for other languages, such as Portuguese, Persian, Arabic. Supervised training, which employs word-level aligned parallel sentences, significantly improves the performance of automatic alignment models. However, word-level aligned corpora are not easy to find or create. The main challenges in this regard are: *Can the alignment models perform sufficiently despite the lack, or total absence, of training data? How can accurate training data be produced? How can crowd-sourcing be helpful?*

Alignment Workflow: Statistical alignment models such as Giza++ and fast_align require a vast amount of training data in the form of parallel sentences in order to produce good results. Further, data sparseness is a central problem, especially for inflected languages like Ancient Greek, which yield a poor performance for less frequent words. However, neural alignment models achieved better performance on modern languages. *Can these models perform sufficiently on historical languages even though the underlying pretrained language models are not trained on historical languages? What is the best training strategy for performance improvement? How can the model be used for cross-lingual applications?*

Evaluation: Systematic evaluation of NLP models requires employing the appropriate evaluation metrics for the task and reference gold standards to compare the model predictions with them. To the best of our knowledge, there is no alignment gold standard for ancient languages for historical languages. Further, *how to assess the quality of automatic alignment models? Is the current practice satisfactory? and How **GOLD** are the available gold standards?*

Deployment & Visualization The alignment model and other developed applications will be used mainly by students, scholars, and literary researchers. *How can we make this tool accessible and usable even for users with no technical knowledge? How can we support*

researchers in exploring and analyzing aligned corpora, comparing the performance of two alignment models, or inspecting the alignment gold standard datasets?

The following chapters are intended to help answer these questions.

1.2 Scope of the Thesis

In this thesis, we study the task of word-level translation alignment applied to historical languages in general and Ancient Greek and Latin in particular. As the title indicates, we address four interdisciplinary aspects of translation alignment.

Figure 1.1 illustrates the scope of the thesis and our contributions. The starting point was developing Ugarit, an interactive annotation tool to perform manual alignment aiming to collect training data to train an automatic alignment model. This resulted in more than 190k accurate translation pairs, which we later used for supervised training. Ugarit has been used in research and also in the classroom at several institutions for teaching and learning ancient languages, which resulted in a large and diverse crowd-sourced aligned corpus that allowed us to conduct experiments and qualitative analyses to detect recurring patterns in annotators alignment practice and the generated translation pairs.

Further, we employed the recent advances in NLP and language modeling to develop an automatic alignment model for historical low-resourced languages, experimenting with various fine-tuning objectives and proposing a training strategy for historical languages. Then, we integrated the alignment model into other development workflows to project cross-lingual annotations and induce bilingual dictionaries from parallel corpora.

Evaluation is essential to assess the quality of NLP models. To ensure employing the best practice, we reviewed the current evaluation procedure, defined its limitations, and proposed two new evaluation metrics. Moreover, we introduced a visual analytics framework to explore and inspect alignment gold standard datasets and support quantitative and qualitative evaluation of translation alignment models.

Besides, we designed and implemented visual analytics tools and reading environments for parallel texts and proposed various visualization approaches to support different alignment-related tasks employing the latest advances in information visualization and best practice.

Finally, we want to emphasize that all the models, processing pipelines, and visualization approaches used and developed in the context of this thesis are not limited to historical languages, but can also be applied to modern languages.

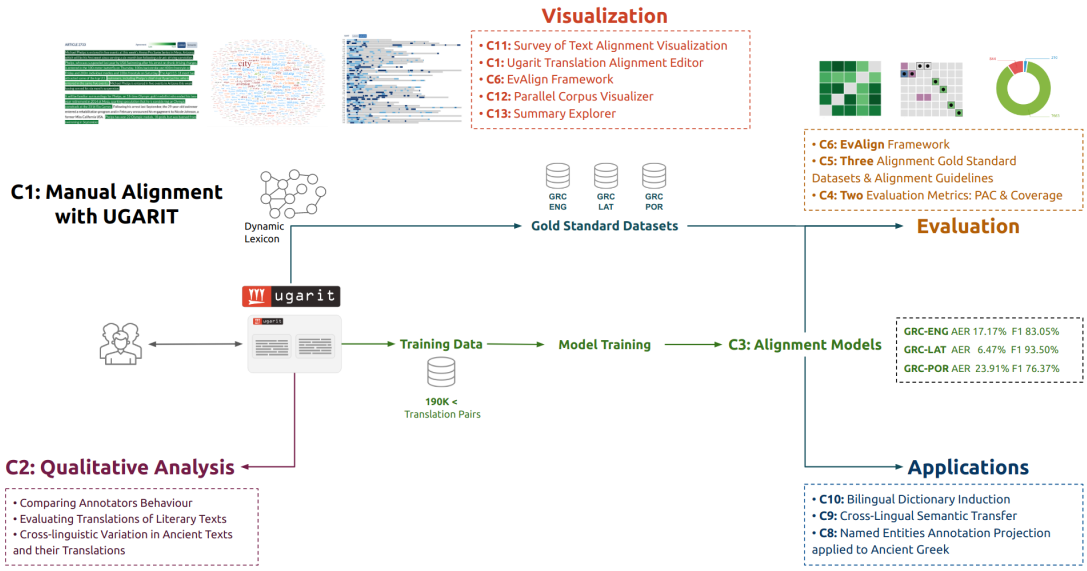


Figure 1.1: The thesis scope and scientific contributions.

1.3 Scientific Contribution

In the following, we briefly outline the main contributions of the thesis:

Manual Alignment and Qualitative Studies

C1) We developed UGARIT alignment editor, an annotation tool to align parallel texts at the word level manually. UGARIT has been widely used for teaching and learning ancient and historical languages (189, 190, 193) and resulted in a large and diverse corpus of crowd-sourced parallel texts in several languages, some of which have never been consistently aligned before. Further, UGARIT has been used to create alignment gold standards and generate training data for the supervised training of automatic alignment models (272).

C2) We made use of the manual alignments available on UGARIT to conduct three experimental studies in collaboration with colleagues from digital classics and philology to qualitatively analyze selected subsets of the corpus. We aimed to understand and compare annotators behaviors and tendencies, evaluate translations of literary texts (191), and study the cross-linguistic variation in ancient texts and their translations (187).

Automatic Translation Alignment

C3) We reviewed the available models and workflows and adopted a state-of-the-art workflow that employs multilingual embeddings and cross-lingual semantic similarity to generate translation pairs. Further, we trained multilingual contextualized language models, fine-tuned them on the translation alignment task, and proposed a training strategy for historical languages (e.g., Ancient Greek) not represented in the state-of-the-art multilingual language models such as mBERT and XLM-RoBERTa (271, 273). The strategy combines supervised and unsupervised training on monolingual and multilingual parallel texts. The models achieved state-of-the-art performance on Ancient Greek-English, Ancient Greek-Latin, Ancient Greek-Persian, and Ancient Greek-Portuguese, even without any Greek-Portuguese parallel sentences in the training dataset.

Evaluation of Automatic Alignment Models

C4) We reviewed the existing benchmark datasets, studied the alignment evaluation process, and defined the limitations and drawbacks of the current practice. Further, we proposed two quantitative evaluation metrics, namely, the *alignment coverage* and *phrase alignment accuracy*, to overcome the limitations.

C5) In order to evaluate the automatic alignment models we trained, we created alignment gold standard datasets for three language pairs, namely, Ancient Greek-English, Ancient Greek-Portuguese, and Ancient Greek-Latin considering the limitations and the failures we found during the work on **C4** (271, 273). Prior to that, we developed alignment guidelines for the language pairs mentioned above in collaboration with domain experts, aiming to reduce inconsistency among annotators and ensure accurate and homogeneous alignments.

C6) We developed EvAligan, a visual analytics framework to support quantitative and qualitative evaluation of translation alignment models. EvAlign allows users to inspect the hosted alignment gold standard datasets and explore quantitative evaluation metrics of the hosted models. It also provides the ability to conduct an extensive qualitative evaluation using different interactive visualization views and filtering options. The framework hosts nine gold standard datasets and the predictions of over 20 alignment models.

Translation Alignment Applications

C7) We trained transformers-based Named-Entities Recognition models for Ancient Greek that achieved a state-of-the-art performance and outperformed the existing models (270).

C8) We developed an alignment-based approach for cross-lingual annotation projection of named entities. The proposed method uses sentence-level aligned parallel corpora of an ancient language and a modern language, for which off-the-shelf NER systems are available. We automatically annotate the text of the modern language using one of the available tools, such as flairNLP. We then employ the alignment workflow and language model we trained previously to find translation pairs and project the annotation to the corresponding tokens in the ancient texts.

C9) We employed translation alignment to investigate the properties of the cross-lingual transfer of semantic frames based on semantic primes (105). Our study[1] focused on two language pairs, English-German and English-Spanish. Further, we employed *Semantic Mirroring* by translating the translations back to the original language to consider the variations of frames evoked by original, translated, and back-translated sentences as clusters of meaning distinctions.

C10) We employed the alignment model we trained to induce bilingual dictionaries from parallel texts. We used a parallel corpus of 400k parallel sentences in six languages, Ancient Greek, Arabic, English, Persian, Latin, and Hebrew, and resulted in five million translation pairs. We modeled the obtained translation pairs in a graph data structure, which allowed us to integrate manual alignments and dictionary entries and facilitate performing clustering and pivoting to generate translation pairs of languages with no direct connections.

Translation Alignment Visualization

C11 We surveyed the existing text alignment visualization approaches, discussed their advantages and drawbacks, and derived design implications. We reviewed 40 visual interfaces that support three text alignment scenarios: collation in textual criticism, text reuse, and translation alignment(267).

[1]This study has been conducted in collaboration with experts in Translation Studies, Citizen Science, and Cognitive Neuroscience. My contributions are listed in Section 5.4

C12 We designed the Parallel Corpus Visualizer to assist scholars and researchers in exploring an aligned corpus and looking closely at the aligned words and phrases and their aggregated statistics. The tool takes as input a manually or automatically aligned corpus in a specific format and parses the corpus to extract translation pairs with their frequency. Further, we used the tool to visualize three manually and automatically aligned corpora.

C13 We created the Summary Explorer (235), an interactive visual analytics framework that supports quantitative and qualitative evaluation of automatic summarization models [1]. The framework offers various visualization views to facilitate navigation and exploration of the hosted datasets and visualizes the alignment between the original documents and the summaries at different granularity levels. The framework hosts the outputs of 55 automatic summarization models covering three benchmark datasets. The datasets hosted on the Summary Explorer include only single-document English summaries. However, the proposed visualization approaches and assessment scenarios can also be employed to visualize the outputs of cross-lingual summarization models.

In the context of this thesis, we have published 15 papers in peer-reviewed conferences and journals in NLP , Digital Humanities , and Visualization. In addition to several papers that are not related to the thesis topics. The full list of publications is available on Google Scholars[2].

1.4 Thesis Structure

The thesis is organized into five chapters in addition to the introduction and conclusion chapters. Each chapter addresses one of the aspects mentioned in the title:

In **Chapter 2**, we introduce Ugarit, a manual alignment editor we developed to collect training data for training an automatic alignment model. We describe the development process and the alignment workflow. Then, we present the different visualization techniques used to visualize different data types. We demonstrate the impact of Ugarit in research and pedagogy afterward and present the qualitative studies we conducted.

[1] This is joint work, and it has been done in collaboration with Shahbaz Syed, who was responsible for preprocessing the data and calculating the evaluation metrics, and Tariq Yousef was responsible for designing and implementing the visualization views.

[2] `https://scholar.google.com/citations?user=KtcjnG8AAAAJ`

	Title	Year	Venue
1	A Survey of Text Alignment Visualization (267) ***T Yousef**, S Jänicke*	2020	IEEE TVCG
2	Summary Explorer: Visualizing the State of the Art in Text Summarization (235) *S Syed, **T Yousef**, K Alkatib, S Jänicke, M Potthast*	2021	EMNLP
3	Translation Alignment with Ugarit (272) ***T Yousef**, C Palladino, F Shamsian, M Foradi*	2022	MDPI
4	An Automatic Model and Gold Standard for Translation Alignment of Ancient Greek (271) ***T Yousef**, C Palladino, F Shamsian, A Ferreira, MF Reis*	2022	LREC
5	Automatic Translation Alignment for Ancient Greek and Latin ***T Yousef**,C Palladino, D J Wright, M Berti* (273)	2022	LREC WS LT4HAL
6	Visual Evaluation for Translation Alignment Data (268) ***T Yousef**, S Jänicke*	2022	EuroVis
7	Transformer-Based Named Entity Recognition for Ancient Greek ***T Yousef**, C Palladino, S Jänicke* (270)	2023	DH
8	EvAlign: Visual Evaluation of Translation Alignment Models (274) ***T Yousef**, G Heyer, S Jänicke*	2023	EACL
9	Named Entities Annotation Projection Applied to Ancient Greek ***T Yousef**, C Palladino, G Heyer, S Jänicke* (275)	2023	LaTeCH

Table 1.1: A list of publications as first author.

In **Chapter 3**, we review automatic alignment models and the current state-of-the-art. Then, we present our work on training a model for aligning historical languages benefiting from the accurate manual alignments available on Ugarit. And discuss the training strategy we followed which led to a state-of-the-art performance.

In **Chapter 4**, we address the evaluation of alignment models, discuss the current practice, define its drawbacks and limitations, and propose two new metrics to overcome them. Further, we present our work to generate alignment guidelines and gold standards. Finally, we present EvAlign, explain its structure and usage scenarios and list our observations on the hosted datasets and alignment models.

In **Chapter 5**, we report on the various applications of text alignments and present our work on three cross-lingual tasks: (1) named-entities annotation projection in a parallel corpus, (2) cross-lingual semantic transfer and frameshifts, (3) multilingual dictionaries induction from parallel corpora.

	Title	Year	Venue
1	How to Read All Languages: Translation Alignment with Ugarit (189) *C Palladino, A Zhang, M Foradi,* ***T Yousef***	2020	DH
2	Translation Alignment for Historical Language Learning: a Case Study *C Palladino, M Foradi,* ***T Yousef*** (190)	2021	DHQ
3	Embodied Metarepresentations (105) *N Hinrichs, M Foradi,* ***T Yousef****, E Hartmann, S Triesch, J Kaßler, J Pein*	2022	Frontiers
4	Using Parallel Corpora to Evaluate Translations of Ancient Greek Literary Texts (191) *C Palladino, F Shamsian,* ***T Yousef***	2022	JCLS
5	Ugarit: Translation Alignment Technologies for Under-resourced Languages (193) *C Palladino,* ***T Yousef****, F Shamsian, N Kanagawa*	2022	DH
6	A Study on Cross-linguistic Variation in Ancient Texts and their Translations (187) *C Palladino,* ***T Yousef***	2023	DSH

Table 1.2: A list of publications as co-author.

In **Chapter 6**, we review visualization techniques used for monolingual and cross-lingual text alignment, discuss their advantages and drawbacks, and deliver design implications. Then, we introduce our work on the Parallel Corpus Visualizer, Summary Explorer, and present our tailored visual solutions for different alignment-related tasks.

In the **Conclusion** 7, we provide a summary of the scientific contributions and the addressed topics and define future research directions.

Remark

Although this dissertation is the work of a single author, the pronoun "we" is used because it is the typical writing style most readers are familiar with. We have explicitly mentioned which dissertation parts have been done in collaboration with other researchers.

Chapter 2

Manual Alignment with Ugarit

Translation alignment is a main task in Digital Humanities and Natural Language Processing. It can be performed at various granularity levels according to the project's context or the research purpose. Several automatic approaches are available to perform the alignment at different levels (74, 114, 239). Nevertheless, manual alignment is still essential despite the advances in automatic translation alignment models, especially for generating training data for supervised training of automatic alignment models (271), creating alignment gold standards (97), teaching and learning new languages (190), and automatic alignment post-editing (95). Manually aligned corpora at the sentence or word level are valuable resources for several applications, such as fine-tuning multilingual language models, training automatic alignment models, cross-lingual annotation transfer, and inducing dynamic dictionaries. However, manual alignment is costly in terms of time and resources and requires domain experts fluent in the languages aligned and purpose-built annotation tools.

In this chapter, we present our work on Ugarit, a public web-based tool for manual annotation of parallel texts for generating translation alignment. We aimed to develop a user-friendly interactive interface to i) create a reading environment of aligned texts to help scholars and new language learners explore parallel texts and learn new vocabulary and their use in context, ii) create training data in the form of translation pairs to train an automatic word-level alignment system for historical languages, iii) create gold standard datasets to evaluate automatic alignment and machine translation systems. Ugarit is now widely used for learning new languages, especially historical languages, and as a reading environment for parallel texts.

Since the tool was made public, the number and variety of languages included by the users have steadily increased and have gone far beyond the original intent. At the

moment this thesis is being written, UGARIT includes over 872,000 translation pairs and over 30,000 aligned texts in 49 languages aligned by 660 active users.

In the following sections, we review the related works and describe the system design and alignment workflow. Then, we provide an overview of the visualization techniques used to visualize the alignment results. Further, we show the impact of UGARIT in research and teaching historical languages, and present how we derived the translation graph from the aligned translation pairs. Moreover, we present the results of three qualitative studies we conducted in collaboration with translation studies, digital classics, and philology experts to analyze the data available on Ugarit and annotators' behaviour. Finally, we discuss the usage limitations of UGARIT, possible improvements, and future development plans. This chapter is based on the following publications:

- **Yousef T**, Palladino C, Shamsian F, Foradi M. *Translation Alignment with Ugarit.* MDPI - Information. 2022 ;13(2):65. (272)

- Palladino C, **Yousef T**. *To Say Almost the Same Thing? A Study on Cross-linguistic Variation in Ancient Texts and their Translations*". Digital Scholarship in the Humanities Journal, 2023. (187)

- Palladino C, Shamsian F, **Yousef T**. *Using Parallel Corpora to Evaluate Translations of Ancient Greek Literary Texts. An Application of Text Alignment for Digital Philology Research.* Journal of Computational Literary Studies. 2022 Dec 13;1(1).(192)

- Palladino C, **Yousef T**, Shamsian F, Kanagawa N. *Ugarit: Translation Alignment Technologies for Under-resourced Languages.* In DH 2022. (193)

- Palladino C, Foradi M, **Yousef T**. *Translation alignment for historical language learning: a Case Study.* Digital Humanities Quarterly. 2021;15(3). (190)

- Palladino C, Zhang A, Foradi M, **Yousef T**. *How to Read All Languages: Translation Alignment with Ugarit.* In DH 2020. (189)

2.1 Related Work

Most word alignment annotation tools are either crafted for a specific project (99, 163), specific language pairs, or specific texts (31). The related works can be grouped into two main groups. The first group includes tools that offer an annotation interface to generate translation equivalents without any interest in visualizing the alignment results. The second group includes tools that provide both annotation and visualization interfaces. From the first group, we can mention the *Blinker Project* (163), which developed the first annotation tool for manual text alignment to align different versions

of the Bible in French and English at the word level. *PLUG Link Annotator* (PLA) has been developed to facilitate the creation of gold standards for evaluating word alignment systems (8). The Linguistic Data Consortium (LDC) has also developed the *LDC Word Aligner* to perform the manual word alignment for Arabic-English and Chinese-English parallel texts from newswire and broadcast mainly (99). *TagAlign* (48) allows users to annotate bilingual texts with a pre-defined tagset and create manual alignments at the sentence level. The tool was initially developed to align Brazilian Portuguese, and English parallel texts. Benner (31) developed a manual word alignment tool to create gold standard for Hebrew-English by aligning the Hebrew bible with its English translation.

On the other hand, *Yawat* (94), *Alpheios* (12), *SWIFT Aligner*(95), and *CLUE-Aligner* (27), which fall into the second group, enable users to create their alignments manually and offer various options for visualizing the aligned texts, such as side-by-side view, the interlinear text view (13), and alignment matrices (27, 94). UGARIT falls in the second group, it provides a manual annotation service and a reading environment for parallel texts. The aforementioned tools use different approaches to link words in the source text with their correspondents in the translation. We can distinguish three main methods:

- Texts are represented as a 2-dimensional alignment matrix. Annotators can select a cell in the grid to create an alignment between the corresponding row- and column-words (27, 47).
- Texts are placed on two parallel columns. Annotators can connect corresponding words of the source and translation texts by drawing lines between them (31, 95).
- Texts are placed side-by-side. Annotators can select the source words and their translation correspondents with mouse clicks (12).

The first two approaches are suitable for short text units (sentences, short paragraphs), where annotators can still capture the context while reading vertical texts, and the third approach is suitable for both short and long texts. Most tools mentioned above are limited to one alignment class between tokens in the original and its translation. Still, *CLUE-Aligner* distinguishes between two main categories, possible and sure alignments, and *Yawat* supports manual labelling of the alignment relations. *SWIFT Aligner* supports Part-Of-Speech and syntactic dependency manual annotation.

2.2 Development Process

The development of annotation tools is a challenging task. It requires a deep understanding of the underlying task and the user's needs and an experience in Human-computer interaction methods and approaches. Before working on UGARIT, we studied the related tools and defined their limitations, we also consulted numerous research papers and surveys (44, 45, 83, 111, 175) that reviewed and analyzed the existing annotation tools and defined design principles and usability recommendations, which helped us to build a primary vision of the tool.

The development of UGARIT has been achieved through the close collaboration of researchers from Computer Science, Digital Humanities, Classical Philology, and Translation Studies, aiming to gain a better understanding of users' needs. The development started in 2017 at the Alexander von Humboldt-Chair for Digital Humanities at Leipzig University. We followed the user-centered design principles during the iterative development process (Figure 2.1), and the usability recommendations for annotation tools (44), leading to a user-friendly, intuitive, easy-to-use tool.

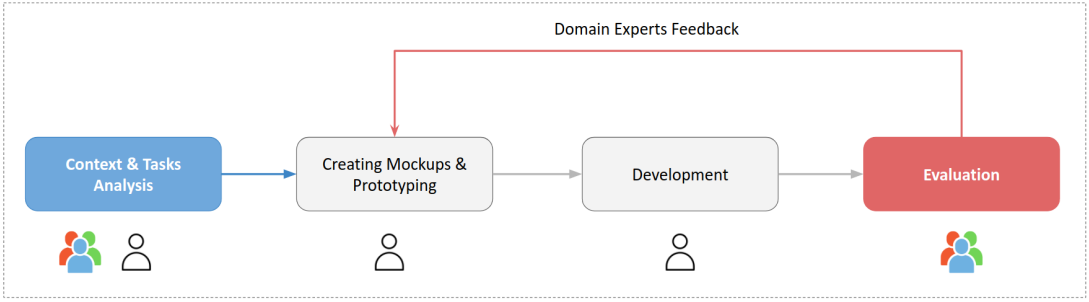


Figure 2.1: User-Centered design process.

The starting point was to define the usage context and analyze the user tasks and requirements. This was done through interviews and discussions with domain experts who explained their needs and illustrated the different usage scenarios. Domain experts defined a list of main features that must be provided in the tool:

- Users can annotate texts in any language, alphabet, or writing system.
- Users can annotate long text passages.
- Users can create different alignment types, one-to-one, one-to-many, many-to-one, and many-to-many.

- Time and clicks required to create the alignment should be minimized to the extent possible.
- Users can analyze and review their alignments via interactive visualization.
- Users can download and share their alignments in different formats.

Next, the development team prepared a set of mockups and prototypes that were discussed and approved with the domain experts. Most of our audience has no technical expertise; therefore, we decided to create a public web-based platform-independent tool that is easily accessible and does not require any prior technical knowledge for installation and use.

Later, the first implementation cycle started and resulted in the first product, which was tested and evaluated by domain experts. The improvement suggestions, problems, and difficulties faced by the domain experts were reported to the developer[1], who worked to solve these issues and enhance the tool's usability. After multiple development iterations, we could reach a stable and reliable version.

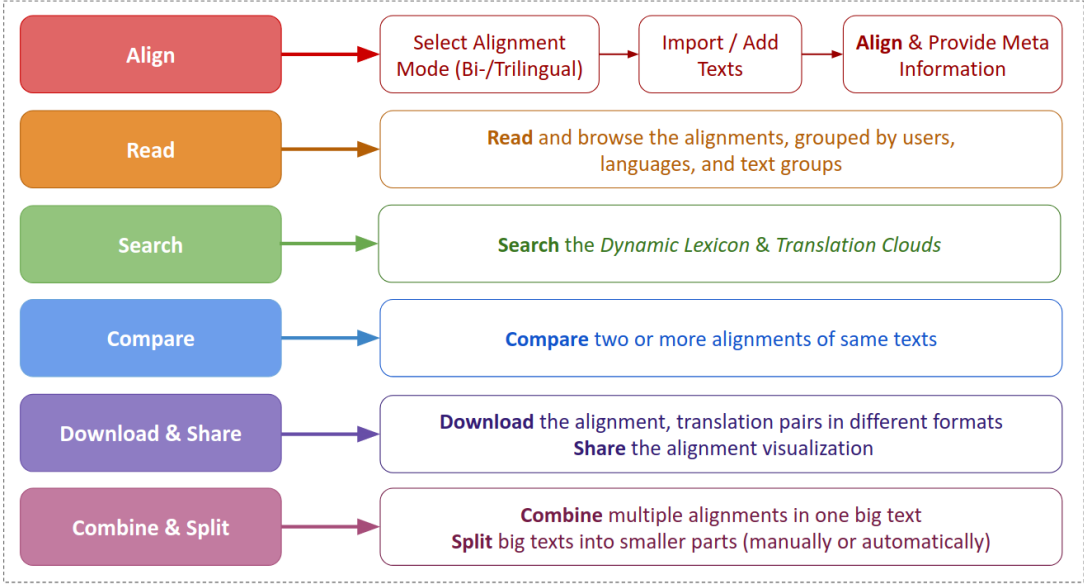


Figure 2.2: Overview of UGARIT's features and functionalities.

UGARIT is implemented in Php on the server-side and Javascript on the client-side; MySql and Neo4J are used to store the data. UGARIT supports UTF-8 text format, which allows users to align texts in many languages and different alphabets such as

[1]I was the only developer of the tool.

Aramic, Ancient Greek, and Coptic. It also shows Right-to-Left texts in the correct direction and provides automatic transliteration for non-Latin alphabets languages. The tool is not designed to be used on mobile devices; however, users can still create alignment for short passages with their mobile devices. For best use, it is recommended to use a computer.

Figure 2.2 summarizes the features and functionalities provided by UGARIT, users can align bi-and trilingual texts; combine their aligned passages in one big text or split a long aligned text into smaller units. Moreover, users can download their alignments, share them, and integrate them into their websites and blogs. Further, Ugarit provides a reading environment for aligned parallel texts, where users can read, search, and compare the aligned passages.

2.3 Alignment Workflow

Annotation is the process of adding information to a text at some level (201). Manual Annotation is a nontrivial task requiring intense focus and attention. Like any chore, it becomes boring after a short while. Manual translation alignment, in particular, is a daunting task; it requires moving eyes between two texts in different languages, finding translation correspondences, and linking them.

When we designed UGARIT interface, we tried to make it easy to use even for users with no experience with annotation tools. The first step is to create an account on UGARIT or log in if the user has already created an account and verified it. To start a new alignment, users have to choose between bi- or trilingual text alignments. Then, users can upload their parallel texts in plain text format or use the canonical text service (CTS) URNs to import texts from the *Perseus Digital Library CTS* [1]repository (22). Alongside, the languages of the texts must be selected. Next, texts will be tokenized and prepared for alignment (Figure 2.3a).

The alignment process is designed to be as simple as possible; we tried to minimize the mouse clicks needed to create and save the aligned units. In the case of bilingual text alignment, the panel is split into three main columns, two columns for the parallel texts and the right-side column for displaying the aligned translation pairs. A progress bar is located on the top of each language panel to show the alignment coverage so far (Figure 2.3b). UGARIT also offers trilingual alignment, where users can align three parallel texts in three different languages (Figure 2.4).

UGARIT uses white space and a punctuation tokenizer to split the text into a sequence of tokens; each token is assigned an id. The current version of Ugarit does not

[1]cts.perseids.org

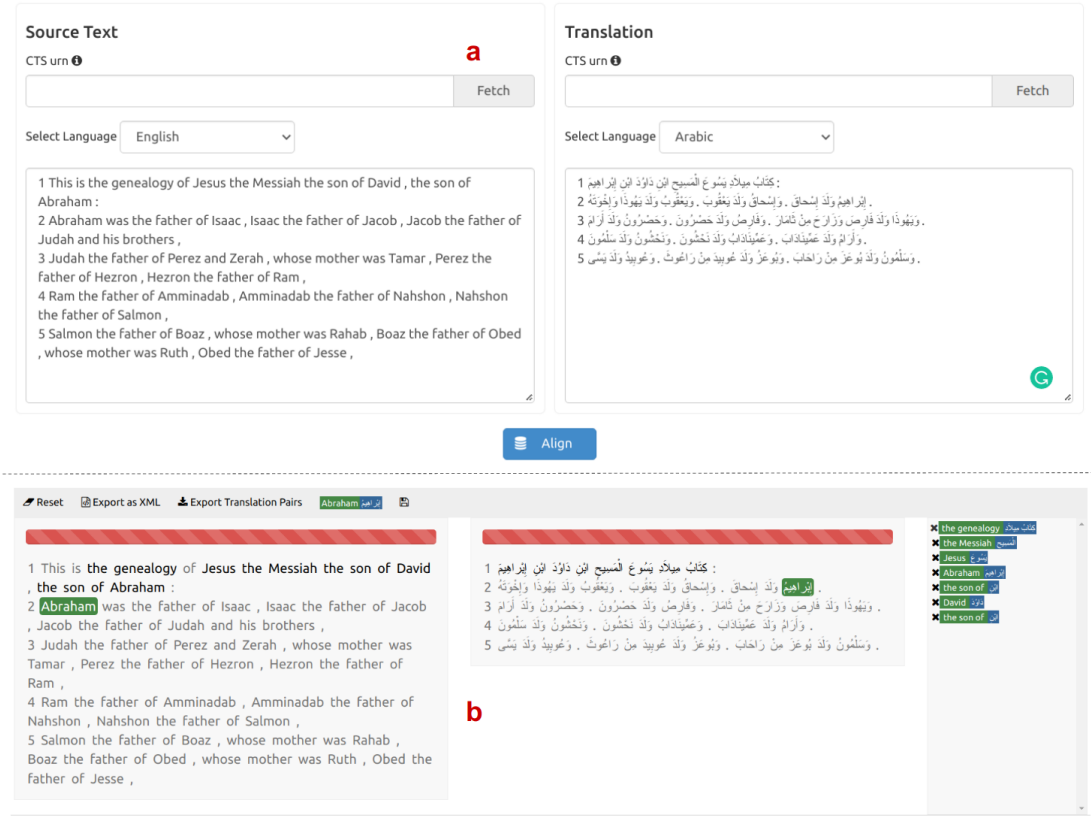


Figure 2.3: UGARIT bilingual alignment editor. The manually created translation pairs are located on the right.

support text post-editing, i.e., once the text is tokenized, it is not possible to update or add new tokens. Each token has two states, namely aligned and unaligned. Unaligned tokens are colored in grey, and the aligned ones are in black.

To align a translation pair, users must select a word/phrase from the original language and then select the corresponding word/phrase in the translated text. Selecting a token with a mouse click will highlight it in green. Clicking a selected token will deselect it and remove the highlighting. The paired tokens will be automatically saved when the user starts aligning a new pair or clicks on the *save* icon. On the right side of the editor, users can see all aligned pairs and have the option to edit or delete any pair. The UGARIT editor allows users to create all types of alignment word-to-word(1-1), word-to-phrase(1-N), phrase-to-word(N-1), and phrase-to-phrase(M-N) alignments. The translation pairs list can be exported in XML or tabular format.

The resulting translation pairs are automatically stored in the database and can be exported in XML or tabular format. Users can also decide whether the alignment is

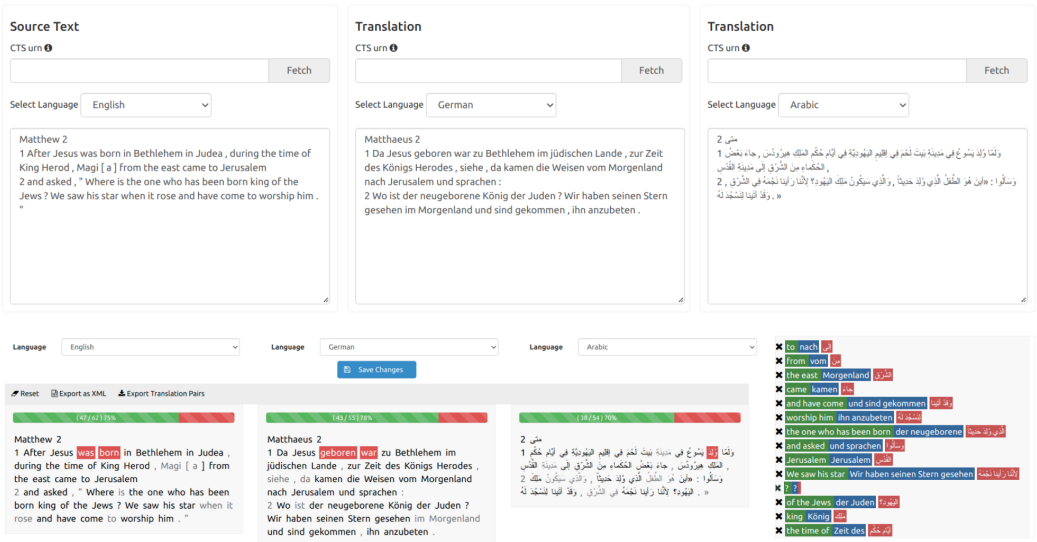


Figure 2.4: UGARIT trilingual alignment editor shows three parallel texts in English, German, and Arabic.

publicly visible on the website or keep it private. Finally, users must provide information about the texts, such as the title, translator, and a short description. Once the user saves the alignment, it will appear on the home page in the *New Alignments* panel.

Alignment Guidelines

Annotation guidelines are a key component of any collaborative annotation work; they ensure annotation consistency and reduce annotation errors. The same applies to translation alignment. However, defining alignment guidelines is challenging; it requires a deep understanding of the linguistic structure of both languages and the relations between them. In general, UGARIT doesn't impose any alignment guidelines, It allows users to tailor their own guidelines according to their projects or alignment purpose. But we share some recommendations and refer to standard guidelines used for the creation of some alignment gold standards (127, 147, 162).

We conducted experiments to measure the impact of the alignment guidelines on the alignment process. The results showed that following alignment guidelines would increase the inter-annotators agreement by 5% to 26% (see 2.5.5.2). However, it is not possible to create guidelines for all language pairs in UGARIT since alignment guidelines need to consider multiple aspects of both the target and the source language at once and should be tailored individually for each language pair by experts.

2.4 Visualization Techniques

Visualization of translation alignment plays an important role in understanding and interpreting the relation between texts and their translations (267) and help analyzing translation equivalents, exploring their usage contexts. UGARIT offers various approaches to visualize aligned texts, translation pairs, and the derived *Dynamic Lexicon*.

2.4.1 Languages Graph

The graph is placed on the tool's home page [1] to give users a quick overview of the languages currently hosted on UGARIT and how they relate (Figure 2.5).

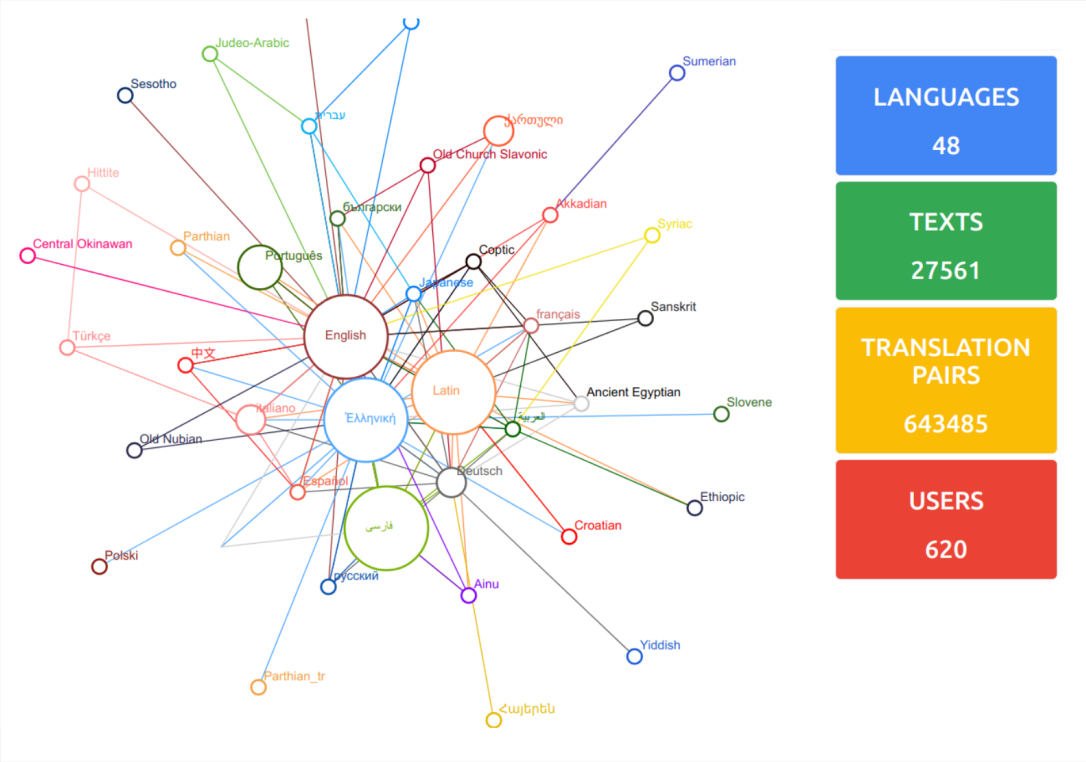


Figure 2.5: Languages graph.

Each node represents a language. The size of the node is proportional to the number of texts in this language stored in the database. Each language is assigned a different color, and nodes are labeled with the language names in the original form. The connection between two vertices means that there are translation alignments between texts in these

[1] http://ugarit.ialigner.com

two languages, and the thickness of the line reflects the number of aligned translation pairs. Clicking a node will show all texts in the language it represents, whereas clicking a link between two vertices will load all aligned texts in these two languages.

2.4.2 Aligned Texts

The side-by-side view is the most intuitive and straightforward approach to display parallel text alignment at different levels. It is also widely used to visualize collation and mono-language alignments (267).
Texts are placed alongside each other and a coloring schema is used to distinguish between aligned and unaligned tokens. Since most tokens are aligned, we used black for aligned tokens and red for the unaligned ones to draw user attention. The alignment between corresponding tokens on the parallel sides is visualized via highlighting. When the user hovers an aligned word, the hovered word/phrase and the paired tokens will be simultaneously highlighted with red color.
A progress bar is located under each text to give the user an overview of how many tokens are aligned and their percentage; the aligned part is colored with green, whereas the unaligned part is colored with red (Figure 2.6).

Embed: This option allows users to generate a link that can be used to share and embed the visualization of their alignments on any blog or website. Furthermore, and users can select which components can be included in the embedded text (progress bar, title, annotator info, text info) (Figure 2.6-1).

Translation Pairs: This option shows an aggregated list of the translation pairs extracted from the aligned text, users can download them in JSON format (Figure 2.6-2).

Alignment Statistics: This option shows statistics of the different alignment categories. This chart provides valuable information on the alignment quality, user's language knowledge, and the relation between the aligned languages (Figure 2.6-3).

Alignment Matrix: with this option, users can view the alignment in the form of a grid: the source text tokens are located on the horizontal axis, and the translation tokens are located on the vertical axis. The blue dots represent alignment between the corresponding column and row tokens. The diagonal dots indicate one-to-one alignments, the vertical ones indicate one-to-many alignments, and the horizontal ones indicate many-to-one alignments (Figure 2.6-4).

Transliteration: UGARIT includes texts in various languages with different alphabets. For better readability, especially for new language learners, UGARIT offers an

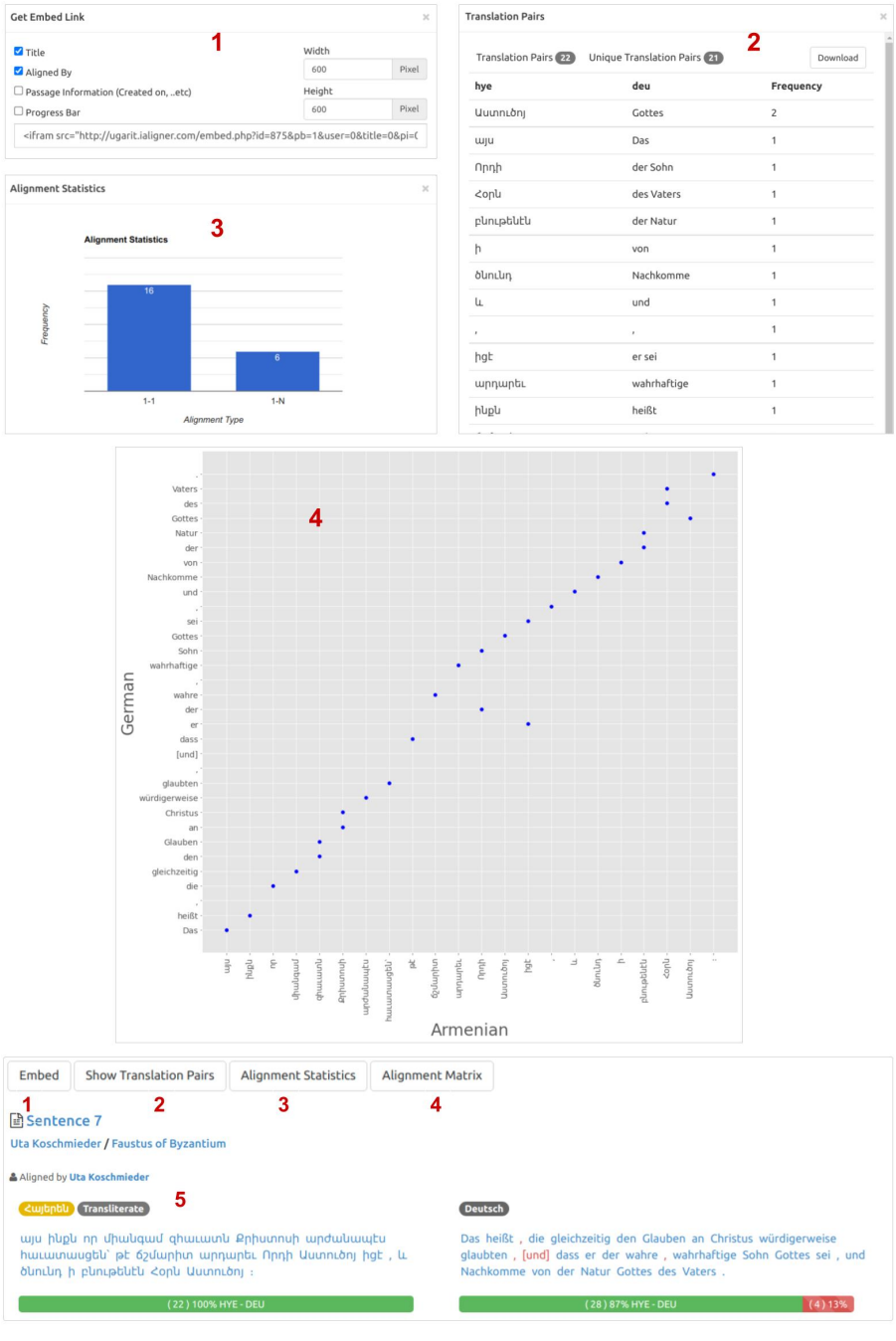


Figure 2.6: Side-by-side visualization of bilingual aligned texts.

automatic transliteration for non-Latin alphabets languages, which is visible when the pointer hovers the aligned word. This feature is currently available for Greek, Arabic, Persian, Armenian, and Georgian (Figure 2.6-5).

Combine/Split Aligned Texts: UGARIT allows users to merge multiple aligned texts into one bigger text. It also lets users split a long aligned text into smaller units (sentences/paragraphs). This feature is beneficial since annotators prefer to align long paragraphs over short ones to avoid copying and pasting them multiple times. On the other hand, splitting long aligned text is useful when users want to create shorter aligned units that can be used later to train or evaluate machine translation or translation alignments models.

2.4.3 Translation Pairs & The Dynamic Lexicon

UGARIT provides a search function to enable users to look for a word or phrase and get detailed information about how this word/phrase occurs in the UGARIT texts collection and how it is translated in different languages. Moreover, UGARIT visualizes the dynamic lexicon search results in two approaches:

- Tree View is a classic branching view that enables users to navigate the hierarchy to filter the results set until they reach the desired subset. Figure 2.7a shows how UGARIT visualizes the search results as three levels tree view. The query word is located at the first level as the root word, corresponding translations are grouped by languages, and languages are placed as nodes at the second level. The aligned translations are placed at the third level. The language nodes are initially collapsed for clarity, and users can expand them to explore the translations by clicking the language label.

- Radial Cluster Dendrogram is similar to the tree view mentioned above; it shows the translation equivalents in the form of a rooted tree and all leaf nodes are placed at the same depth (Fig 2.7).

These two views illustrate the idea of the *Dynamic Lexicon* (265), which is an automatically derived from aligned texts at the word/phrase level. *Dynamic Lexicon* uses triangulation to create new dynamic lexica using pivot languages based on the assumption that two words/phrases are likely to be translations if they are translations of the same word in a third language (7). In the two examples shown in Figure 2.7, triangulation considers all leaf nodes synonyms (if they are in the same language) or translation equivalents since they share the exact English translation.

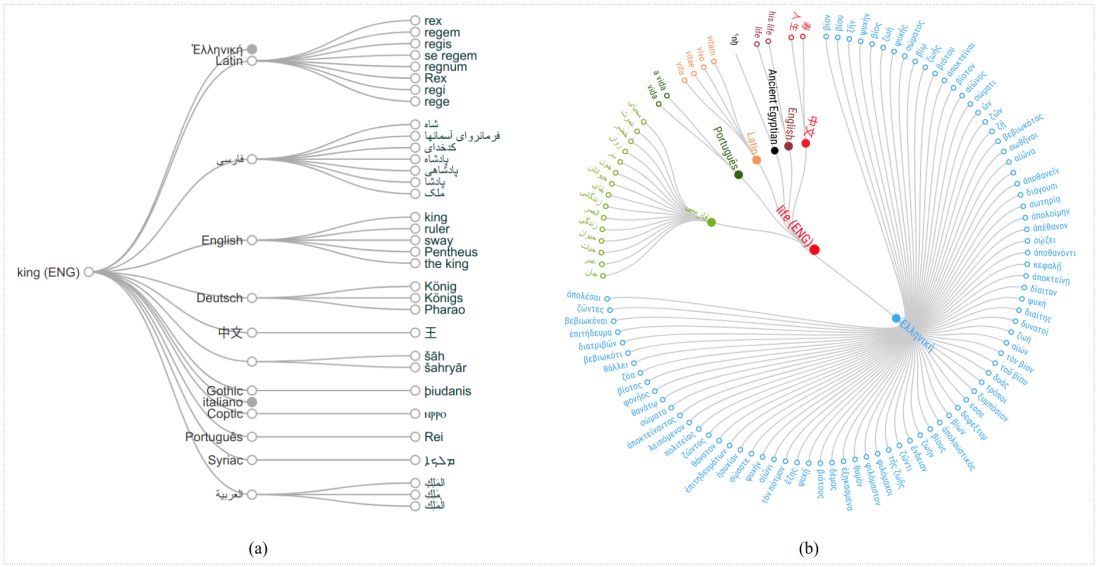


Figure 2.7: Visualization of translation pairs search results. a) Tree View, b) Radial Cluster Dendrogram View.

2.5 Ugarit in Practice

2.5.1 Research and Pedagogy

The complexities involved in the translation alignment process often reflect the dynamic relationship between the original texts and translation, and between different languages. Ugarit provides an environment where users can engage with those complexities analytically, on both a small and big scale.

On a small scale, Ugarit offers a visually powerful environment, where the reader can analytically compare texts token by token, and at the same time, observe the results through interactive visualization. Moreover, the opportunity to publish the results online provides a way to be part of a broader conversation on the reception and significance of a text over time. For these reasons, Ugarit is currently used in studies on translation and reception and language teaching, particularly for Ancient Greek and Latin, across the world. In pedagogy, translation alignment is often integrated with grammatical and syntactical observations to emphasize the complex interplay between the language of the translation and the target language, and students are assigned various alignment tasks and exercises to empower the analytical approach to the text (184, 190). On a bigger scale, Ugarit also provides manually aligned parallel corpora across languages that had never been compared before, such as Coptic, Ancient Greek, Arabic, Persian, Latin, Egyptian, Georgian, etc. The analysis of these aligned datasets and recurrent patterns in

Institution	# Texts
Leipzig University - Germany	2903
Furman University - USA	1211
São Paulo State University - Brazil	763
Catholic Memorial School - USA	697
Tufts University - USA	345
University of Zagreb - Croatia	183
Liceo G. Peano-Tortona - Italy	100
Augustana College - USA	82
University of Göttingen - Germany	79
Ivane Javakhishvili Tbilisi State University - Georgia	65

Table 2.1: Top ten institutions from which UGARIT users come.

word matching can provide insights into how cross-linguistic and cross-cultural dynamics are affected by different language structures, cultural differences, text genres, and even language proficiency (186).

Teaching

UGARIT has been used for vocabulary learning, morpho-syntactic comparisons and learning assessment both in the classroom and in self-study. One case in point is an online course for teaching Ancient Greek to Persian speakers (216), where UGARIT was used extensively. After a five-minute tutorial, all course participants could successfully register and use UGARIT to align translation. During this course, UGARIT was used as a reading environment for parallel texts by the educator during the class, as a vocabulary learning tool by the students, and as an assessment/practice environment where the educator could assess and examine the students' understanding of the text. Furthermore, the trilingual alignment provided a suitable environment for using a third language as a bridge between the source language and the target language, which in the case of the Ancient Greek course would be English.

To the best of our knowledge, Ugarit have been regularly used for teaching classical and ancient languages at Leipzig, Tufts , and Furman Universities, and Boston Latin Academy, and it is currently under experimentation at UNESP Brazil (Araraquara), University of the Sciences Philadelphia, Dickinson College, and across some high schools in Italy, Germany, and the USA.

Further, Didakta[1], a platform for data-driven learning of the Ancient Greek Language, uses the embed option available on Ugarit interface to integrate the alignments created on Ugarit into the platform. Table 2.1 shows the top ten institutions from which Ugarit users come sorted by the number of aligned texts.
Ugarit has also been present during all Sunoikisis Digital Classics[2] courses since 2018 and in several Digital Classicist London Seminars[3] since 2020. My colleagues [4] and I presented translation alignment with Ugarit, usage scenarios, applications, and its impact on research and pedagogy in over eight sessions live streamed via Youtube and still available on the Sunoikisis DC and Digital Classicist London Seminars Youtube channels[5] [6]. Further, we organized two workshops during the Digital Humanities conferences DH2020 (189) and DH2022 (194). In the DH2020 workshop[7], we illustrated the importance of translation alignment in the field of language learning. We also provided a short theoretical overview of the principles of translation alignment, together with a hands-on tutorial on Ugarit.
In the DH2022 workshop[8], participants learned the fundamentals of Translation Alignment and how to use Ugarit. We aimed to expand our community and available datasets by targeting East Asian languages in particular. The workshop included 12 participants from different fields (such as Applied Language Studies, Translation Studies, History, Digital Humanities, Linguistics, and NLP), and the majority hold a PhD.

Research

The *Digital Rosetta Stone Project* [9] utilized Ugarit to align the Hieroglyphic, Demotic, and Ancient Greek texts of the Rosetta Stone with the English and German translations. The project linked the text passages on the photograph of the Rosetta Stone with the corresponding manually aligned passages on Ugarit. Further, Maia Shukhoshvili[10] aligned her translation of Plato's "*Theaetetus*" from Ancient Greek to Georgian using

[1] `https://didakta.vercel.app/`

[2] An International Consortium of Digital Classics Programs, `https://sunoikisisdc.github.io/`

[3] Seminar series affiliated with the Digital Classicist are hosted by the Institute for Classical Studies, University of London.

[4] Chiara Palladino, Farnoosh Shamasian, and Maryam Foradi

[5] `https://www.youtube.com/channel/UCjwSzxMtwDaPu6JvG1GquOw`

[6] `https://www.youtube.com/channel/UCIamtu1Z62wL5XRk2mE8HKw`

[7] The workshop was entitled: *"How to Read All Languages: Translation Alignment with Ugarit."*

[8] The workshop was entitled: *"Ugarit: Translation Alignment Technologies for Under-resourced Languages"*

[9] `http://rosetta-stone.dh.uni-leipzig.de/`

[10] Associate Professor at East European University, who translated Plato's trilogy: "Sophist", "Statesman", "Theaetetus" from Ancient Greek in Georgian.

UGARIT. She also introduced her alignment strategy and summarized her observations and conclusions (220). The aligned corpus comprises 65 texts with over 20.000 aligned translation pairs. Also, Several articles have been published discussing the results of translation alignments of texts in various languages(82), its applications(185) and its pedagogical potential (190).

We also conducted several experiments and analyses on the data available on UGARIT, which will be discussed in section 2.5.5. In addition, several Bachelor, Master, and PhD Students used Ugarit in the context of their finished or in-progress theses (73, 78).

2.5.2 Translations Graph

UGARIT is characterized by its simplicity and ease of use. For this reason, it has been used by many users and in several projects, which led to continuous growth in the number of aligned texts and translation pairs. We realized that we could achieve more benefits if we represented the translation pairs as a graph, allowing us to create new translation pairs and connect languages that have never been connected via exploiting the graph structure and triangulation. The proposed model can be described as follows: Let $G = (V, E)$ be a translations-graph. where V is a set of all words/phrases in the the UGARIT aligned texts, and E is a set of all translation pairs. Let $v_1 \in V$ be a word/phrase in language l_1, and v_2 be a word/phrase in language l_2. $\{v_1, v_2\}$ be a translation pair that occurs n times in the data set.
Each node $v \in V$ has four attributes, an internal id id, label t, language l, and n the frequency of this word/phrase in the whole dataset $v\{id, t, l, n\}$.
The *translation* relation between $v_1\{id_1, t_1, l_1, n_1\}$ and $v_2\{id_2, t_2, l_2, n_2\}$ can be represented as an edge $e \in E$, e encapsulates a single value n which indicates the frequency of this translation pair in the dataset. We have used the Neo4j database (65) to model the translations graph.

Translation Clouds

Visualizing the translation graph as a connected graph was not possible due to the large number of nodes and crossing edges, which would affect the translations' clarity and readability. The solution was to keep the labels and remove the lines that connect them, resulting in a word cloud representation.
Translation cloud shows the translation equivalents to a given word, these equivalents are extracted from the translation graph up to the fifth level. Figure 2.8 shows the translations cloud of the root word *city*.

Figure 2.8: Translations cloud of the root word *city*.

The view consists of three components, the translation cloud, the languages bar above it, and the concrete translation path on the right side. The search word is located at the cloud center, surrounded by its equivalents in other languages. Further, translation equivalents are grouped by language, and every language is assigned a unique color. However, languages are listed in the languages bar; clicking on a language label will hide all translation equivalents in other languages and keep the ones in the selected language. Labels are displayed in different font sizes according to the distance to the root word. Since the translations cloud visualizes connected translation equivalents up to the fifth level, the distance varies between 1 to 5; the lower the distance value, the higher the font size. Moreover, clicking a translation label will show the path between the selected word and the root. Also, a tooltip is used to show information about the label, such as language and frequency.

Using the Neo4j graph database at the back-end to store the translation pairs has shown many advantages; it reduced the response time of search queries, enabled us to perform complex queries, and facilitated the dynamic lexicon production.

2.5.3 Creating Training Data

When we started working on UGARIT, our main objective was to provide a crowd-sourcing user interface to easily create training data and develop an automatic alignment system for historical and low-resourced languages. Unlike modern languages, ancient languages lack big parallel corpora that could be used to train automatic models. The Bible is the most popular parallel corpus available in numerous historical and modern languages. Still, it is limited to a specific period, text genre, and vocabulary. The Bible can be used for *Unsupervised* training, which uses parallel sentences or paragraphs with no further information about the correspondence among the tokens of the source text and its translation. However, *supervised* training, which is more accurate, requires manually annotated parallel sentences at token level, which makes UGARIT a great resource of accurate training data for supervised training.

The tool has been used in several contexts by a wide range of people of different levels of expertise, students, teachers, scholars, researchers, translators, ..etc. Therefore, in order to obtain high-accuracy training dataset, we defined a list of 25 experts and teachers who we know personally and we know their level of experience to use their alignments as training data. Table 2.2 provides an overview of the dataset we used for the supervised training. It contains over 3800 texts in 16 languages and 24 language pairs, with over 190.000 manually aligned translation pairs. However, the dataset can be expanded by including more users, which requires extensive research on their background and knowledge.

2.5.4 Creating Alignment Gold Standard

Gold standard datasets are essential for evaluating the performance of automatic alignment models. Gold standards are manually aligned parallel sentences performed by multiple domain experts (at least two experts) according to predefined alignment guidelines to ensure consistency and homogeneity of the alignments. We used UGARIT to generate alignment gold standard for three language pairs Ancient Greek-English, Ancient Greek-Portuguese, and Ancient Greek-Latin. Two experts aligned the same set of parallel sentences for each language pair. Then, we merged the alignments and calculated their Inter-Annotator Agreement. In Chapter 4, we discuss our work on generating alignment gold standards and guidelines in more detail.

2.5.5 Qualitative Analyses and Experiments

UGARIT is a valuable resource of manually aligned texts, allowing us to conduct some experiments and analysis on the data available on Ugarit, and draw conclusions

Language-Pair	# Texts	# Translation Pairs
FAR-ENG	563	59,407
GRC-ENG	944	31,018
GRC-LAT	507	21,676
GRC-KAT	64	16,909
GRC-FAR	551	10,644
GRC-ITA	106	10,327
GRC-POR	344	8,915
LAT-ITA	61	7,071
LAT-POR	25	5,507
LAT-ENG	67	3,991
GRC-DEU	64	3,758
GRC-HRV	156	2,754
EGY-DEU	67	2,090
COP-ENG	39	1,962
EGY-ENG	66	1,870
GRC-ARA	14	1,258
FAR-DEU	75	928
GRC-ESP	1	884
DEU-HYE	66	823
COP-FRA	12	471
DEU-ENG	13	261
GRC-JPN	5	238
AIN-FAR	3	95
GRC-COP	3	52
Total	3,828	193,196

Table 2.2: An overview of the training dataset.

on annotators' behaviour. In the following, we present three qualitative studies we conducted in collaboration with experts in translation studies and digital classics and philology, namely, Chiara Palladino and Farnoosh Shamasian.

2.5.5.1 Experiment 1: A study on cross-linguistic variation in ancient texts and their translations

UGARIT hosts a large and diverse corpus of crowd-sourced parallel texts in several languages, some of which have never been consistently aligned before. Moreover, the texts were aligned manually, which gave the annotators complete control over their choices. This offered an opportunity to investigate what factors influence the establishment of word correspondences between different languages, and whether these trends could tell us something about the relationship between ancient texts and their translations. We analyzed a subset of the corpus, selecting reliable users, and measured various kinds of translation pairs ratios to investigate the resulting patterns.

Corpus selection:

When we selected the dataset, we tried to minimize inconsistency and noise by selecting exclusively the alignments made by reliable expert users, such as teachers and researchers. Moreover, we considered that all the chosen subsets included scholarly translations, primarily published in authoritative editions, from relatively recent periods (mostly 18th-19th century). Further, we filtered out language pairs which have less than 1000 translation equivalents. For consistency, we limited our observations to bilingual alignments.

Our most extensive subset consists of Classical Persian-English with 60k translation pairs, Ancient Greek-English 24k and Greek-Georgian 17k. All the considered subsets included between 500 and 30 independent and complete alignments. Because of the diversity of the corpus, some subsets are considerably bigger than others, even though the number of translation pairs was sufficient to reveal some patterns. For this reason, we refrain from making comparative assessments and consider each language pair as an independent unit.

Methodology:

We measured translation pairs ratios distinguishing between one-to-one (1-1) pairs, one-to-many (1-N), many-to-one (N-1), and many-to-many (N-N). We also measured the length[1] ratio between the two languages (L1/L2: L1 being the source, L2 the target), considering the number of words of each corpus, excluding punctuations and diacritics. However, the datasets presented in this experiment were not created under similar conditions, even though the annotators were experts, there were several of

[1]Token-based length.

		1-1	1-N	N-1	N-N	J-S Divergence	J-S Distance
GRC-ENG	User 1	63.11%	31.43%	1.81%	3.65%	0.01046	0.10226
	User 2	63.41%	31.72%	2.27%	2.60%		
GRC-LAT	User 1	88.83%	8.03%	2.68%	0.46%	0.01613	0.12701
	User 2	89.21%	6.39%	2.96%	1.44%		
GRC-POR	User 1	72.52%	20.19%	2.76%	4.52%	0.02910	0.17059
	User 2	74.59%	21.55%	1.77%	2.09%		

Table 2.3: Jensen-Shannon divergence and distance for three Gold Standard datasets created under ideal conditions.

them, operating during various periods of time, for different purposes, and on different texts. To make sure our datasets were sufficiently homogeneous, we performed tests for statistical divergence to verify whether random subsets would still show similar patterns. We picked random subsets of 10-30 complete alignments, depending on corpus size, calculated their alignment ratios, and measured them against the main corpus using the *Jensen-Shannon Divergence*, a statistical divergence measure which calculates the similarity between two probability distributions, according to this formula:

$$JSD(P||Q) = \tfrac{1}{2}D(P||M) + \tfrac{1}{2}D(Q||M)$$

$$\text{where } M = \tfrac{1}{2}(P+Q) \quad , \quad D(P||M) = \textstyle\sum_{x\epsilon X} P(x) log\left(\frac{P(x)}{M(x)}\right)$$

$D(P||M)$ is the Kullback–Leibler divergence and the square root of the *Jensen–Shannon Divergence* is the *Jensen–Shannon Distance*. We based the calculation on the indicators of 1-1, 1-N, N-1, and N-N alignments.

One of the problems of running such an evaluation is that we do not have a baseline reference that gives us an idea of what a "good" divergence looks like. To solve this issue, We measured the divergence and the distance between the annotators who annotated the Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Portuguese alignment gold standard datasets because these aligned corpora were created under optimal conditions, and experienced annotators aligned the same corpus following the same alignment guidelines. The results indicate a distance between 0.1 and 0.17 (Table 2.3), We consider this range as baseline for the annotation divergence.

Results

In our analysis, we considered two groups of subsets, namely, *G1* the English translations of ancient languages and *G2* Ancient Greek translated into other languages. An extensive and detailed qualitative analysis of the obtained results from linguistic and philological

Languages	Texts	TPs	L1/L2	1-1	1-N	N-1	N-N	NA L1	NA L2
FAR-ENG	561	59,308	53.68%	57.72%	38.45%	2.29%	1.54%	3.83%	24.56%
GRC-ENG	741	23,908	76.75%	64.84%	28.89%	2.01%	4.26%	12.96%	14.76%
LAT-ENG	41	2,654	64.28%	57.20%	36.17%	1.62%	5.01%	17.60%	23.26%
COP-ENG	27	1,597	67.17%	42.64%	47.90%	3.38%	6.07%	2.88%	2.60%
GRC-ENG Poetry	157	3,415	81.22%	57.36%	34.58%	1.52%	6.53%	8.29%	10.95%
GRC-ENG Prose	320	11,886	75.54%	70.54%	24.28%	1.79%	3.39%	12.44%	12.73%

Table 2.4: Alignment ratios of various subsets of translations from ancient languages into English. NA L1, NA L2 indicate the percentage of not aligned tokens in the source text (L1) and translation (L2)), respectively. The last two rows compare a subset of poetry and a subset of prose from GRC-ENG.

perspectives is performed by Chiara Palladino and published in our paper (187). In the following, we provide a numerical analysis of the experimental results:

English translations of ancient languages. The predominance of English translations in the corpus allowed us to establish the first group of alignments with English as a translation language for various ancient texts. We measured link types occurring between English translations and texts in three Indo-European languages (Classical Persian, Ancient Greek, Latin), and one non-Indo-European language, Coptic. Overall, in the first three subsets considered, which all include Indo-European languages, there are common tendencies in which the source language maps against the target (Table 2.4) . In Persian and in Latin, we have 57% of the cases where one word maps against one word in the translation, while in Greek, the ratio goes up to 64.84%. On the other hand, there is a strong tendency to map one word in the source language with multiple words in the translation: the highest point here is reached by Persian, where the ratio is 38.45%, while in Greek it is 28.89% and 36.17% in Latin.

The opposite of this dynamic almost never occurs, as multiple words in the source are only matched against one word in the translation in about 2% of the cases. The cases where multiple words in both texts are matched as groups are also relatively marginal, although more significant in percentage for Greek and Latin than for Persian. The comparison of 1-N and N-1 ratios suggests that the linguistic structure of the languages plays a part in this phenomenon. Ancient Greek, Latin, and Classical Persian are highly inflected languages, which naturally construct meaning by adding morphemes to words. Another relevant difference is shown by the length rate, measured in number of tokens, between the source text (L1) and the translation (L2): the value here is 0.64 for Latin and 0.53 for Persian. This ratio indicates that the translations of Classical

Languages	Texts	TP	L1/L2	1-1	1-N	N-1	N-N	NA L1	NA L2
GRC-ENG	741	23,908	76.75%	64.84%	28.89%	2.01%	4.26%	12.96%	14.76%
GRC-ITA	23	2,552	81.80%	61.79%	29.86%	4.70%	3.64%	9.19%	6.71%
GRC-FAR	206	5,228	73.68%	54.80%	30.62%	8.05%	6.52%	8.94%	12.15%
GRC-LAT	317	7,850	105.47%	89.87%	5.27%	4.25%	0.60%	19.27%	18.00%
GRC-KAT	64	16,909	110.57%	53.07%	15.21%	22.28%	9.44%	0.16%	0.12%

Table 2.5: Alignment ratios of various subsets of translations from Ancient Greek.

Persian literature used almost twice as many words as the source: this phenomenon is peculiar to the FAR-ENG subset.

Ancient Greek translated into other languages. The second biggest corpus consists of aligned texts in Ancient Greek against translations in various languages, including English, Italian, Modern Persian, Latin, and Georgian.

Table 2.5 shows a high ratio of 1–1 TPs ($\approx$ 65%), and a significant 1-N ratio ($\approx$ 29%)in the Ancient Greek-English subset, reflecting grammatical characteristics and the inflected nature of Greek, in contrast with the less inflected target language. Further, the results reveal that the mapping of Ancient Greek onto Latin is almost perfect: the 1–1 alignment ratio is by far the highest in the entire corpus (89.87%), with marginal quotas for 1-N (5.27%), N-1 (4.25%), and almost no N–N (0.60%).
In Greek-Persian, the 1–1 ratio and the length ratio are the lowest of the Indo-European subsets, at 54.8% and 73%, respectively.
In Greek–Georgian, the pattern of 1-N and N-1 ratios that were regularly observed in Indo-European translations is inverted: one word in Greek is translated into multiple words in Georgian only in 15.21% of the cases, while one word in Georgian can render multiple words in Greek in 22.28% of the cases.

The results suggest the presence of general patterns that do not depend on the presence of guidelines or individual choices. Further, the statistical divergence indicates that different users who aligned various texts at different moments had common tendencies in establishing the four alignment categories. The experiment provides insight into the impact of linguistic and cultural factors in manual translation alignment and crowd-sourced annotation: the type of language pair, the origin of the translations, and cultural interplay between textual traditions can all be determining factors in the creation of aligned corpora.

2.5.5.2 Experiment 2: Comparing Annotators' Behaviour

The assumption behind using a collection of manually aligned corpora is that annotators, particularly in the context of citizen science projects, will be consistent and perform their alignments by respecting a set of predefined rules to guide them in creating correspondences across languages. However, several linguistic and contextual constraints affect the selection of translation pairs. For instance, how to align wordplays that are translated in different ways linguistically and syntactically? How to work with allusions and semantic complexities that cannot be exhausted by the simple operation of matching tokens? Moreover, there are contingent factors to consider when a diverse group of users is involved in the collection of aligned pairs, including their familiarity with the given text and context, and their linguistic background.

In this experiment, we demonstrate that annotator behavioural patterns are not significantly affected by following *Alignment Guidelines*, and they tend to make decisions based on a set of contingent factors that cannot be entirely addressed in advance through guidelines. We propose a comparison of various users, and try to analyse the reasons behind different results in their alignments.

Methodology. We collected alignments performed by specific groups of users, divided in two main categories: students with instructors, and expert users working individually. We observed behavioral patterns of each group, particularly considering the following contingent factors:

- The annotator's proficiency in the language (expert to student).
- Their native language and linguistic background.
- Their individual alignment choices (preference towards 1-1 pairs, etc.).
- The context where the annotation was performed.
- The linguistic peculiarities of the texts aligned.
- The translation style (literal, semantic, free, etc.)
- The use of externally imposed guidelines.

For each group, we compared the ratios of the different translation pairs, we also considered the annotation agreement among the users, measuring where there was overlap and where they made individual choices.

Cast Study 1: Instructors and language students.

Since Ugarit is used by many language instructors, particularly in the area of Ancient Greek and Latin, we collected data from groups of students who aligned the same passages for class work and assignments, and measured the levels of agreement and disagreement between them and with the instructor. In this experiment, the students were not given proper alignment guidelines, but operated under the control of a language expert. We considered groups at two levels: advanced and intermediate. We considered three groups:

- *Group 1*: two Master students with similar levels of proficiency in Ancient Greek, against the instructor, aligning a passage from Herodotus, *Histories* against a modern English translation (Figure 2.9A).

- *Group 2*: three Master students with similar levels of proficiency in Ancient Greek against their instructor, aligning a passage in Ancient Greek from Plato's *Symposium*, against a modern English translation (Figure 2.9B).

- *Group 3*: three students with **intermediate** levels of proficiency in Latin, against the instructor, aligning two passages from Apuleius' *Metamorphoses* against modern English translations (Figure 2.9A).

We can summarize our observations in the following points:

- Less expert users have more tendency towards lexical error in the establishment of the translation pairs. This is less likely to happen with advanced students: the lower the level, the more likely it occurs.

- Less language expertise results in more inconsistency in the alignment strategy, e.g. in the establishment of grammatical groups to align together.

- A significantly lower-than-average one-to-one alignment rate indicates less expertise and more discomfort with the text, especially when not compensated by higher one-to-many and many-to-many ratios.

- More expert users have a higher tendency to 1-1 alignments, compared with less proficient users.

Cast Study 2: Expert users with similar backgrounds

In the experiment, we had three user groups with three different languages pairs, we tried to cover more languages (Persian-English, Ancient Greek-Italian, Ancient Greek-English):

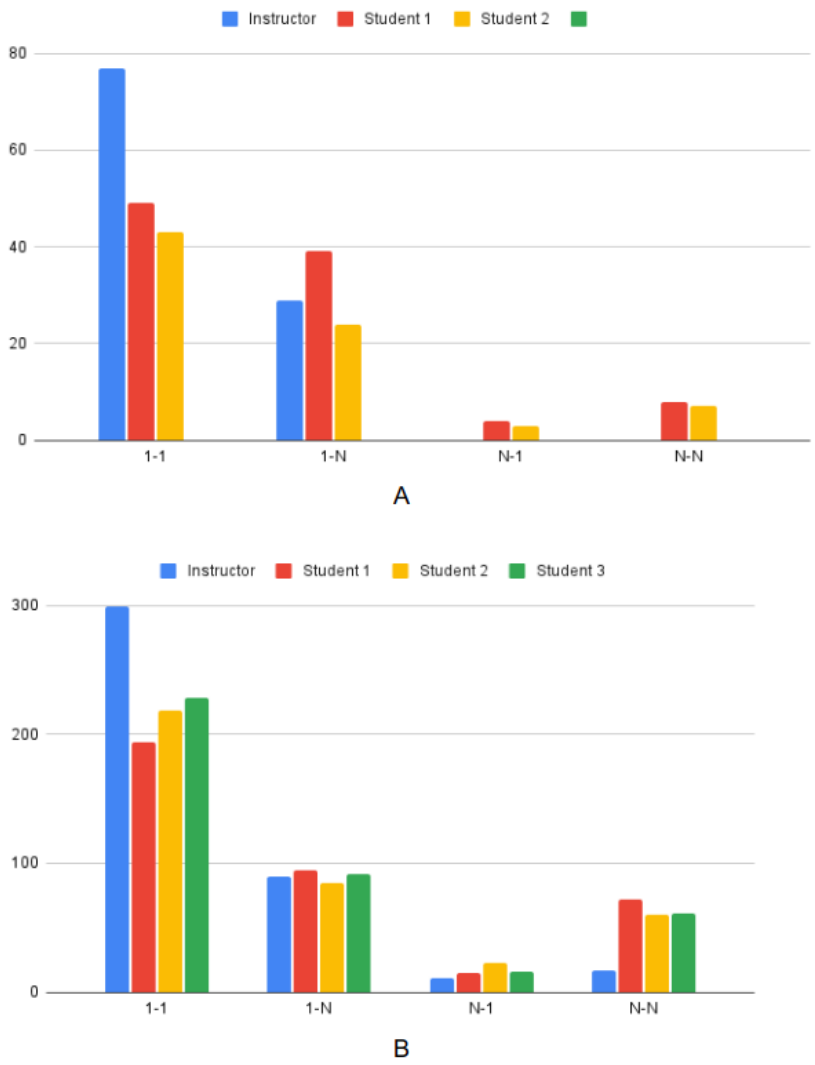


Figure 2.9: Experiment 2, Case Study 1: instructors vs students. A) A comparison of translation pairs categories of Users Group 1. B) A comparison of translation pairs categories of Users Group 2.

- *Group 1*: two expert users, native speakers of Persian, aligned a selection from the *Masnavi* poems (in Persian) against a translation in English. The first alignment was performed without guidelines; the second alignment was performed with a given set of guidelines.

- *Group 2*: two expert users, native speakers of Italian, both educated in Italy with the tradition of Classical language teaching, aligned the same portion of Cicero's *First Speech Against Catilina* against an Italian translation without guidelines.

User Group	Text		User 1	User 2	Overlap	Agreement	J–S Distance
Group 1	Masnavi	No Guidelines	222	190	98	47%	0.2620
		With Guidelines	218	213	157	73% (+26%)	0.0894
Group 3	Odyssey	No Guidelines	295	270	195	69%	0.0374
		With Guidelines	284	277	208	74% (+5%)	0.0010
	Herodotus	No Guidelines	124	141	78	58%	0.2363
		With Guidelines	126	122	96	77% (+19%)	0.0374

Table 2.6: Experiment 2, Case Study 2, Users Group 3. A comparison of established translation pairs by the two users and their overlap and agreement.

- *Group 3*: two users with two different linguistic backgrounds (Italian and Persian respectively), with English as second language and proficiency in Ancient Greek, aligned a passage from the *Odyssey* (1.1-50) and a passage from *Herodotus*'s Histories (1.1) against two popular English translation. The first alignment was performed autonomously, without guidelines; the second alignment was performed with a given set of guidelines.

We manually inspected the intersection data between the users to understand if, and to what extent, providing guidelines would have provided for more homogeneity in the results. Our inspection revealed that even remarkably similar users tend to have important **disagreements** in their alignment choices, even in very short texts. While some of this disagreement is attributable to the lack of guidelines, there is evidence that guidelines can not address every possible issue, and there is always a certain degree of editorial choice, regardless of how affine the users are. We can summarize our observations in the following two points:

- Translation alignment guidelines must appropriately consider the structural linguistic specificities of source and target languages, in order to predict possible patterns of discrepancy. However, the more different the languages, the less likely it is to achieve perfect analytical correspondence, precisely because of these structural factors.

- Alignment Guidelines can not address all ambiguities. There is a high degree of subjectivity in the choices of annotators, which can only be partly adjusted through external control mechanisms such as guidelines. In many cases, the complexity of the texts compared, coupled with external factors, create important inconsistencies in the output.

- Following Alignment guidelines resulted in a higher Inter-Annotator Agreement on all compared texts and reduced the Jensen–Shannon Distance between the

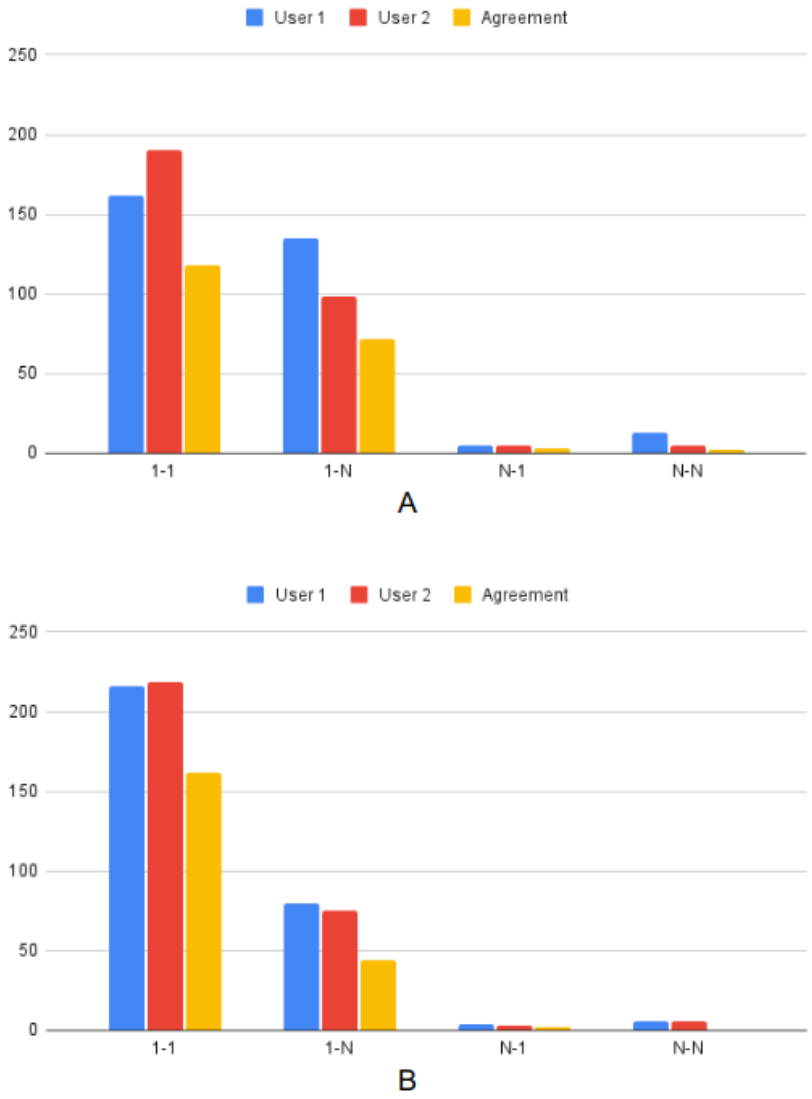


Figure 2.10: Experiment 2, Case Study 2, Users Group 3, Alignments of Hetrodotus's *Histories* (1.1). A) without guidelines B) with guidelines.

alignment categories distribution (1-1, 1-N, N-1, and N-N) (Table 2.6). We can also see that clearly in Figures 2.10 and 2.11.

Even annotators that are virtually identical in terms of experience and background, are not able to achieve perfect consistency with each other. In fact, this is a condition that is almost unattainable in any crowd-sourcing framework: typically, there will be external factors, such as the different linguistic proficiency, background, personal understanding of the context, and even correct understanding of the guidelines provided,

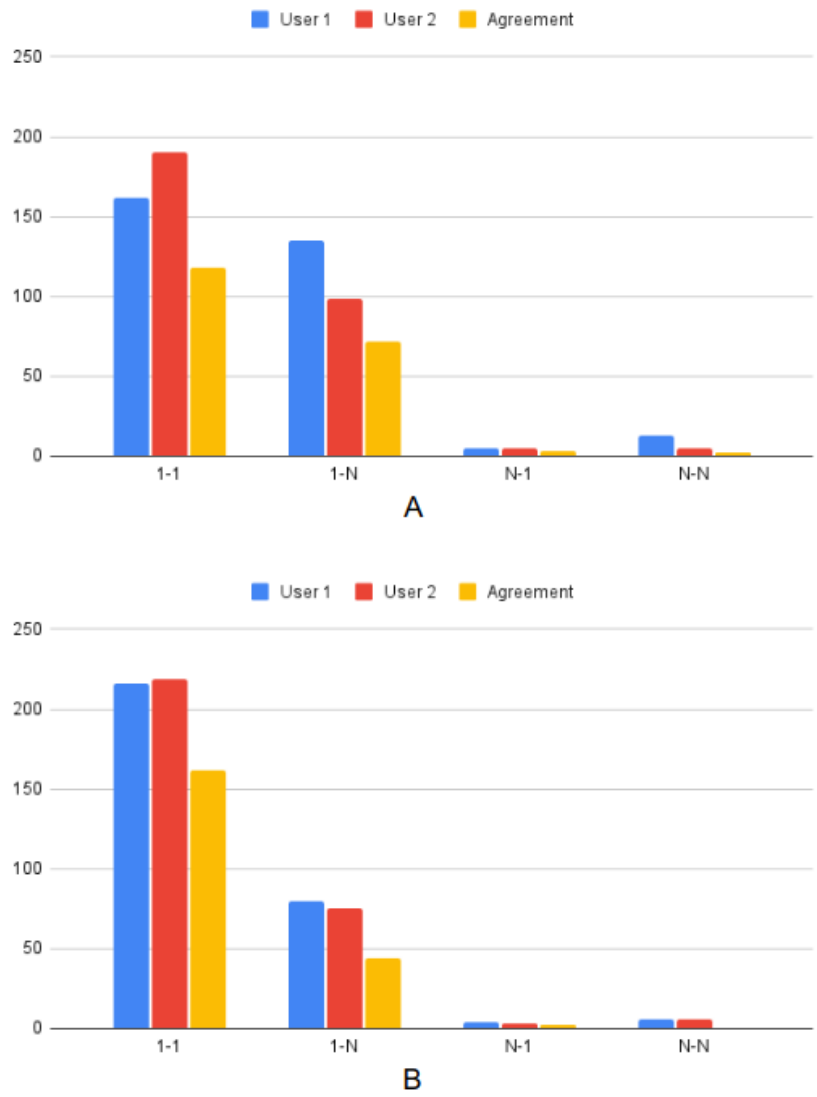


Figure 2.11: Experiment 2, Case Study 2, Users Group 3, Alignments of *Odyssey* (1.1-50). A) without guidelines B) with guidelines.

that cannot be completely controlled.

2.5.5.3 Experiment 3: Evaluating Translations of Ancient Greek Literary Texts

We proposed a workflow to examine how the alignment of parallel texts at word level can support a comparative analysis and the individuation of certain translation phenomena. We analytically compared translations of Ancient Greek texts into English and Persian

through the analysis of parallel corpora aligned manually at word level using UGARIT aiming to observe how close the translations were to each other and to the original text.

We selected samples from Greek tragedy (Euripides' Hippolytus) and Homeric epic (Iliad). The texts were aligned against four competing translations in English (Euripides)[1] and three competing translations in Persian (Iliad), a direct translation[2] and two indirect translations derived from the French translation[3].

The same annotator completed each category of alignments (Ancient Greek vs. English and Ancient Greek vs. Persian), ensuring that a consistent strategy was adopted in the alignment of translation pairs. Each annotator followed a set of guidelines designed for that particular language pair to provide more homogeneous alignments and reduce the chance of mistakes to a minimum.

We extracted all translation pairs and analyzed the following variables:

- **The ratio of non-aligned words in both languages**: the lack of alignment indicates that the user did not find an acceptable correspondence for an individual word or group of words the other text. When the number of non-aligned tokens in the original is significantly lower than the translation, it suggests that something in the original was omitted or overlooked. When the translation includes significantly more non-aligned tokens than the original, it indicates that there is a tendency to expansion and paraphrasing in the translation (Figure 2.12A).

- **Word intersections**: to investigate the semantic overlap across the translations being compared.

- **Translation pair ratios**: measuring how many times one word in the original matched against one word in the translation (one-to-one) or against more than one word (one-to-many) and vice versa (many-to-one), and how many times groups of words were aligned in both texts (many-to-many) (Figure 2.12B).

We used a combination of these criteria to examine trends across the translations. For Hippolytus, the observations led to a surprising conclusion that a 1910 translation, despite a completely different set of stylistic and linguistic choices, was in fact more literal and adherent to the original than the modern and academic ones. For the Iliad, the application of these criteria supported the isolation of phenomena specific to indirect translations, such as the peculiar rendering of proper names.

[1] E.P. Coleridge (1910),D. Kovacs (1995), G. Theodoridis (2010), and I. Johnston (2016)
[2] F. Shamsian (2020)
[3] S. Nafisi (1958) (171), M.J. Kazzazi (1998) (124)

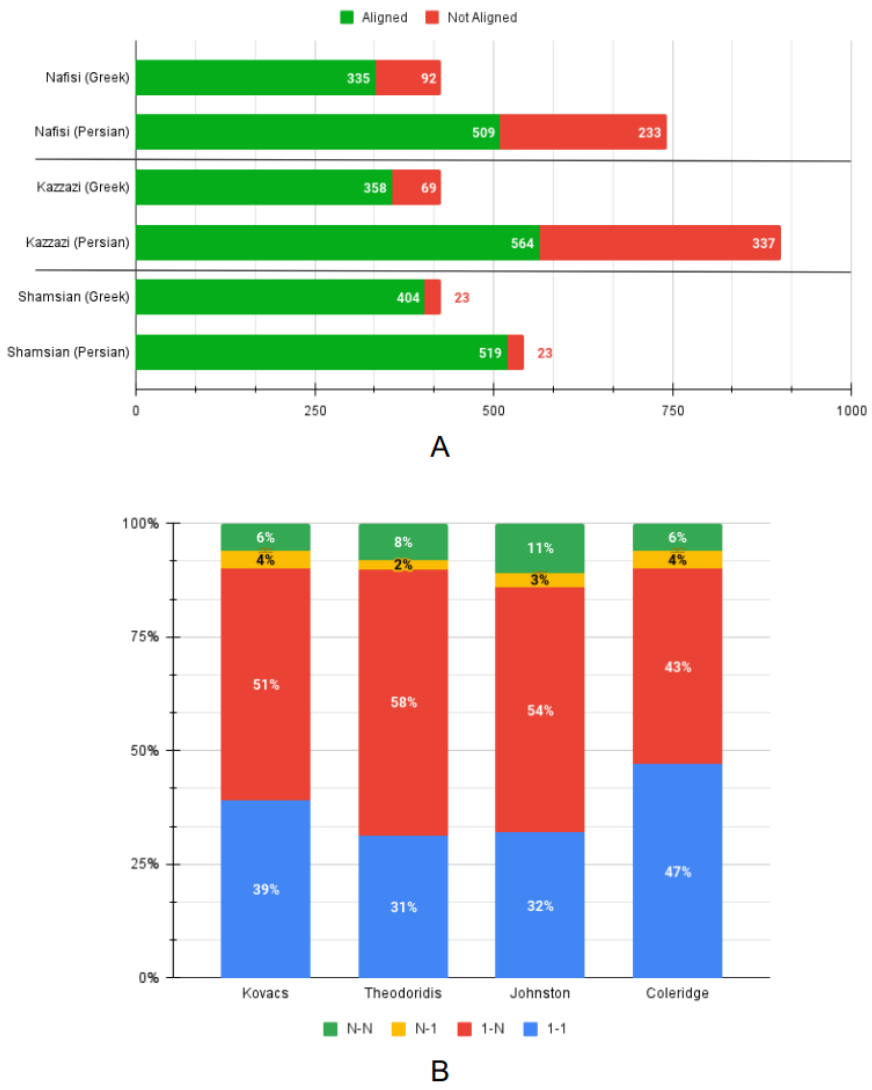


Figure 2.12: Comparison criteria of Experiment 3. A) Ratio of aligned and non-aligned tokens across all Persian translations of the Iliad 1-67. B) Translation Pair ratios across the four translations of Hippolytus, vv. 616-638.

This experiment is part of a more significant effort to upscale the functionalities of Ugarit, and it is conducted in parallel with the development of alignment guidelines for various types of projects, mainly with Ancient Greek as a source language. Alignments, guidelines and data used in this experiment are available on Github[1]. Further, the description of the employed methodology and an extensive analysis is presented in our

[1] `https://github.com/UgaritAlignment/JCLS22-Paper`

paper (192) published in the Journal of Computational Literary Studies. Our future work is oriented towards considerably expanding the dataset of parallel corpora to apply this methodology to a larger group of texts.

2.6 Future Work

When we created UGARIT, we aimed to collect training data for automatic alignment systems and visualize the automatically aligned texts. Still, UGARIT has also been used in many research projects and for different purposes. Experts, who have been using UGARIT regularly, have provided us with some ideas and improvement suggestions to enhance its usability. Therefore, new features and functionality should be implemented and developed to keep pace with users' needs. We can sum them up as follows:

- *Enhance the user interface:* the current user interface was developed in 2017 using Php, plain JavaScript, and Bootstrap CSS. We are going to create a new version employing modernist frameworks and libraries.

- *User roles:* the next version of UGARIT will offer different user roles such as *expert, instructor, student*, which would help create accurate training data by considering the alignments created by experts and instructors, since they are supposed to produce correct and precise alignments. In contrast, students in the learning phase could make some alignment mistakes, and these mistakes should not affect the accuracy of the dynamic lexicon and training data sets.

- *Teaching:* Further, experts and instructors will be able to create groups, add students to the groups, and create assignments. Instructors can upload these assignments in the form of plain parallel texts; students will be asked to align them with deadlines, with the possibility of uploading the correct alignment to allow the system to evaluate the assignments automatically and give notes to every student.

- *Automatic alignment:* we will integrate our automatic alignment model and offer the users to align their texts automatically and correct the obtained alignments manually.

- *Collaborative alignment:* in the current version, users can only align their texts; however, the next version will provide an option for the collaborative alignment of long texts where multiple users can work on the same text.

Chapter 3

Automatic Translation Alignment

In the previous chapter, we showed how alignment can be done manually using tailored annotation tools and how accurate alignments can be employed for various analyses and experiments. On the other hand, manual alignment is expensive in terms of time and resources and requires prior knowledge of both languages. Therefore, automatic alignment models are indispensable. The work on automatic alignment began 30 years ago in the context of developing statistical machine translation models, which was the main driving force behind the development and evolution of automatic alignment models. Later, word-level alignment was essential in various NLP tasks, mainly cross-lingual transfer and annotation projection. In this chapter, we review automatic alignment models and the current state-of-the-art and discuss the challenges. Then, we present our work on training a model for aligning historical languages in general and Ancient Greek and Latin in particular, benefiting from the accurate manual alignments available on Ugarit. This chapter is based on the following publications:

- **Yousef T**, Palladino C, Shamsian F, Ferreira AD, dos Reis MF. *An automatic model and Gold Standard for translation alignment of Ancient Greek.* In Proceedings of the Thirteenth *Language Resources and Evaluation Conference* (LREC) 2022.

- **Yousef T**, Palladino C, Wright DJ, Berti M. *Automatic Translation Alignment for Ancient Greek and Latin.* In Proceedings of the Second Workshop on *Language Technologies for Historical and Ancient Languages* (LREC) 2022.

3.1 Overview of Automatic Alignment Models

Figure 3.1 illustrates the evolution of translation alignment models over the last 33 years. Automatic alignment models can be categorized into three main categories:

3.1.1 Statistical Models

The work on automatic word alignment started in the early nineties (41, 86, 87, 123). Brown et al.(1993) introduced five statistical lexical alignment models, known as IBM models, to build statistical machine translation systems. The models have been widely used and various modifications and improvements have been introduced to enhance their robustness and quality (64). Vogel et al.(1996) proposed using Hidden Markov Model (HMM) to model the word alignment problem by making the alignment probabilities dependent on the relative alignment positions rather than on the absolute positions.

In the beginning of 2000s, Och and Ney(2003) introduced Giza++, a log-linear combination of the best models of Brown et al. (42) Vogel et al. (248), which has been considered the state-of-the-art in the field for a long time and still used as a baseline. Callison-Burch et al.(2004) achieved significant improvements in the alignment and translation quality by incorporating word-aligned texts in training, which were more effective than bilingual dictionaries. Moore(2005) proposed a discriminative alignment method produced comparable or better performance compared to IBM models.

Vogel(2005) developed a phrase alignment model to extract phrase pairs from bilingual corpora. Blunsom and Cohn(2006) proposed a discriminative model that uses supervised training set to induce alignments from parallel corpus using Conditional Random Field (CRF) models.
Liang et al.(2006) (Berkeley Aligner) proposed an unsupervised training approach by jointly training two simple HMM models, and outperformed Giza++ on English-French parallel corpus. Gao and Vogel(2008) provided a multi-threading implementation of Giza++ (MGiza++) to accelerate the training process. Further, Junczys-Dowmunt and Szał(2012) proposed SyMGiza++, a symmetrized version of MGiza++. Lardilleux et al. (140) proposed *Anymalign*, an effecient associative sub-sentential alignment model that makes use of rare words and can align any number of languages simultaneously.

Later, Dyer et al.(2013) introduced a fast and effective log-linear implementation of IBM Model 2, called $fast_align$, outperforming IBM Model 4. Then, Östling and Tiedemann(2016) proposed EfLomAl, an efficient and accurate word alignment model using a Bayesian model with Markov Chain Monte Carlo (MCMC) inference.
To this day, Giza++, fast_align (77), and EfLomAl (182) are still used as baseline models in almost all literature about automatic translation alignment.

Data Sparseness Problems

The performance of statistical alignment models relies on the presence and size of training datasets containing parallel sentences. In the absence of such parallel sentences,

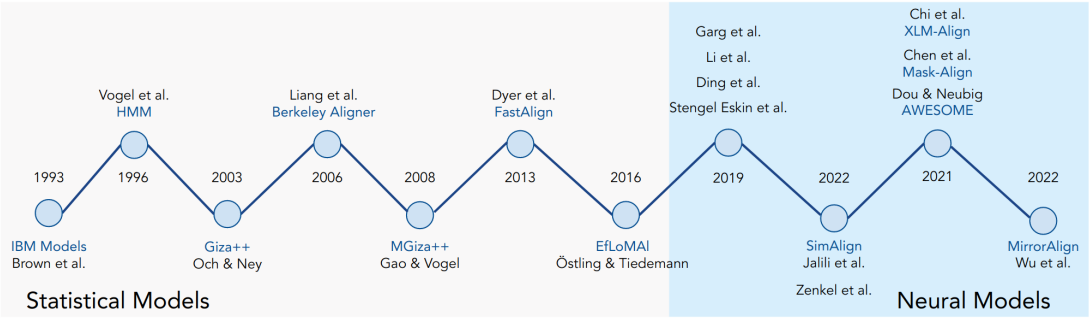


Figure 3.1: Overview of automatic alignment models.

these models exhibit poor performance. Statistical alignment models depend mainly on word frequency and co-occurrences to generate their alignment probabilities. Therefore, words with high frequency have more chance to be correctly aligned, whereas less frequent words, which appear once or twice in the training corpus, are more likely to be misaligned or not aligned at all. Rare words can also be aligned to too many tokens in the parallel sentence, this phenomena is called the *"garbage collectors effect"* (167). The Sparseness problem appears clearly in highly inflected languages, such as Ancient Greek and Latin. Therefore, various syntactic and morphological normalizations must be applied in order to improve the alignment performance and reduce data sparseness (63, 109, 126, 144, 157)

Training. Unsupervised training means the model is trained on parallel sentences with no information about word-level alignments. In contrast, supervised training means the model is trained on parallel sentences aligned at the word level. Supervised training is expensive in terms of time and resources since it requires manual alignments. But it enhances the performance of the alignment model significantly (47, 172, 230).

3.1.2 Neural Models

Translation alignment has benefited from the advances in neural machine translation (NMT) and the rise of transformers models. Neural alignment models have achieved significantly better performance than statistical models, even with no or small training datasets. Different workflows have been proposed recently; we can distinguish between two main paradigms:

Embedding-based Models Models in this category leverage multi-lingual contextualized embeddings to compute cross-lingual semantic similarity among tokens and

employ normalization heuristics to extract the alignments from the similarity matrix. Various training objectives have been proposed to fine-tune the contextualized embeddings to produce better alignments (14, 53, 74, 172). SimAlign (114), Aweasome (74), xlm-Align (53), and MirrorAlign (259) are the state-of-the-art in this category and have achieved great performance on several language pairs. Since the workflow uses pre-trained embeddings, these models can produce good alignments without any training data.

Attentional NMT Models Neural machine translation models employ a sequence to sequence approach that implements encoder-decoder architecture (234). Various attention mechanisms have been proposed to enhance their performance (23, 156). However, attention has been used to induce alignments at word level. Bahdanau et al. (23) extended the basic encoder–decoder architecture and proposed a model that learns to align and translate jointly. Similarly, Garg et al. (92) proposed a model that employs multi-task learning objective for training NMT model and extract the alignment from the attention weights. Various researchers employed this paradigm and proposed a variety of attention variants (71, 148, 277). Stengel-Eskin et al. (230) introduced a discriminative neural alignment model that leverages similarity matrices of encoder-decoder representations to predict word alignments. Zenkel et al. (278) added an alignment layer to the pretrained NMT and the obtained results outperformed Giza++ on different benchmark datasets. Chen et al. (51) proposed Mask-Align a self-supervised neural model with two attention variants.

3.1.3 Hybrid Models

These models combine statistical and neural approaches aiming for a better performance by using the output of statistical models as supervision to train neural models (11), or employing various statistical and neural models and taking their majority vote (228). Marchisio et al. (160) introduced the embedding-enhanced Giza++ that outperformed the traditional Giza++ on all evaluation datasets. Garg et al. (92) incorporated alignments from Giza++ in his multi-task training objective and achieved better alignments than Giza++.

3.2 Automatic Translation Alignment Applied to Historical Languages

All proposed alignment models are designed and optimized to align texts in modern languages, primarily European languages. In this thesis, we leverage the recent advances in neural alignment models and apply them to historical texts and languages. We employ a state-of-the-art workflow that utilizes pre-trained contextualized language models to generate word alignments based on cross-lingual semantic similarity. Further, we fine-tune a language model that can align Ancient Greek, Latin, and English with a training strategy that combines unsupervised training over monolingual and bilingual datasets with supervised training over accurate alignments provided by UGARIT. The trained model significantly outperformed all statistical models such as Giza++, Elfomal, and fast_align on Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Portuguese parallel texts even with the absence of any training data for Ancient Greek-Portuguese.

3.2.1 Word Embedding

Word embedding is a numerical representation of a word. Embedding encodes the meaning of the word as a vector of numeric values; therefore, words with similar meanings have closer representations in the embedding space. There are several models to train word embeddings using a text corpus, such as Word2Vec (166), Fasttext (38), and GloVe[1] (197). These models take a text corpus as input and produce an embedding space. However, these models provide static embeddings, i,e. they encode each word in the language as a single embedding vector regardless of its context. Several approaches have been proposed to map the monolingual embeddings into a shared multilingual space in order to create bilingual or multilingual embeddings. Supervised approaches employ seed bilingual dictionaries to bootstrap the mapping process(72, 142, 262), and unsupervised approaches rely on adversarial training (17, 60, 164, 279).

Recently, with the advances in language modelling and the advent of transformer models (245), such as ELMo[2] (200), BERT[3] (68) and RoBERTa[4] (152), it was possible to generate contextualized embeddings, which provide varying word embeddings according to the context of the sentence. Moreover, contextualized embeddings outperformed

[1] Global Vectors
[2] Embeddings from Language Models
[3] Bidirectional Encoder Representations from Transformers
[4] Robustly Optimized BERT Pre-training Approach

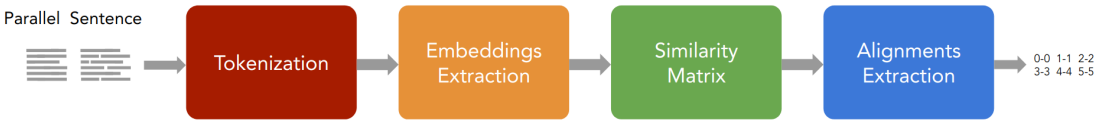


Figure 3.2: The alignment workflow.

static embeddings on a wide range of NLP task (114, 173, 249). Pretrained transformers-based multilingual models, such as multilingual Bert (mBERT) (68) and multilingual RoBERTa (XLM-RoBERTa) (61) achieved significant performance improvements for numerous cross-lingual tasks. Both models are trained unsupervised with the Masked Language Modeling (MLM) training objective, mBERT is trained on Wikipedia texts from 104 languages, whereas XLM-RoBERTa is trained on CommonCrawl data (253) from 100 languages.

Conneau and Lample (59) proposed a Translation Language Modeling (TLM) training objective, a supervised method that leverages parallel data to improve cross-lingual pretraining of multilingual language models. Conneau and Lample (59) found that transformer-based cross-lingual trained unsupervised with MLM outperforms other unsupervised cross-lingual embeddings (60, 138) when computing the cross-lingual word similarity. Chi et al. (53) proposed Denoising Word Alignment, a pretraining task to enhance cross-lingual transferability in multilingual language models. Alqahtani et al. (14) proposed using Optimal Transport (OP) as alignment objective to fine-tune multilingual contextualized embeddings. Nagata et al. (172) modeled the alignment problem as a question-answering problem, by mapping the token in the source sentence (question) to a span (answer) in the target sentence, and they fine-tuned mBERT with manual word-level alignments.

3.2.2 Alignment Workflow

Word alignment is the process of finding word-level translation equivalents between the source sentence $S = (s_1, s_2, .., s_n)$ and its translation $T = (t_1, t_2, .., t_m)$ (42). The alignment process takes S and T as inputs, and generate as output a set of translation pairs $A = \{(s_i, t_j) : s_i \in S, t_j \in T\}$ where s_i is a translation equivalent of t_j.

Figure 3.2 illustrates the core concept of recent studies (74, 114, 230), which leverage the pre-trained multilingual contextualized language models such as mBERT (68) and XLM-RoBERTa (61) or fine-tuned versions of them. The similarity matrix can be derived based on distance/similarity metrics that calculate the similarity for every two tokens based on their embeddings. Then, the word-level alignments can be predicted by employing an extraction algorithm over the similarity matrix.

Similarity Matrix

-	The	son	of	Let	##o	and	Zeus
Λ	252.83	279.58	335.76	408.33	367.15	256.09	213.56
##η	255.65	282.71	335.34	390.42	356.16	249.17	226.16
##το	238.15	265.66	344.19	394.22	404.42	244.34	223.14
##ῦ	239.53	260.33	346.32	366.3	397.53	245.61	206.13
##ς	257.24	278.64	343.31	355.94	408.2	255.69	216.0
καὶ	240.56	230.0	200.73	192.18	213.52	393.37	185.44
Δ	154.59	168.28	270.35	272.57	247.79	223.47	477.98
##ι	167.47	167.86	286.2	271.2	286.11	234.91	445.36
##ὸ	181.58	194.33	284.9	271.77	300.86	250.05	436.31
##ς	176.77	181.93	303.07	276.98	313.56	250.8	419.29
υ	348.62	430.52	186.0	170.84	166.64	197.67	235.27
##ὶ	354.82	437.62	179.53	172.94	151.7	172.85	206.11
##ός	369.75	438.88	201.11	181.84	179.07	179.78	222.28

(a) Similarity Matrix

Forward

-	The	son	of	Let	##o	and	Zeus
Λ	0.0	0.0	0.0	1.0	0.0	0.0	0.0
##η	0.0	0.0	0.0	1.0	0.0	0.0	0.0
##το	0.0	0.0	0.0	0.0	1.0	0.0	0.0
##ῦ	0.0	0.0	0.0	0.0	1.0	0.0	0.0
##ς	0.0	0.0	0.0	0.0	1.0	0.0	0.0
καὶ	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Δ	0.0	0.0	0.0	0.0	0.0	0.0	1.0
##ι	0.0	0.0	0.0	0.0	0.0	0.0	1.0
##ὸ	0.0	0.0	0.0	0.0	0.0	0.0	1.0
##ς	0.0	0.0	0.0	0.0	0.0	0.0	1.0
υ	0.0	1.0	0.0	0.0	0.0	0.0	0.0
##ὶ	0.0	1.0	0.0	0.0	0.0	0.0	0.0
##ός	0.0	1.0	0.0	0.0	0.0	0.0	0.0

Backward

-	The	son	of	Let	##o	and	Zeus
Λ	0.0	0.0	0.0	1.0	0.0	0.0	0.0
##η	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##το	0.0	0.0	0.1	0.0	0.02	0.0	0.0
##ῦ	0.0	0.0	0.86	0.0	0.0	0.0	0.0
##ς	0.0	0.0	0.04	0.0	0.98	0.0	0.0
καὶ	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Δ	0.0	0.0	0.0	0.0	0.0	0.0	1.0
##ι	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ὸ	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0
υ	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ὶ	0.0	0.22	0.0	0.0	0.0	0.0	0.0
##ός	1.0	0.78	0.0	0.0	0.0	0.0	0.0

Intersection

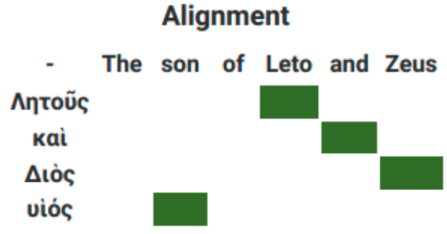
-	The	son	of	Let	##o	and	Zeus
Λ	0.0	0.0	0.0	1.0	0.0	0.0	0.0
##η	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##το	0.0	0.0	0.0	0.0	1.0	0.0	0.0
##ῦ	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ς	0.0	0.0	0.0	0.0	1.0	0.0	0.0
καὶ	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Δ	0.0	0.0	0.0	0.0	0.0	0.0	1.0
##ι	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ὸ	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0
υ	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ὶ	0.0	1.0	0.0	0.0	0.0	0.0	0.0
##ός	0.0	1.0	0.0	0.0	0.0	0.0	0.0

(b) Alignment Extraction (Softmax)

Intersection

-	The	son	of	Let	##o	and	Zeus
Λ	0.0	0.0	0.0	1.0	0.0	0.0	0.0
##η	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##το	0.0	0.0	0.0	0.0	1.0	0.0	0.0
##ῦ	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ς	0.0	0.0	0.0	0.0	1.0	0.0	0.0
καὶ	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Δ	0.0	0.0	0.0	0.0	0.0	0.0	1.0
##ι	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ὸ	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0
υ	0.0	0.0	0.0	0.0	0.0	0.0	0.0
##ὶ	0.0	1.0	0.0	0.0	0.0	0.0	0.0
##ός	0.0	1.0	0.0	0.0	0.0	0.0	0.0

(c) Alignment Result



(d) Alignment Result

Figure 3.3: The alignment process and an example illustrates its workflow (mBERT).

Similarity Matrix

-	_The	_son	_of	_Le	to	_and	_Ze	us
_Λ	203.42	196.18	203.6	300.45	260.95	113.87	118.47	103.44
η	217.12	189.28	202.26	309.75	289.45	117.93	107.32	96.33
το	199.43	198.26	204.66	278.22	316.79	133.37	106.09	106.83
ῦ	208.72	212.32	237.3	272.84	299.03	150.06	120.55	132.07
ς	221.05	214.08	212.2	250.81	279.8	137.81	112.96	128.24
_καὶ	114.75	128.73	115.04	124.37	121.93	309.73	132.18	139.06
_Δι	93.57	123.67	152.19	157.72	152.62	184.61	322.99	304.64
ὸ	97.46	124.1	172.57	145.83	158.94	181.67	306.42	316.77
ς	98.68	117.08	166.69	134.47	150.66	172.81	295.54	319.58
_	209.92	249.06	96.56	85.4	83.15	134.22	136.64	126.19
υ	193.34	257.83	87.04	71.53	70.34	117.35	135.1	123.1
ἰ	192.84	260.3	89.7	78.09	74.31	124.16	143.55	124.48
ός	204.28	249.47	96.56	81.02	89.68	137.43	145.6	147.82

(a) Similarity Matrix

Forward

-	_The	_son	_of	_Le	to	_and	_Ze	us
_Λ	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
η	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
το	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
ῦ	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
ς	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
_καὶ	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
_Δι	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
ὸ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
_	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
υ	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
ἰ	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
ός	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

Backward

-	_The	_son	_of	_Le	to	_and	_Ze	us
_Λ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
η	0.02	0.0	0.0	1.0	0.0	0.0	0.0	0.0
το	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
ῦ	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
ς	0.98	0.0	0.0	0.0	0.0	0.0	0.0	0.0
_καὶ	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
_Δι	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
ὸ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.06
ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.94
_	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
υ	0.0	0.08	0.0	0.0	0.0	0.0	0.0	0.0
ἰ	0.0	0.92	0.0	0.0	0.0	0.0	0.0	0.0
ός	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Intersection

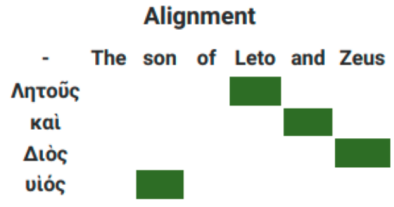
-	_The	_son	_of	_Le	to	_and	_Ze	us
_Λ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
η	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
το	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
ῦ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
_καὶ	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
_Δι	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
ὸ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
_	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
υ	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
ἰ	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
ός	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

(b) Alignment Extraction (Softmax)

Intersection

-	_The	_son	_of	_Le	to	_and	_Ze	us
_Λ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
η	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
το	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
ῦ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
_καὶ	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
_Δι	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
ὸ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
ς	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
_	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
υ	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
ἰ	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
ός	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

(c) Alignment Result



(d) Alignment Result

Figure 3.4: The alignment process and an example illustrates its workflow (XLM-RoBERTa).

3.2.2.1 Tokenization & Embeddings Extraction

Tokenization aims to split the sentences into single tokens in order to find their translation correspondences. Traditional tokenizers use white spaces or punctuations as token boundaries to break the text. However, contextualized language models use different approaches. For instance, mBert uses a *WordPiece* tokenizer, which splits words either into their complete forms or into frequent word pieces (subwords) when that is possible (e.g. multitasking would be divided to multi, ##task, ##ing) (Figure 3.3a). XLM-RoBERTa uses a *SentencePiece BPE*(135), a subword tokenizer similar to *WordPiece*, but converts white spaces to "_" and considers the whole sentence as one large token (Figure 3.4a). After tokenizing the sentence, embedding vectors can be extracted for each token from the language model. Both mBERT and XLM-RoBERTa (Base) generate a 768-length embedding vector for each token. For similarity matrices generation, we tested two embeddings alternatives: the *subwords-level embeddings* as provided by the language models and the *word-level embeddings* where each word is represented by the average of all its subwords embedding vectors. In all experiments, we extracted embedding vectors from the 8-th layer of mBERT and XLM-RoBERTa since they have achieved the best performance.

3.2.2.2 Similarity Matrix

Suppose S_{src}, S_{tgt} be two parallel sentences with lengths n, m. and $SIM_{n\times m}$ the similarity matrix of these two sentences. Using the embedding vectors derived from multilingual transformers models, the similarity matrix can be filled as in equation 3.1:

$$\sum_{i}^{n}\sum_{j}^{m} SIM(i,j) = F_{sim}(t^{i}_{src}, t^{j}_{tgt}) \tag{3.1}$$

Where t^{i}_{src} is the embedding vector of the ith token in S_{src}, t^{j}_{tgt} is the embedding vector of the jth token in S_{tgt}, and F_{sim} is the cross-lingual word similarity, which can be computed using a similarity function between the two vectors such as *Cosine Similarity*, *Dot Product*, *Euclidean distance*, or other functions (53). However, Our experiments revealed that *Dot Product* slightly outperforms *Cosine Similarity*. Figure 3.3a shows an example of a similarity matrix computed using the *Dot Product* over the (sub)word embeddings extracted from the 8-th layer of mBERT.

3.2.2.3 Alignments Extraction

Once the similarity matrix is computed, alignments can be extracted using a heuristic to find best matching. The alignment task can also be seen as a classification problem

where every source token s_i must be classified (aligned) into a set of classes (tokens in the target sentence $t_1, t_2, ..t_n$) and vice versa. Therefore, common mathematical functions that have been used in machine and deep learning can be used for this task as well. Chi et al. (53) modeled the alignment problem as optimal transport in order to induce the alignments from the similarity matrix. Dou and Neubig suggested two probability thresholding-based methods to extract alignments from the similarity matrix, namely, $Softmax$ and $Entmax$ (199). Chi et al. (53) employed *Optimal Transport* (OT) heuristic. Further, Jalili Sabet et al. (114) suggested three methods: $Argmax$, $Itermax$, and $Match$, a graph-based method.

Argmax (arguments of the maxima) is a mathematical function that returns the argument(s) that returns the maximum value from the target function. It is usually used in single-class classification tasks. In our scenario, it returns the index corresponding to the highest similarity in a given vector.

Itermax applies *Argmax* iteratively[1]. Unlike *Argmax*, *Itermax* allows that one token to be aligned to multiple tokens. Our qualitative analysis confirms this observation and we mention it in chapter 4 when discussing *phrase alignment accuracy*.

Softmax is a mathematical function that converts a vector of numbers into a probability distribution; it is a smooth approximation to the *Argmax* function. *Softmax* is very common in the field of Machine and Deep Learning and unlike *Argmax* it is used in various multi-class classification tasks.

αentmax is a differentiable generalization of *Softmax* (62) proposed by Peters et al. to solve *Softmax*'s limitations, but is unfortunately slower than *Softmax* (238).
Dou and Neubig (74) utilize a threshold to filter out the small probabilities and replace them with zeros when when applying *Softmax* and *Entmax*.

Further, these functions are applied in two directions (forward and backword) and the intersection is calculated as final result (Figures 3.3b and 3.4b). That means the function will be applied to each row and column in the similarity matrix, classifying/aligning each token in the source sentence to the possible tokens in the target, and then classifying/aligning each token in the target to.

[1]The full description of the algorithm is available in (114).

Match is a graph-based approach the considers the similarity matrix as a bipartite graph, and the tokens of the two parallel sentences are disjoint sets of nodes connected through weighted edges, where the weight is the semantic similarity between the two connected tokens. The Match algorithm tries to find the maximum matching in the graph (89, 114).

OT Chi et al. (53) used this approach to extract translation pairs from similarity matrix. It formulates the alignment problem as Optimal Transport problem and uses Sinkhorn algorithm for this purpose (202) .

Figures 3.3a and 3.4a show how the alignment can be computed on the subword level. However, the target is to perform word-level alignment, therefore, we employed the heuristic principle *"two words are aligned if any of their subwords are aligned"* to convert subword-level alignments to word-level alignments similar to Jalili Sabet et al. and Dou and Neubig (Figures 3.3c and 3.4c).

3.2.3 Training Process and Objectives

The experiments we conducted on the pre-trained mBERT and XLM-RoBERTa (Zero-Shot) showed significantly poor performance on Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Portuguese datasets. Therefore, fine-tuning those models was necessary to achieve better performance. Due to the availability of parallel sentences and in order to obtain the best outcome from the training process, we conducted several experiments employing multiple training objectives[1] in order to find the best training strategy for our case study:

- **Masked Language Modeling (MLM):** it randomly masks some of the tokens from the input sentence (e.g. 15% of all word pieces), and the model has to predict the original masked tokens (68).
- **Translation Language Modeling (TLM):** this objective concatenates the source and target sentences and perform MLM on the concatenated string (59).
- **Self-training Objective (SO):** this objective enhances the symmetricity between forward and backward alignments and encourages words aligned in the first alignment pass to have closer embeddings.

[1] A detailed description of the training objectives is available in (53, 59, 68, 74)

- **Parallel Sentence Identification (PSI):** this objective encourages parallel sentences to be more similar than random unaligned sentences, which enables the overall alignments of embeddings to be closer together.

- **Denoising Word Alignment (DWA):** this objective estimates the alignments of the input sentences on-the-fly and tries to predict the alignments from the perturbed version of the input sentence aiming to minimize the cross-entropy between the alignment probabilities and the self-labeled alignments (53).

Table 3.1 provides an overview of the conducted experiments and the employed training objectives and training datasets.

Experiment	Input Models	Epochs	Training Objectives	Languages	Data Size	Source
EX 1	mBERT, XLM-R	1	MLM, SO, TLM, PSI	GRC-ENG GRC-LAT	40.500 sentences	Perseus DL DFHG
EX 2	EX2 fine tuned models	5	MLM	GRC Monolingual	12 Millions Tokens	Perseus DL, First1kGreek, TreeBanking
EX 3	EX1 fine tuned models	5	MLM, SO TLM, PSI	GRC-ENG GRC-LAT GRC-KAT	45.000 sentences	Perseus DL, DFHG, UGARIT
EX 4	EX2 fine tuned models	5	MLM, SO TLM, PSI			
EX 5	EX4 fine tuned models	30	SO	Mixed dataset	5000 sentences 190k TPs	UGARIT

Table 3.1: An overview of the conducted experiments.

Experiment 1

In this experiment, we performed **unsupervised** fine-tuning of the mBERT and XLM-RoBERTa using 40.500 Ancient Greek-English and Ancient Greek-Latin parallel sentences with the training objectives MLM, TLM, SO, PSI. The parallel sentences used in this experiment are taken from Perseus Digital Library[1] (Iliad, Odyssey, Xenophon, New Testament) and the Digital Fragmenta Historicorum Graecorum project (DFHG)[2] (32).

[1]https://github.com/PerseusDL/canonical-greekLit

[2]The DFHG is the digital open version of the five volumes of the first big printed collection of ancient Greek fragmentary historians edited by Karl Müller in the 19th century. https://www.dfhg-project.org.

Experiment 2

Multilingual language models mBERT and XLM-RoBERTa are not trained on Ancient Greek texts but on modern Greek, which is very different. Therefore, we suggested to train them with monolingual Ancient Greek texts to allow models learning Ancient Greek embeddings. For this purpose, we used 12 million Ancient Greek tokens to train the fine-tuned models we obtained in *Experiment 1* with Masked Language Model (MLM) training objective. We collected the training dataset from Perseus Digital Library, the first thousand years of Greek project[1], and the PROIEL, Perseus[2], and Gorman[3] treebanking datasets.

Experiment 3 & 4

Experiment 3 and 4 aim to inspect the impact of training the models on monolingual texts; therefore, the two experiments use the same training data, but they differ by the pre-trained model. *Experiment 3* fine-tunes the model obtained in *Experiment 1*, whereas *Experiment 4* fine-tunes the model obtained in *Experiment 2* on 45.000 parallel texts (The dataset in *Experiments 1* in addition to 4.000 further parallel sentences taken from UGARIT database). For training, we used MLM, TLM, SO, PSI training objectives. The texts are in different languages mainly (Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Georgian).

Experiment 5

In this experiment, we perform supervised training for the fine-tuned model obtained after *Experiment 4* using a word-level manually aligned dataset provided by UGARIT. We selected accurate and complete alignments done by manually selected list of teachers, and Experts. The dataset consists of 5,000 parallel texts and almost 190k translation pairs, and we trained the model for 30 epochs.

Other experiments

We have also conducted various experiments and tested different variants of supervised and unsupervised monolingual and bilingual training to determine which training strategy achieves the best performance. Further, we performed supervised training for the fine-tuned model obtained after *Experiment 5* using dictionary entries from

[1] https://opengreekandlatin.github.io/First1KGreek/
[2] universaldependencies.org
[3] https://vgorman1.github.io/

the Ancient Greek-English MORPHEUS[1] aiming to test if fine-tuning the model with bilingual dictionary entries would enhance the performance. The dataset consists of 43k entries; each entry consists of a single Greek token with meanings in English. However, the results reveal a drop in performance.

3.2.4 Evaluation

Evaluation Metrics

We evaluate the performance of the alignment models against three gold standard datasets using *Precision*, *Recall*, $F1$, and Alignment Error Rate (AER) (180), which can be computed as in equations 3.2.

$$Precision = \frac{|A \cap P|}{|A|} \quad , \quad Recall = \frac{|A \cap S|}{|S|}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.2}$$

$$AER = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|}$$

Where A indicates the alignments set predicted by the model, P and S indicate respectively the *Possible* and *Sure* alignment sets in the gold standards, and $|.|$ denotes the length of the set. A detailed discussion of evaluation datasets, the current evaluation practice, its limitations, and the suggested solutions is presented in Chapter 4.

Baseline Models

We compare the performance of our models to three popular statistical word alignment models, namely, Giza++, fast_align, EfLoMAl. The baseline models require training data in the form of parallel sentences. We trained them on the 35.000 Ancient Greek-English and 8000 Ancient Greek-Latin sentences, and evaluated their performance on our gold standard datasets. The poor performance on the Ancient Greek Portuguese dataset is because of the absence of the training data; the models are trained only on 183 sentences of the gold standard dataset which is insufficient amount to produce good results.

[1]Morpheus is a morphological parsing and lemmatizing tool produced by the Perseus Project.

		Precision		Recall		F1		AER	
Baseline	Giza++	37.25%		29.26%		32.78%		67.01%	
	FastAlign	37.37%		35.64%		36.48%		63.47%	
	EfLoMAl	47.17%		42.93%		44.95%		54.95%	
		mBERT				XLM-RoBERTa			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
Zero-Shot	Softmax	37.14%	21.09%	26.90%	72.70%	37.59%	11.84%	18.01%	81.80%
	Entmax	42.58%	17.34%	24.64%	74.94%	46.74%	8.67%	14.63%	85.20%
Ex1	Softmax	52.98%	38.21%	44.40%	55.28%	54.61%	28.21%	37.20%	62.46%
	Entmax	59.84%	37.04%	45.76%	53.78%	65.14%	23.49%	34.53%	65.07%
Ex2	Softmax	54.03%	39.68%	45.76%	53.94%	53.58%	24.21%	33.35%	66.33%
	Entmax	60.35%	35.29%	44.54%	54.99%	61.99%	18.54%	28.54%	71.11%
Ex3	Softmax	65.06%	48.08%	55.30%	44.33%	65.22%	36.39%	46.72%	52.88%
	Entmax	68.23%	46.01%	54.96%	44.58%	73.42%	30.34%	42.94%	56.57%
Ex4	Softmax	74.71%	54.61%	63.10%	36.50%	77.17%	54.42%	63.83%	35.81%
	Entmax	77.60%	52.22%	62.43%	37.11%	82.84%	48.84%	61.45%	38.11%
	Match	58.09%	65.03%	61.36%	38.81%	62.46%	72.70%	67.19%	33.05%
	Argmax	78.72%	51.39%	62.18%	37.33%	84.73%	46.65%	60.17%	39.36%
	Itermax	69.64%	58.35%	63.50%	36.25%	77.07%	58.28%	66.37%	33.32%
Ex5	Softmax	80.80%	56.91%	66.78%	32.72%	92.62%	66.85%	77.65%	21.88%
	Entmax	83.86%	53.76%	65.52%	33.93%	94.17%	64.35%	76.46%	23.06%
	Match	65.42%	72.76%	68.90%	31.31%	79.22%	**87.26%**	**83.05%**	**17.17%**
	Argmax	84.95%	52.47%	64.87%	34.57%	**94.44%**	63.32%	75.81%	23.70%
	OT	78.83%	47.71%	59.44%	39.57%	91.50%	54.97%	68.68%	30.78%
	Itermax	78.43%	64.08%	70.53%	29.14%	91.05%	71.65%	80.19%	19.42%

Table 3.2: Evaluation results on Ancient Greek-English gold standards.

		Precision		Recall		F1		AER	
Baseline	Giza++	25.59%		24.60%		0.2509		74.88%	
	fast align	25.62%		30.14%		27.70%		72.47%	
	EfLoMAl	34.84%		35.59%		35.21%		64.81%	
		mBERT				XLM-RoBERTa			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
Zero-Shot	Softmax	30.08%	26.66%	28.27%	71.66%	21.68%	13.53%	16.66%	83.16%
	Entmax	33.65%	22.67%	27.09%	72.73%	23.55%	9.27%	13.30%	86.44%
Ex5	Softmax	63.84%	61.27%	62.53%	37.40%	76.11%	75.61%	75.86%	24.13%
	Entmax	65.49%	57.41%	61.18%	38.61%	**77.45%**	72.69%	74.99%	24.89%
	Match	50.00%	72.61%	59.22%	41.50%	58.79%	**86.17%**	69.89%	31.01%
	Argmax	66.01%	54.92%	59.96%	39.76%	77.25%	71.10%	74.05%	25.81%
	OT	62.33%	46.55%	54.11%	44.26%	75.27%	64.52%	69.48%	30.28%
	Itermax	59.67%	64.06%	61.79%	38.35%	72.22%	81.02%	**76.37%**	**23.91%**

Table 3.3: Evaluation results on Ancient Greek-Portuguese gold standards.

		Precision		Recall		F1		AER	
Baseline	Giza++	55.03%		67.61%		60.67%		39.48%	
	fast_align	51.64%		70.51%		59.62%		40.67%	
	EfLoMAl	76.79%		78.12%		77.45%		22.57%	
		mBERT				XLM-RoBERTa			
		Precision	Recall	F1	AER	Precision	Recall	F1	AER
Zero-Shot	Softmax	55.40%	51.52%	53.39%	46.55%	49.35%	42.10%	45.44%	54.49%
	Argmax	80.25%	34.86%	48.61%	50.87%	62.10%	41.88%	50.02%	49.77%
Ex5	Softmax	85.67%	84.64%	85.15%	14.83%	94.64%	92.39%	**93.50%**	**6.47%**
	Entmax	87.39%	82.39%	84.82%	15.12%	95.39%	91.30%	93.30%	6.64%
	Match	62.18%	87.97%	72.86%	27.55%	80.61%	**96.30%**	87.76%	12.50%
	Argmax	88.46%	80.80%	84.46%	15.09%	**95.52%**	91.38%	93.40%	6.55%
	OT	84.44%	71.78%	77.70%	21.64%	94.44%	88.12%	91.17%	8.76%
	Itermax	81.27%	84.78%	82.99%	17.06%	92.21%	93.33%	92.77%	7.25%

Table 3.4: Evaluation results on Ancient Greek-Latin gold standards.

3.2.5 Results

The results reported in Tables 3.2,3.3, and 3.4 are based on subwords embeddings. The results show the superiority of the embedding-based models over the statistical models (Giza++, EfLoMAl, fast_align). Also, The performance of the models derived from fine-tuned XLM-RoBERTa models are superior to those derived from mBERT-fine-tuned models for all gold standard datasets. Further, training the model on monolingual texts in *Experiments 2* decreased the $F1$ and increased the AER on both models.

As mentioned before, in *Experiments 3 and 4*, we used the same training datasets and the same number of epochs. Table 3.2 reveals significant differences in the results, as all the evaluation metrics values of *Experiment 4* are superior to *Experiment 3* values in both models with a large margin, which demonstrates the importance of training the models on monolingual datasets. Moreover, supervised training with word-level aligned sentences in *Experiment 5* enhanced the performance significantly, and the AER decreased by (10%-15%). Table 3.3 shows that the proposed training strategy also achieved good results on Ancient Greek-Portuguese dataset with no supervised or unsupervised training on Ancient Greek-Portuguese parallel texts.

Regarding the alignment extraction approaches, *Argmax* outperformed other models on Ancient Greek-English and Ancient Greek-Latin datasets regarding Precision, and achieved the second highest Precision on Ancient Greek-Portuguese after *Entmax* with a slight difference (0.2%). *Match* always achieved the highest Recall and the lowest Precision. Figure 3.5 confirms this observation and shows that *Match* generates the highest number of translation pairs among the other approaches, which explains why

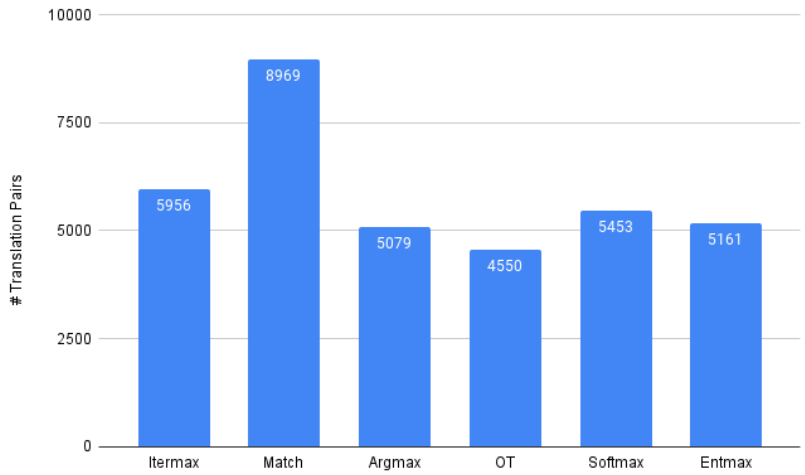


Figure 3.5: A comparison among the alignment extraction approaches regarding the number of translation pairs they produce based on the XLM-R fine-tuned model obtained in Experiment 5.

Match consistently achieves the highest Recall and the lowest Precision as reported in tables. Moreover, OT generates the lowest number of translation pairs. Therefore, it achieves always the lowest Recall and the highest Precision.

The results show that there is no superior model regarding the AER, while $Itermax$ achieved the lowest AER on Ancient Greek-Portuguese and the second lowest AER on Ancient Greek-English, it achieved the fourth best AER on Ancient Greek-Latin with 0.78% more error rate than the best model. $Softmax$ achieved better AER, $F1$, $Recall$ than $Entmax$ in all experiments, while $Entmax$ is always superior regarding $Precision$.

Further, We compared the alignment performance when using subword embeddings and averaged word embeddings, and the experiments showed that subwords-level embeddings always outperform word-level embeddings on different alignment extraction approaches with mBERT and XLM-RoBERTa (Table 3.5).

We also tested the model performance on Ancient Greek-Persian, Ancient Greek-Arabic and Arabic-English and the results were excellent. Unfortunately, we could not measure and quantify the performance due to the absence of alignment gold standard datasets for these language pairs.

3.2.5.1 Qualitative Evaluation

In addition to the quantitative evaluation, we conducted a qualitative evaluation in collaboration with domain experts on 50 Ancient Greek-English sentences, 52 Ancient

	mBERT		XLM-RoBERTa	
Tokenization	Subword	Word	Subword	Word
Softmax	32.72%	34.47%	21.89%	24.07%
Entmax	33.93%	35,81%	23,69%	25,42%
Match	31.31%	37.32%	18.72%	29.06%
Argmax	34.57%	36.29%	24.54%	26.16%
Itermax	29.14%	29.62%	19.73%	19.14%

Table 3.5: An AER comparison between two different embeddings alternatives: the subword embeddings as they are provided by the language models, and the word-level embeddings which is calculated for each word as the average embeddings of its subwords. This comparison is based on the embeddings provided by the Greek-English fine-tuned models obtained in Experiment 5.

Greek-Portuguese sentences, and 50 Ancient Greek-Latin fragments for a detailed observation of the errors and mismatch patterns.

For Greek-English alignments, from 527 translation pairs, there were 30 incorrect pairs (5.6%). The most frequent inaccuracy was in partially correct pairs. They included an unrelated token that should have been excluded, or they were missing a token in the English translation to match correctly. This partially incorrect matching counts for 20 instances out of 30 errors. Of this number, 13 instances included an extra token in the English translation (Example: ἄναχτος: lord strikes), two instances included an extra token in the Greek (example: ἐξεπράθομεν δέδασται : we apportioned), and five instances were missing token(s) in the English translation (Example: ἐΐχτην : were, missing the token “like” in the English translation). The rest of the errors (10 out of 30) were tokens that did not match at all, most commonly, particles and conjunctions (5 instances) that are often difficult to translate or align. The rest of the errors can be attributed to the complexity and ambiguity of the translation or the text.

For Greek-Portuguese alignments, from 509 translation pairs, there were 51 incorrect pairs (10%). As in Greek-English, the most frequent inaccuracy (24 out of 51 errors) was found in pairs that were partially correct (e.g.: ἄλλοι: os outros; ὅσσε: olhos pareciam; οὕτω: tão cedo). Another 11 mismatches were related to missing a word in Portuguese translation (e.g.: γλυκίων: mais; καρτερός: vigoroso). Among the 14 tokens that were not a complete match, the inverted alignments seemed to occur where they were in close proximity either in the original or in the translation (e.g.: φηρσὶν: os monteses, and ὀρεσκῴοισι: centauros). Also, the ambiguity of the Portuguese term ”a”

(preposition and definite article) may have caused another type of error (e.g. πρὸ: a). In general, the errors appeared to be more critical when Portuguese shows a two-word comparative, periphrastic, or multi-word expression translation.

For Greek-Latin alignments, from 748 translation pairs, there were 40 incorrect pairs (5.35%). The model correctly aligned 54 of 54 prepositions (100%), 18 of 18 adverbs (100%), 186 of 188 Named-Entities (98.94%), 53 of 54 adjectives (98.15%), 53 of 54 conjunctions (98.15%), 40 of 41 pronouns (97.56%), 119 of 125 verbs (95.20%) and 125 of 133 substantives (93.98%).

Most recurrent errors are due to the absence of articles in Latin: Greek articles are sometimes incorrectly aligned with contextual Latin adjectives, pronouns, and substantives. Other limits are also due to elliptical constructions, where finding a certain match is more complex. Finally, Greek particles are variously aligned with Latin conjunctions and adverbs.

3.2.6 User Interface

In order to make the alignment model usable and accessible for users with no technical expertise, we implemented an online tool to allow them to align their texts. The tool is available under `http://ugarit-aligner.com`. Users can enter the source text and translation and get the alignment. Users can align up to five parallel sentences at once by putting each sentence in a single line. The tool provides three visualization options: side-by-side view, table view, and grid view. Users can also export the obtained alignments in JSON format. Moreover, the alignment model (fine-tuned XLM-RoBERTa) is also published on Huggingface[1] and can be downloaded and used locally.

3.2.7 Summary

The evaluation results revealed that fine-tuning the multilingual models on monolingual datasets played a key role in enhancing the performance of the model. The results can be further enhanced by performing supervised and unsupervised fine-tuning even on language pairs that are different from the target languages such as Ancient Greek-Italian and Ancient Greek-French. The results also showed the great impact of supervised training with word-level manual alignments which decreased the AER from 33.05% to 17.17% on fine-tuned XML-RoBERTa model and from 36.25% to 29.14% on fine-tuned mBERT model on Ancient Greek-English dataset. This leads us to conclude that fine-tuned XMl-RoBERTa models are more sensitive to supervised fine-tuning than fine-tuned mBERT models.

[1] https://huggingface.co/UGARIT/grc-alignment

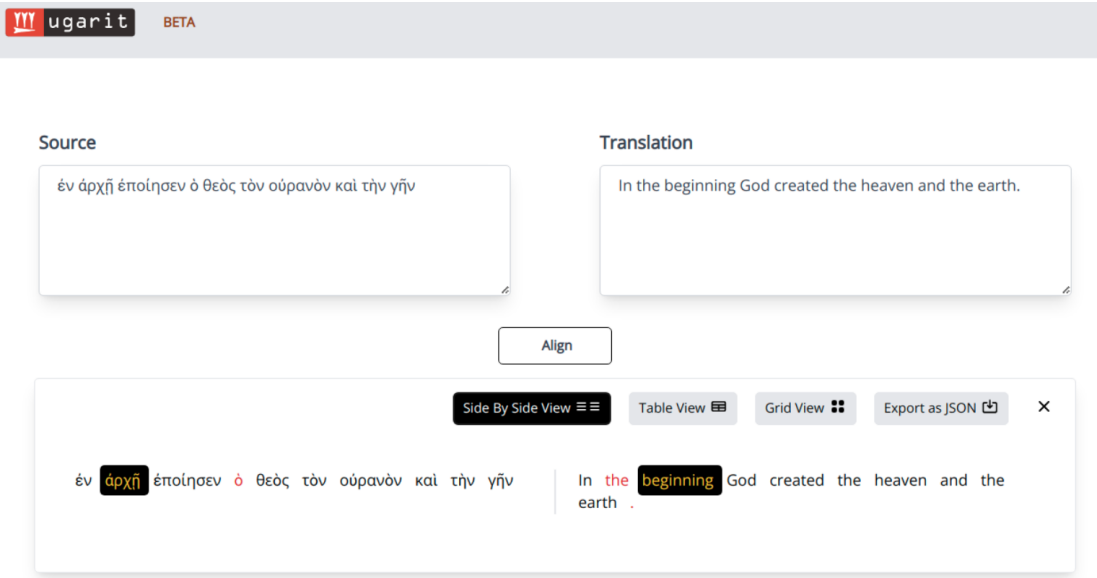


Figure 3.6: Ugarit automatic alignment tool.

Training Strategy for historical and low-resourced languages:

During the development process, we conducted numerous experiments to find the best training strategy. We can summarize our observations in the following points:

- Pretrained multilingual language models are trained on many languages[1], but not on all languages. Therefore, for languages that are not part of mBERT and XLM-RoBERTa languages, we recommend starting with unsupervised training with monolingual texts using MLM training objective. Nevertheless, For languages that use alphabets not included in the vocabularies of mBERT, and XLM-RoBERTa, such as Syriac and Coptic, we have to train a multilingual model from scratch.
- If parallel corpora are available, we suggest to perform unsupervised training with bilingual sentences after the training on monolingual texts to achieve the best outcome.
- If manual alignments at the word level are available, we suggest to perform supervised training at the last phase after unsupervised training to achieve better results. The experiments showed that running unsupervised training after supervised training will result in higher AER.

[1]mBERT is trained on 104 languages and XLMRoBERTa on 100 languages

Chapter 4

Translation Alignment Evaluation

Evaluation plays a crucial role in the development cycle of any system or software. It supports performance assessment and estimating the quality and reliability of the model and aids developers and practitioners in fine-tuning and optimizing experiments' parameters and exploring areas for performance improvement. Nevertheless, assessing models' performance requires error-free reference data where the output is already known for a given input. This kind of data is called gold standards. Performance evaluation can be conducted quantitatively and qualitatively. Quantitative evaluation employs various measures to generate numerical values that describe the overall behavior of the model and reflect its quality; performance comparison among models is based on these numerical values. Qualitative evaluation allows for in-depth investigation and understanding of the studied models and involves human judgment.

Word alignment is a central component in statistical machine translation systems, therefore, several research papers studied the performance evaluation of word alignment models and its impact on the performance of machine translation models (20, 85, 233). However, Alignment evaluation is a challenging task and subject to several factors (8), such as the target application of the alignment, language pairs, text genre, and text granularity level. Blinker (163) and ARCADE (246) projects made the first efforts to define evaluation metrics for the alignment task.

In this chapter, we address the evaluation task of translation alignment models, review the available benchmark datasets, discuss the current practice, define its drawbacks and limitations, and propose two new metrics to overcome them. Further, we present our work to generate alignment guidelines and gold standards for three language pairs Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Portuguese. Finally, we introduce EvAlign, a visual analytic framework to support quantitative and qualitative evaluation of alignment models and gold standards. We explain the framework structure,

visual design, and usage scenarios and list our observations on the hosted datasets and alignment models. This chapter is based on the following publications:

- **Yousef T**, Jänicke S. *Visual Evaluation of Translation Alignment Data.* InProc. **EuroVis** 2022 (Vol. 22) (268).

- **Yousef T**, Heyer G, Jänicke S. EvAlign*: Visual Evaluation of Translation Alignment Models.* **EACL** system demonstration 2023 (274).

4.1 Alignment Gold Standards

Gold standards are essential for evaluating the performance of NLP models. They are used as references to compare models' predictions to assess their quality and reliability. Developing gold standards for translation alignment is a challenging task. It requires annotating a selection of bilingual sentences manually. The annotation process must be accomplished by multiple domain experts (at least 2) annotating the same parallel sentences to avoid any bias. Several tools are available for generating manual alignments, such as UGARIT(272), CLUE-Aligner (27), SWIFT Aligner(95), and Yawat (94).

Creating alignment gold standards for a language pair also requires defining annotation guidelines to be followed by annotators to ensure high agreement and consistency. The guidelines must discuss and solve all ambiguous and challenging cases when the translation is not literal or sentences have very different grammatical structures. Moreover, Inter-Annotator Agreement (IAA) can be computed to validate the reliability and quality of the alignment guidelines and gold standard. IAA reflects the overlap among annotators' choices and can be calculated using F1 score (Dice coefficient (70)) and/or Cohen's Kappa Coefficient (58)

The gold standard dataset is a set of aligned sentences, and each parallel sentence is provided with a set of translation pairs assigned one of two categories, Sure (S) or Possible (P). Sure translation pairs indicate clear and unambiguous alignments, whereas Possible translation pairs occur when translations are inaccurate or non-literal or have allusions and idiomatic expressions. The categorization of translation pairs is accomplished during the annotation process if the annotation tool offers a such option. Or when merging the alignments of multiple annotators. In this case, the overlap among annotators' choices is considered as Sure alignments and their union as Possible alignments. Figure 4.1 shows the most commonly used format for gold standards. It consists of three files, the source sentences, target sentences, and translation pairs files. Each sentence is assigned a unique ID, and each translation pair consists of the

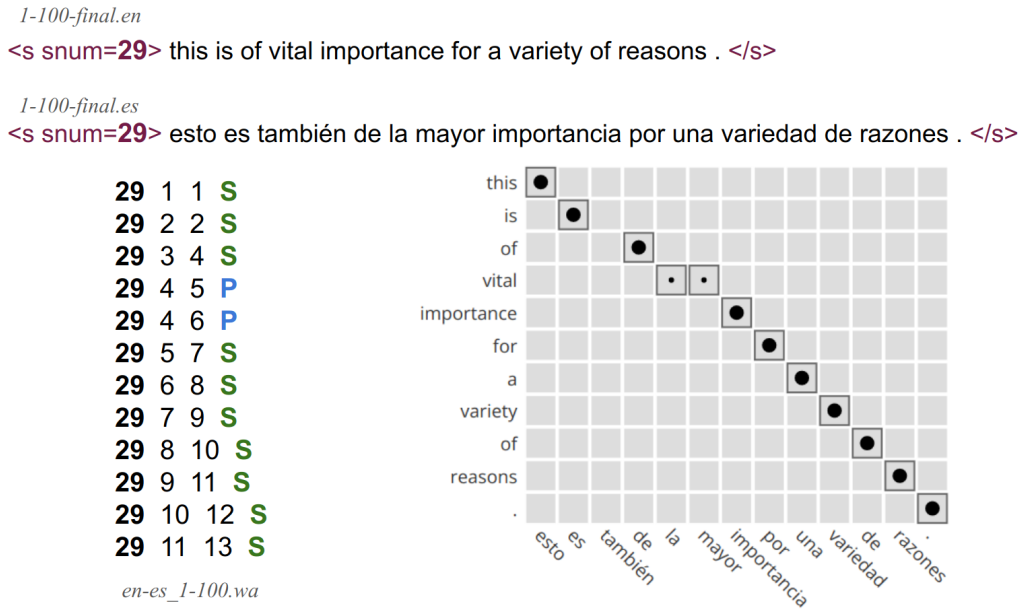


Figure 4.1: Example of gold standard format, Sentence number 29 from the English-Spanish dataset (97).

sentence ID, source token ID within the source sentence, target token ID within the target sentence, and link type.

While there are several alignment gold standard datasets for modern European languages, including English, there is a lack of such datasets for ancient or classical languages. To bridge this gap and to be able to evaluate our automatic alignment model, we developed alignment gold standards and guidelines for Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Portuguese.

Table 4.1 provides an overview of the available evaluation datasets with information about the corpus size, IAA, and text source. An extended list with download links is available on GitHub [1]. Further, We surveyed literature published in the last three years to review the evaluation practice and the employed evaluation datasets. Table 4.2 summarizes the survey results and shows that most researchers evaluated their models mainly on three alignment datasets, German-English (180), English-French, and Romanian-English (165). However, several datasets are available in various languages but have not yet been used for performance evaluation (Table 4.1).

[1] `https://github.com/TariqYousef/Word-Alignment-Gold-Standards`

Source	Language Pair	# Sentences	IAA	Text Type
(179)	English-German	854		Verbmobil
(165)	Romanian-English	248		
	English-French	447		Hansard
(137)	English-Spanish	500		Europarl
(134)	Czech-English	2400	93%	Penn Treebank corpus (WSJ)
(98)	English-Portuguese	100	89.5 %	Europarl
	English-Spanish	100	86.7 %	Europarl
	English-French	100	90.8 %	Europarl
	Portuguese-Spanish	100	93.2 %	Europarl
	Portuguese-French	100	93.5 %	Europarl
	Spanish-French	100	96.5 %	Europarl
(158)	Dutch-English	1500	84-94 %	Journalistic, Newsletters, and Medical Reports
(107)	English–Swedish	1164	91.3%	Europarl
(228)	Icelandic-English	604		ParIce Corpus Project
(271)	Ancient Greek-English	275	86.17%	Perseus Digital Library
	Ancient Greek-Portuguese	183	83.31%	Perseus Digital Library
(273)	Ancient Greek-Latin	100	90.50%	DFHG Project
(102)	Myanmar-English	500	91.56%	Myanmar- English ALT parallel corpus

Table 4.1: An overview of the existing alignment gold standard datasets.

4.1.1 Alignment Gold Standard For Ancient Greek

In the context of our work on developing an automatic alignment system for historical languages (271, 273), we developed alignment gold standard datasets for Ancient Greek-English, Ancient Greek-Latin, and Ancient Greek-Portuguese[1] in collaboration with domain experts[2]. For every language pair, two experts, who had previous experience with the manual alignment in UGARIT, aligned the same parallel passages following predefined alignment guidelines.

[1]We are currently working on a gold standard dataset for Ancient Greek-Persian. The guidelines are already available on GitHub.

[2]The guidelines and manual annotation of the selected sentences is done by Chiara Palladino, Farnoosh Shamasian for Greek-English, Anise Ferreira and Michel Reis for Greek-Portuguese, and Chiara Palladino and David Wright for Greek-Latin. Tariq Yousef was responsible for the data pre- and postprocessing, merging the alignments, producing the gold standard datasets, and calculating the IAA.

Paper	EN-CZ	EN-DE	EN-FR	EN-HI	EN-RO	EN-JA	EN-ZH	EN-IC	EN-VI	EN-FA	EN-AR
(230)							✓				x
(92)		✓	✓		✓						
(71)		✓	✓		✓						
(114)	✓	✓	✓	✓	✓					✓	
(278)		✓	✓		✓						
(52)		✓	✓		✓		✓				
(106)			✓		✓						
(263)			✓	✓	✓						
(172)		✓	✓		✓	✓	✓				
(74)		✓	✓		✓	✓	✓				
(281)	✓	✓									
(228)	✓	✓	✓					✓			
(160)		✓	✓		✓						
(176)	✓	✓	✓		✓	✓			✓		
(51)		✓	✓		✓		✓				
(53)		✓	✓	✓	✓						
(259)		✓	✓		✓						

Table 4.2: An overview of alignment gold standard datasets that have been used for performance evaluation in the literature.

Corpus

The aligned dataset for **Ancient Greek-English** and **Ancient Greek-Portuguese** language pairs includes three Ancient Greek texts and their translations into English and Portuguese. The annotators aligned passages from the *Iliad*, Plato's *Crito*, and Xenophon's *Cyropedia*. The chosen texts provided sufficient diversity of language (Homeric to Koine Greek) and of text genre (poetry, prose, and dialogue). The selected translations were mostly modern for both English and Portuguese (46, 159, 170, 254). For **Ancient Greek-Latin**, we randomly selected 100 fragments from the *Digital Fragmenta Historicorum Graecorum* (DFHG) corpus[1]. The corpus includes more than eight thousand quotations and text-reuses (*fragments*) in Ancient Greek with their Latin translation (33, 34).

Alignment Guidelines and Workflow

For each language pair, both the gold standard and the guidelines were created by the same two experts. The first draft of the guidelines was created through multiple meetings, prior to aligning the corpus. The basic structure was developed starting from an already existing model (137), and considerably expanded to include various

[1] `https://www.dfhg-project.org`

language-specific issues that the experts had previously encountered. Then, the experts aligned a subset of the corpus in order to test the general consistency and feasibility of the guidelines. During this phase, there was a brief discussion for each new issue, and a preferred annotation style or an improvement in the guide was agreed upon. Each change was incorporated in the final version, and the alignments were revised accordingly. After the subset was completed, the experts completed the alignment without further discussions, to ensure that the consistency and efficacy of the guidelines could be appropriately tested.

Our alignment guidelines do not distinguish between sure and possible alignments as proposed by Och and Ney, but when combining the alignments of the two annotators, we defined sure and possible alignment sets for every sentence as follows:

$$S = A_1 \cap A_2 \quad , \quad P = A_1 \cup A_2$$

A_1 and A_2 are the alignment sets created by the first and second annotators, respectively. S denotes sure alignments which include all translation pairs where both annotators agree. P denotes possible alignments where the translation pairs are aligned by at least one annotator.

Inter-Annotator Agreement

We measured the IAA on the manually annotated texts, considering the agreement between the annotators on the aligned tokens and the unaligned ones. Since Ugarit allows annotators to create multi-word alignments (1-to-N, N-to-1, and N-to-N), we converted multi-word alignments to 1-to-1 pairs to consider the partial matching of the translation pairs. For example, if the phrase "The son" is aligned to "υἱός", it is converted to two 1-1 alignment pairs (The, υἱός) and (son, υἱός). Let A_1 and A_2 be the flattened translation pairs created by the first and second annotators, respectively, we employed Sørensen–Dice coefficient to compute the IAA as follows:

$$IAA = \frac{2 * |A_1 \cap A_2|}{|A_1| + |A_2|} \tag{4.1}$$

Table 4.3 provides and overview of the IAA results and the gold standards datasets. The alignment guidelines and gold standard datasets are available on Github[1].

[1] https://github.com/UgaritAlignment/Alignment-Gold-Standards

	Greek-English	Greek-Portuguese	Greek-Latin
Sentences	274	183	100
Greek Tokens	5.208	3.216	1.766
Greek Types	2.285	1.587	1.028
Eng/Por/Lat Tokens	7.284	3.710	1.623
Eng/Por/Lat Types	1.580	1.355	1.081
Sure Alignments	6.036	3.028	1.380
Possible Alignments	1.388	864	213
IAA	86.08%	83.31%	90.50%

Table 4.3: An overview of the gold standards datasets and their Inter-Annotator Agreement.

4.2 Evaluation Metrics

Translation alignment can be considered an information retrieval task (8); therefore, we can employ information-retrieval evaluation metrics such as Recall, Precision, and F1 for assessing the performance of automatic alignment models. Ahrenberg et al. (8) introduced different implementations of Recall and Precision within the ARCADE and PLUG projects. Further, Och and Ney (180) introduced the Alignment Error Rate (AER), trying to combine and balance Recall and Precision. These metrics are based on the overlap between the model's predictions A with the Sure (S) and Possible (P) alignment sets of the gold standards (Equation 4.2).

$$
\begin{aligned}
Precision = \frac{|A \cap P|}{|A|} \quad , \quad Recall = \frac{|A \cap S|}{|S|} \\
F1 = \frac{2 * Precision * Recall}{Precision + Recall} \\
AER = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|}
\end{aligned}
\tag{4.2}
$$

Lambert et al. (137) studied the influence of the amount of Sure and Possible alignments on the AER and concluded that AER would be smaller when the S/P ratio is low, and vice versa. Figure 4.2 shows the correlation between AER and the S/P ratio on six gold standard datasets, the line chart confirms Lambert et al.'s conclusion.
Ayan and Dorr (20) proposed a new metric, the Consistent Phrase Error Rate

(CPER), to evaluate translation alignment in the context of phrase-based machine translation aiming to assess the impact of automatic alignment on machine translation performance. Lardilleux et al. (139) suggested a protocol to evaluate alignment models by comparing their alignments against bilingual lexicon.

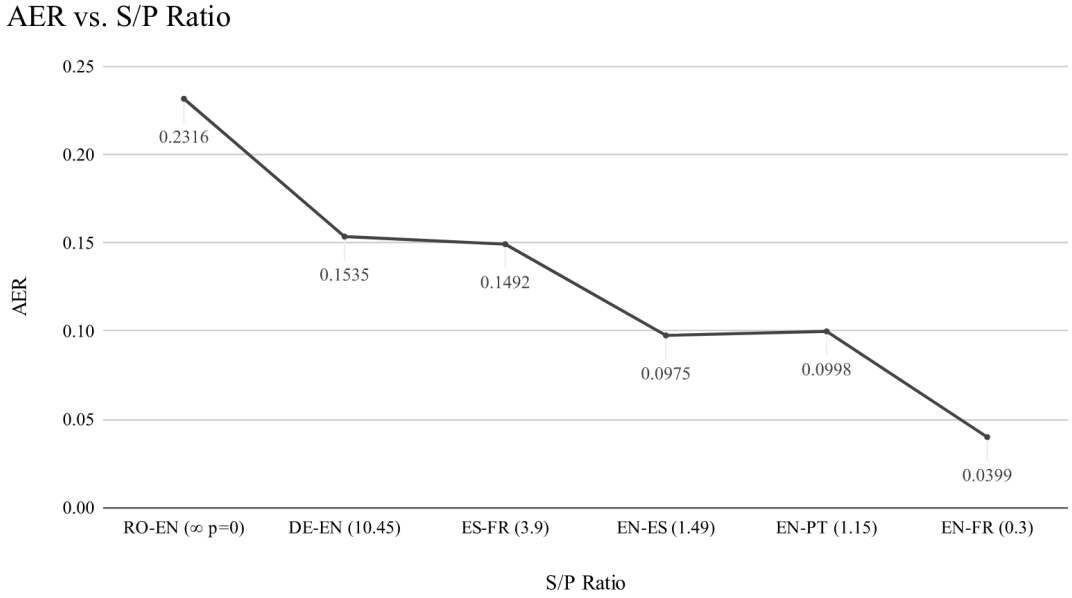


Figure 4.2: The correlation between AER and S/P Ratio. The alignment model used for this illustration uses Argmax method with fine-tuned mBERT Embeddings.

4.2.1 AER Limitations

Evaluating the performance of translation alignment models is a complex task, even for humans. In many cases, it is challenging to tell if an alignment between two tokens/phrases is entirely correct because that relies on several factors, mainly the text genre, context, translation quality, and human annotator's knowledge.
Although AER is commonly used for evaluating automatic alignment models, it does not reflect the actual quality of the model performance. While developing and evaluating automatic alignment models, we encountered several cases where AER was limited and failed to capture the real error rate. We summarize our observations in the following:

- AER is highly affected by the gold standard dataset, i.e., the selection of sure and possible translation pairs and their proportions of the whole dataset. And the gold standard alignments are subject to the alignment guidelines and annotators' agreement, which is also influenced by the characteristics of the selected corpus, the annotators'

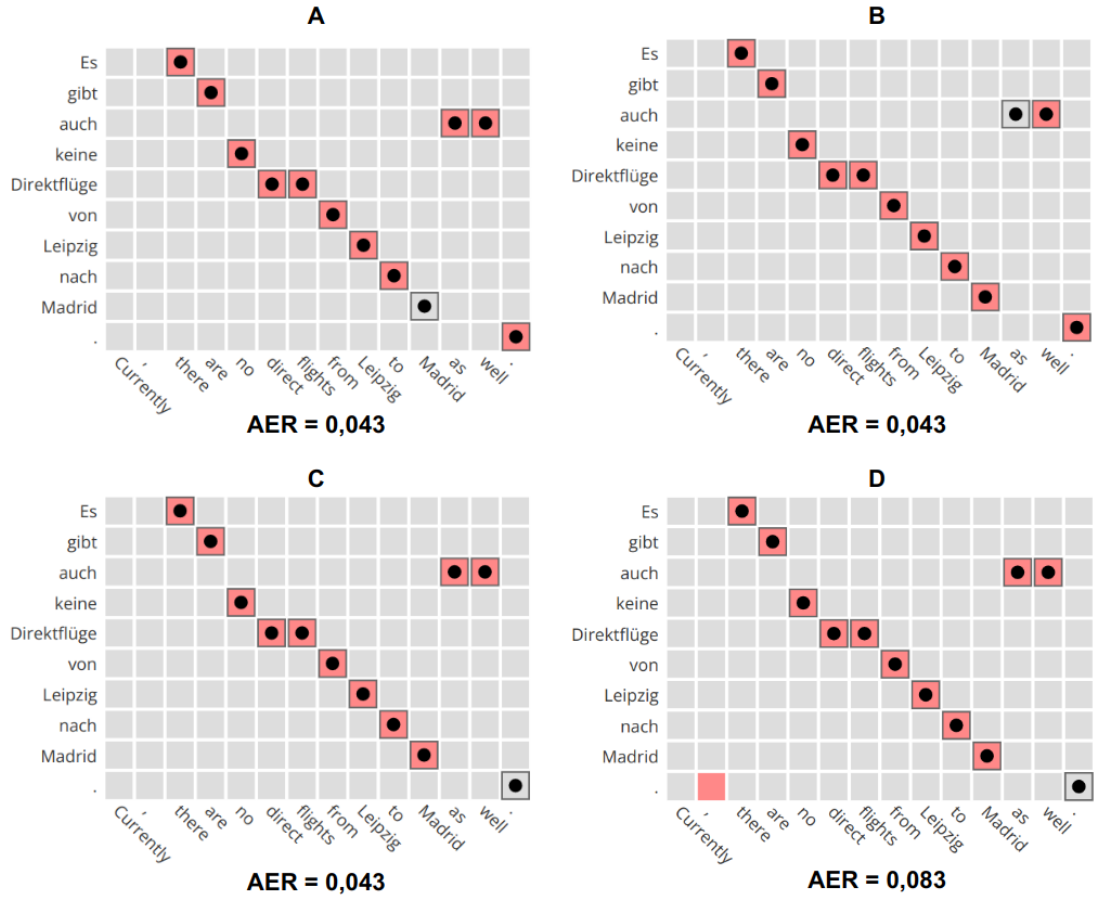


Figure 4.3: AER Limitations, the bold circles means gold standard sure alignments and colored cells represent model's output. A) The model failed to align *Madrid*. B) The model failed to align *auch* to *as well*. C) The model failed to align ".". D) The model aligned "." incorrectly.

knowledge, and the target application. That means it might be possible to have correct alignments predicted by the models, but the gold standards do not consider them. Thus, AER will treat them as incorrect alignments. We encountered such cases repeatedly while inspecting the existing gold standard datasets.

- AER is intolerant; it considers all tokens equally important and there is no distinction between function words and context words. In the example illustrated in Figure (4.3A and C), AER penalizes a missing alignment of the full-stop the same as missing alignment of *Madrid*.

- AER fails to capture phrase misalignments. In Figure 4.3B, the German word

auch must be aligned to the English phrase *as well*, producing two sure alignments $auch - as$ and $auch - well$. Nevertheless, If an alignment model aligns only a part of the phrase, $auch - well$, this will be considered a correct alignment, while it is not because there is no constraint saying that the model must produce the two sure alignments together in order to count them as correct alignments.

- AER does not consider null-alignments, i.e. tokens with no translation equivalents in the parallel sentence. In Figure 4.3, the word *currently* has no translation equivalent in the German translation (i.e. null-alignment). The model also did not align it to any German word. However, AER doesn't reward the model's success in not linking null-alignment with any correspondence.

- Incorrect alignments have more impact on AER than missing alignments. While a missing alignment reduces the Recall slightly, incorrect alignment reduces both Precision and Recall (Figure 4.3C, D).

4.2.2 Alignment Coverage and Phrase Alignment Accuracy

We proposed two metrics to overcome the limitations above. The Alignment Coverage represents the portion of the aligned tokens out of all tokens in the dataset. It can computed for the gold standard dataset and for models' predictions. We compute *Coverage* as follows:

$$Coverage = 1 - \frac{|S_n| + |T_n|}{|S| + |T|} \tag{4.3}$$

Where S and T are the sets of all tokens in the source and target sentences, respectively, S_n and T_n are the sets of null-alignments in the source sentences and target sentences. When the gold standard dataset has high coverage, it indicates that the translation is almost literal and finding translation correspondences was straightforward. Moreover, low coverage indicates that the translation is nonliteral, uses a different grammatical structure, or that the alignment guidelines are so strict.

Figure 4.4 compares nine datasets regarding alignment coverage. The English-French and Spanish-French datasets have the highest coverage (95%), whereas the German-English and English-Portuguese datasets have the lowest coverage (90%). Figure 4.5 shows the alignment coverage of 20 alignment models with the German-English dataset. The dotted line represents the gold standard coverage. The coverage of *Itermax_FT_mBERT* matches the gold standard, and the Match model with different embeddings achieved higher coverage than the gold standard. If the model has higher

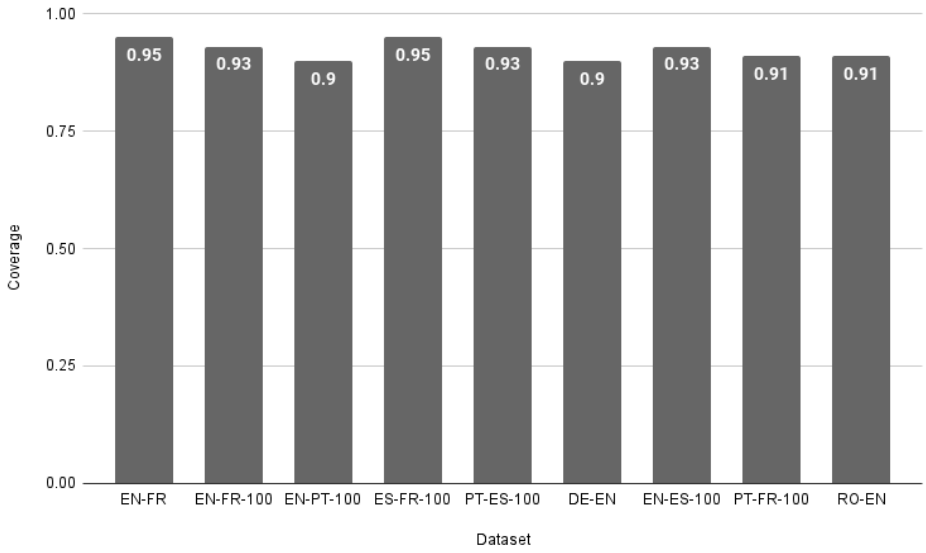


Figure 4.4: A comparison among nine benchmark datasets regarding the alignment coverage.

alignment coverage than the gold standard, that means the model tries to find translation correspondences for the gold standard null-alignments and that increases the AER.

The Phrase Alignment Accuracy (PAC) measures the model's ability to align phrases. Phrase alignment appears when a token in one sentence is aligned to multiple tokens in the corresponding translation (one-to-many or many-to-one), or when multiple tokens in one sentence are aligned to multiple tokens in the corresponding sentence (many-to-many). Our definition of phrase does not constrain that the tokens must be consecutive. However, the phrase is correctly aligned if all its tokens are aligned with each other. For instance, the English phrase *public health policy* and the German equivalent *Gesundheitspolitik* are aligned correct if, and only if the model predicts $public - Gesundheitspolitik$, $health - Gesundheitspolitik$ and $policy - Gesundheitspolitik$ pairs. Because all tokens contribute to the meaning of the phrases, and missing any token changes the meaning or make it incomplete. We compute PAC as stated in Equation 4.4:

$$PAC = \frac{|P_m \cap P_{gs}|}{|P_{gs}|} \tag{4.4}$$

Where P_{gs} is the aligned phrases set of the gold standard, and P_m is the set of predicted aligned phrases by the model.

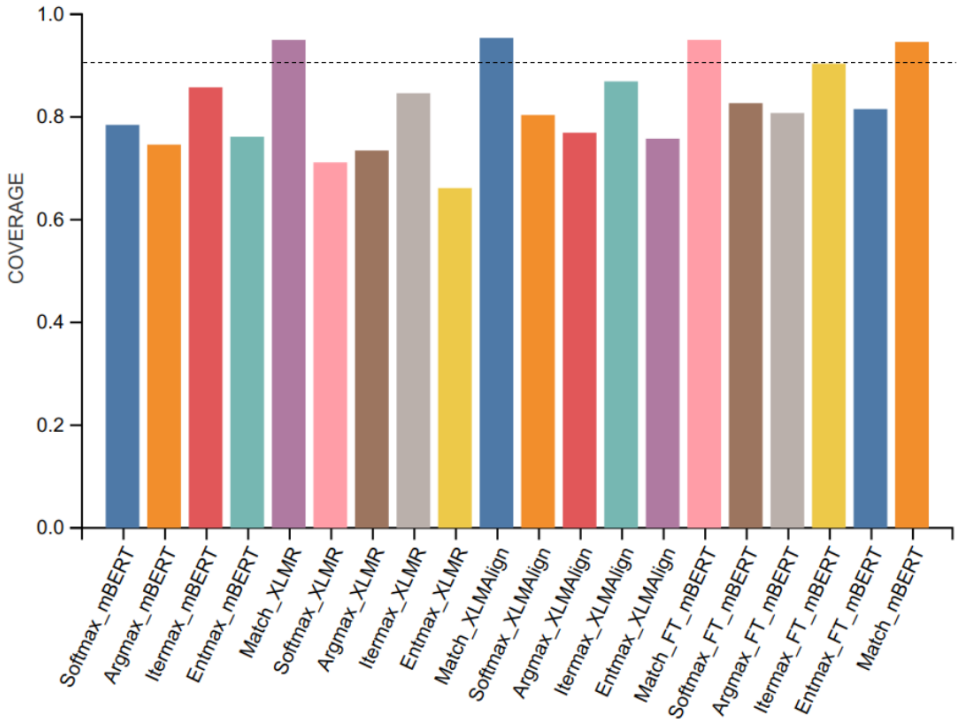


Figure 4.5: A comparison among 20 alignment models regarding the alignment coverage on the German-English dataset.

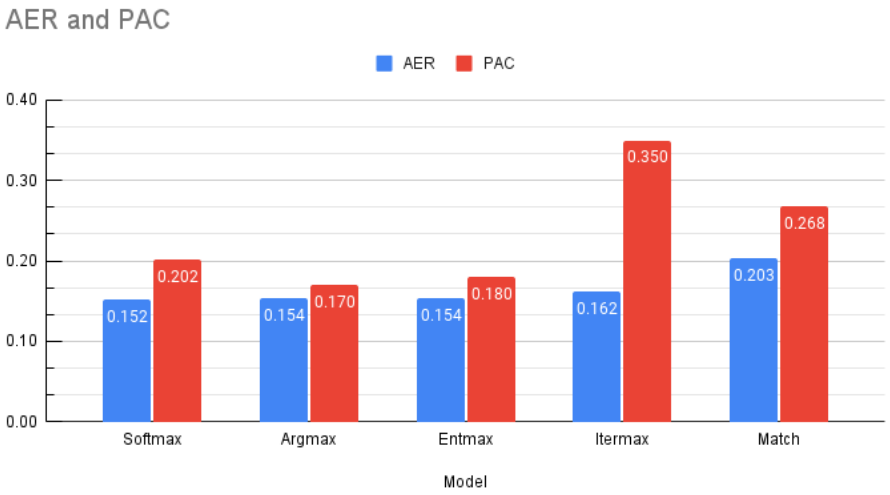


Figure 4.6: Comparison among five alignment models on the German-English dataset regarding AER and PAC.

Figure 4.6 compares the performance of the best five alignment models on the German-English dataset, the models use the fine-tuned mBERT embeddings. As we can see, Softmax achieved the lowest AER and slightly outperformed Itermax with 0.01 points. Nevertheless, Itermax outperformed Softmax and other models significantly in aligning phrases.

4.3 Visual Evaluation of Translation Alignment Models

The aforementioned limitations of the quantitative evaluation metrics revealed the need for qualitative evaluation for better performance assessment. For this purpose, we proposed EvAlign, a visual analytics framework for quantitative and qualitative evaluation of automatic translation alignment models. EvAlign offers several visualization views enabling developers to visualize their models' predictions and compare the performance of their models with other baseline and state-of-the-art models. Via different search and filter functions, developers and practitioners can also inspect the frequent alignment errors and their positions. EvAlign hosts nine gold standard datasets and the predictions of multiple alignment models. Our usage scenarios document that the insights gained with our solution can be pivotal for improving automatic alignment models. The tool is extendable, and adding additional datasets and models is straightforward. EvAlign can be deployed and used locally and is available on GitHub[1] and online[2].

4.3.1 Implementation Details

EvAlign hosts nine benchmark datasets in total, German-English[3], French-English, and Romanian-English [4], in addition to six other datasets[5] (98) (English-French, English-Spanish, English-Portuguese, Spanish-French, Portuguese-Spanish, and Portuguese-French) that have not been used before for evaluation.

Regarding the alignment models, we selected embeddings-based $Softmax$, $Entmax$ (74), $Argmax$, $Itermax$, and $Match$ (114) with different contextualized embeddings, namely, multilingual-BERT (mBERT) (68), XLM-RoBERTa (61), fine-tuned mBERT[6], and XLM-Align[7] (53). In addition to Giza++, fast_align, and EfLoMAl for the German-English, English-French, and Romanian-English datasets. Our selection was subject to

[1] https://github.com/TariqYousef/EVALign
[2] http://evalign.info/
[3] https://www-i6.informatik.rwth-aachen.de/goldAlignment/
[4] https://web.eecs.umich.edu/~mihalcea/wpt/index.html
[5] https://www.hlt.inesc-id.pt/w/Word_Alignments
[6] https://github.com/neulab/awesome-align
[7] https://huggingface.co/microsoft/xlm-align-base

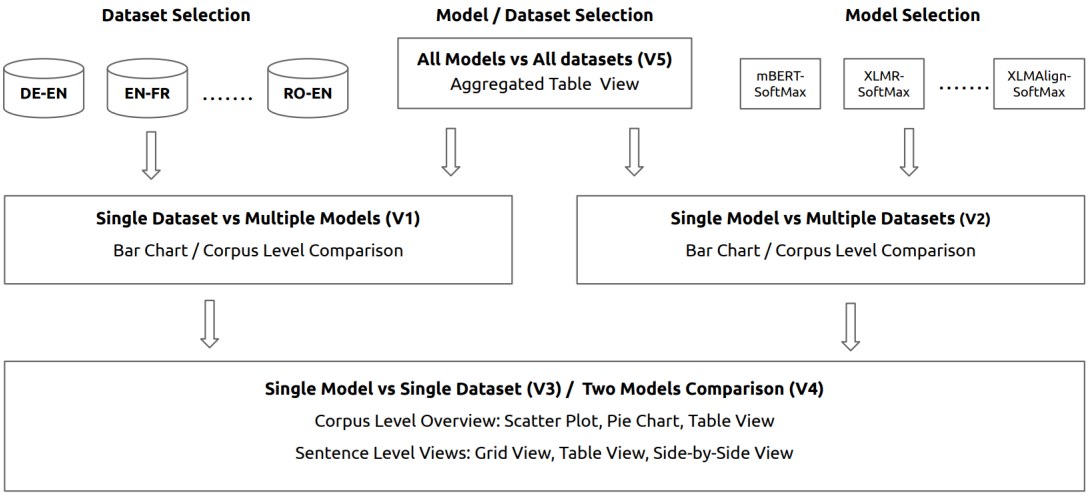


Figure 4.7: Overview of EvAlign

implementation availability and reproducibility. Unfortunately, many literature papers do not provide any information about the source code or their models' predictions. We used the default implementation provided by authors in their GitHub repositories with the default parameters.The backend API is implemented using Django framework and Postgres database, while the visualization views are created with React JS and D3.js.

4.3.2 User Interface

Figure 4.7 illustrates the usage workflow; users start navigating through the tool by selecting a dataset from the landing page which lists all hosted benchmark datasets, selecting a model from the models page which lists all hosted models, or selecting a model or dataset from the aggregated overview. EvAlign offers five main views:

Single Dataset vs Multiple Models (V1). This view provides a performance overview of the alignment models hosted on EvAlign over the selected dataset using a *bar chart*. The overview allows users to select among different quantitative evaluation metrics, namely, Precision, Recall, F1, AER, Coverage, PAC, and the number of translation pairs. The view also visualizes all sentences of the selected dataset using a grid view allowing users to inspect the possible and sure alignments and assess their correctness and coverage (Figure 4.8A) . From this view, users can select a single model to inspect its performance on a specific dataset.

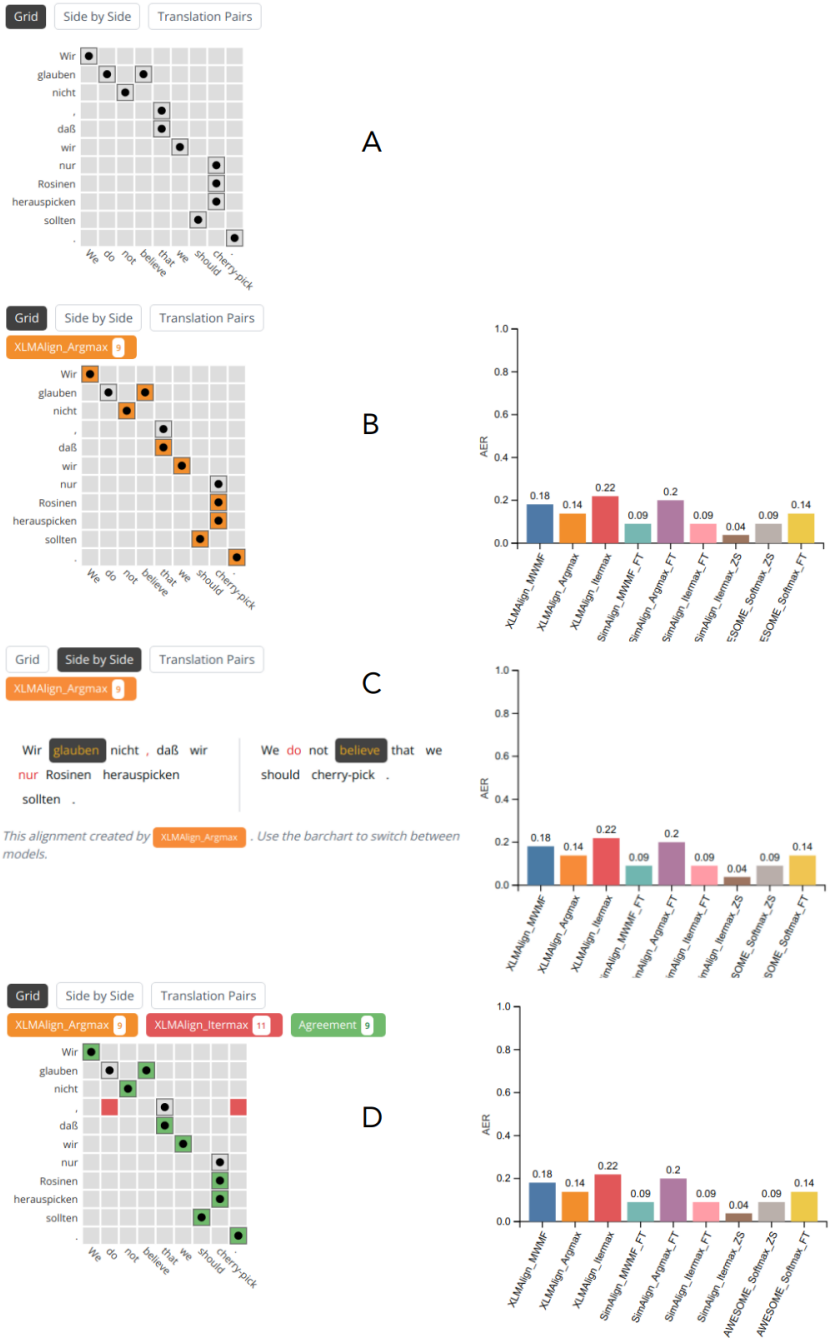


Figure 4.8: Sentence-level views, A) the default view when no model is selected, showing sure (big dots) and possible (small dots) alignments. B) visualizing the alignments of a single model. C) the side-by-side view. D) the gird view visualizes the alignments of two models and their agreement.

Single Model vs Multiple Datasets (V2). This view provides a summarized performance overview of the selected alignment model over different benchmark datasets using a *bar chart* that allows switching among different evaluation metrics. Selecting a dataset will forward the user to *Single Models vs Single Dataset* view.

Single Model vs Single Dataset (V3). This view offers various corpus-level and sentence level visualization views providing the user with all needed functions to inspect the the dataset sentences and explore the alignments predicted by the selected model. This view aggregate wrong alignments, missing alignments and correct alignments to facilitate the analysis of the model performance. Further it shows the relation of the different evaluation metrics with sentence lengths. The predicted alignments are visualized with *Grid*, *Side-by-side*, and *Table* views. Moreover, it offers various sorting, filtering and searching options to support qualitative evaluation (Figures 4.8B and 4.8C).

Two Models vs Single Dataset (V4). From V3, users can select a model to compare it to a reference model at sentence-level. The view visualizes the agreement and disagreement between the predictions of the two models using the *Grid*, *Side-by-side*, and *Table* views (Figure 4.8D and 4.9).

Figure 4.9: Table view shows the agreement of two models predictions at sentence level. Sentence *1* form DE-EN dataset.

All Models vs all Datasets (V5). In this aggregated view, all hosted datasets and models are presented in a table. Users can switch between different quantitative metrics with different sorting options (Figure 4.10).

AER | Precision | Recall | F1 | # Translation Pairs

Models/Datasets	EN-FR	EN-FR-100 ↑	EN-PT-100	ES-FR-100	PT-ES-100	PT-FR-100	DE-EN	EN-ES-100	RO-EN
Softmax_FT_mBERT	0.0407	**0.133**	0.1027	0.1403	0.0589	0.1564	0.1524	0.0958	0.226
Entmax_FT_mBERT	0.0392	**0.1354**	0.1013	0.1446	0.0611	0.1593	0.1539	0.0954	0.2293
Argmax_FT_mBERT	0.0399	**0.1359**	0.0998	0.1492	0.0659	0.16	0.1535	0.0975	0.2316
Itermax_FT_mBERT	0.067	**0.1362**	0.1207	0.1328	0.0739	0.1528	0.1616	0.109	0.221
Softmax_mBERT	0.0514	**0.1487**	0.1179	0.1478	0.0777	0.1595	0.1828	0.1093	0.2457
Entmax_mBERT	0.0511	**0.149**	0.1162	0.1533	0.0799	0.1686	0.1889	0.1107	0.2547
Argmax_mBERT	0.0519	**0.1579**	0.1245	0.1613	0.0797	0.1701	0.1971	0.1219	0.2645
Itermax_mBERT	0.0757	**0.1612**	0.1397	0.1559	0.0918	0.1747	0.1943	0.1292	0.2413
Argmax_XLMAlign	0.0569	**0.165**	0.1336	0.1721	0.0873	0.1937	0.1694	0.1165	0.2535
Softmax_XLMAlign	0.0792	**0.1697**	0.146	0.1637	0.0937	0.1944	0.1846	0.1272	0.2527
Entmax_XLMAlign	0.0736	**0.1714**	0.1407	0.172	0.0949	0.1974	0.1879	0.1221	0.2664
Itermax_XLMAlign	0.0903	**0.1753**	0.1631	0.1693	0.1109	0.2113	0.1845	0.1375	0.2473
Match_FT_mBERT	0.097	**0.1774**	0.1502	0.1831	0.1265	0.2187	0.203	0.1372	0.2493
Match_mBERT	0.1132	**0.1837**	0.1827	0.1887	0.1304	0.2233	0.2264	0.1499	0.2635
Argmax_XLMR	0.0655	**0.1924**	0.1324	0.1768	0.0743	0.2022	0.1923	0.1222	0.2591
Itermax_XLMR	0.0902	**0.1951**	0.1564	0.1784	0.1018	0.2204	0.1986	0.1418	0.2537
Softmax_XLMR	0.0918	**0.1953**	0.157	0.1771	0.0891	0.2162	0.2306	0.1406	0.2864
Entmax_XLMR	0.0901	**0.2115**	0.1693	0.1923	0.0925	0.2248	0.245	0.1463	0.302
Match_XLMAlign	0.1317	**0.2122**	0.1907	0.1976	0.1378	0.2523	0.2262	0.1711	0.2753
Match_XLMR	0.1357	**0.2192**	0.1864	0.2016	0.1329	0.2541	0.2369	0.1704	0.2781

Figure 4.10: Aggregated table view allows to compare the quantitative metrics of all models on all datasets.

4.3.3 Visual Design

EvAlign offers a variety of corpus and sentence level visualization views in addition to several searching and filtering functions. When designing EvAlign, we consulted the text alignment visualization survey (267) and adapted Schneiderman's Information Seeking Mantra (219) "Overview first, zoom and filter, then details-on-demand" to facilitate interactive navigation through the benchmark datasets and alignment models.

4.3.3.1 Corpus-level Views

Corpus-level views provide comprehensive overviews of the compared models by visualizing aggregated statistics and evaluation metrics at the dataset level. A ***bar chart*** on the dataset page is placed in the upper left corner, allowing users to compare available alignment models. A button bar is located above the bar chart, allowing users to switch

between seven evaluation metrics: AER, Precision, Recall, F1, Number of Translation Pairs, Coverage, and PAC. Each model is assigned a unique color (Figure 4.11A) and hovering a bar will show a tooltip with the corresponding information.

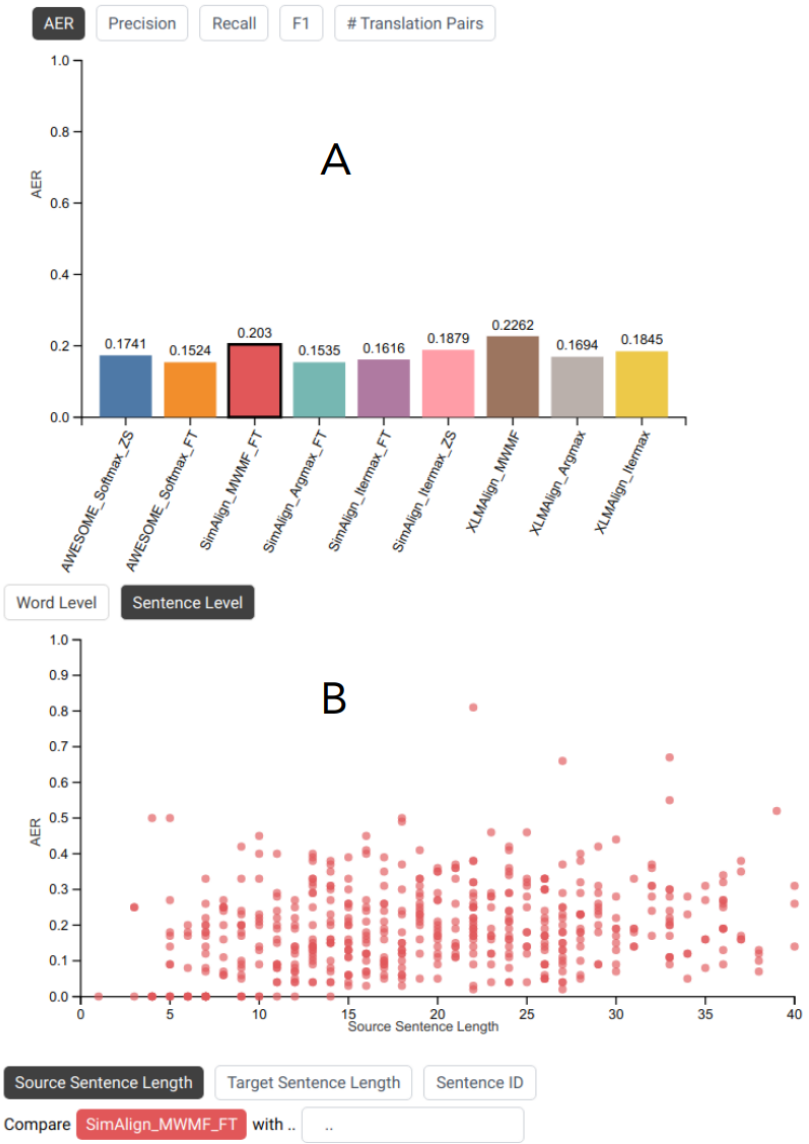


Figure 4.11: Single Dataset-Single Model Corpus-Level view, A) Bar Chart to compare the performance of different alignment models according to a selected metric. B) Scatter Plot shows the relation between sentences length and the selected metric.

Selecting a model with a mouse click on the corresponding bar will load a ***pie chart*** that shows a comparison between the model predictions and the gold standard. We

distinguish among three main sets (Figure 4.13): i) *correct alignments*, where the model predictions match the gold standard, shown in green. ii) *wrong alignments*, where the model failed to align the translation pairs correctly, shown in red. iii) *missing alignments*, the pairs the model was supposed to align, shown in orange. Clicking on any of the three sets will load the corresponding translation pairs in the neighbouring table, which aggregates the translation pairs and shows them with their frequency. Moreover, the translation pairs are clickable, and the corresponding gold standard sentences with sentence-level views will be displayed when clicked. Further, users can switch between the *pie chart* and the *scatter plot*, which displays the relation between the sentence length (x-axis) and the selected evaluation metric (y-axis) of the selected alignment model; each sentence is presented as one dot (Figure 4.11B).

The *scatter plot* helps users detect outliers and interesting observations, such as the relation between AER and the sentence length. Moreover, a range selector allows filtering of the dataset by selecting multiple sentences to be visualized at the sentence level for more detailed inspection. Further, the evaluation metrics will be calculated for the selected sentences and displayed under the scatter plot. This allows users to eliminate subsets (for example, short or long sentences) and see how these subsets affect the quantitative evaluation metrics. The selected sentences will be displayed as paginated list of sentence-level views. The tool provides sorting options according to the selected metric.

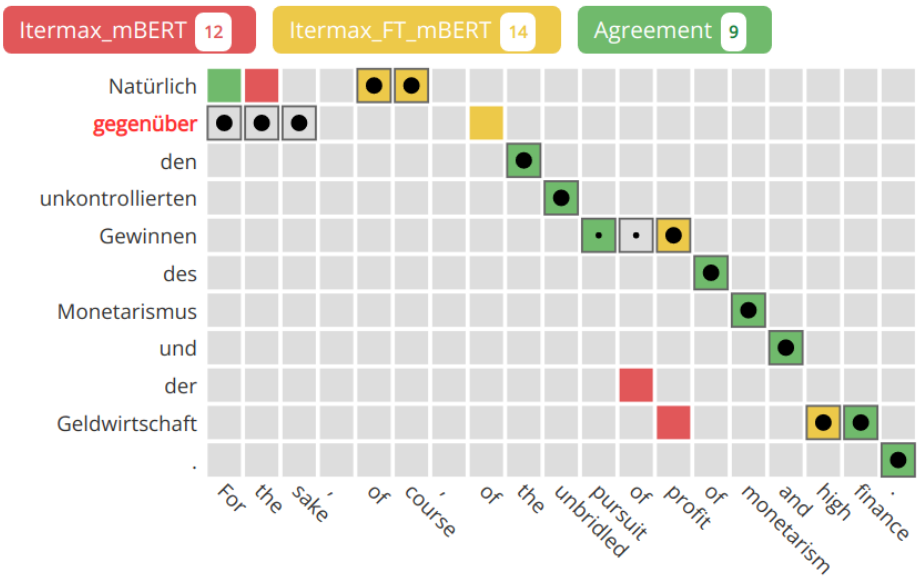


Figure 4.12: A Grid View to visualize the alignment at sentence level.

4.3.3.2 Sentence-level Views

The sentence-level views aim to show the alignment among words of the source and target sentences. The framework provides two sentence-level views, namely, *grid view* and *side-by-side view*. The views are accompanied with a *bar chart* showing the sentence-level evaluation metrics of the hosted models and enabling users to select a model to visualize its output for the corresponding sentence. The *grid view* places the two sentences as a grid. The source sentence tokens are placed vertically, and the target tokens are placed horizontally. The gold standard Sure and Possible alignments will be displayed in the corresponding cells as big and small dots, respectively. The *grid view* is suitable for visualizing the alignments of a single model by coloring the corresponding cells with the model's unique color. It is also appropriate to visualize the alignments of two models and their agreement (Figure 4.12).

The *side-by-side view* places the two sentences alongside each other; it utilizes the mouse hover to highlight the hovered token and the aligned tokens in the parallel sentence. The current implementation of this view allows visualizing the alignment of a single model, and users can switch between models via a neighboring *bar chart*.

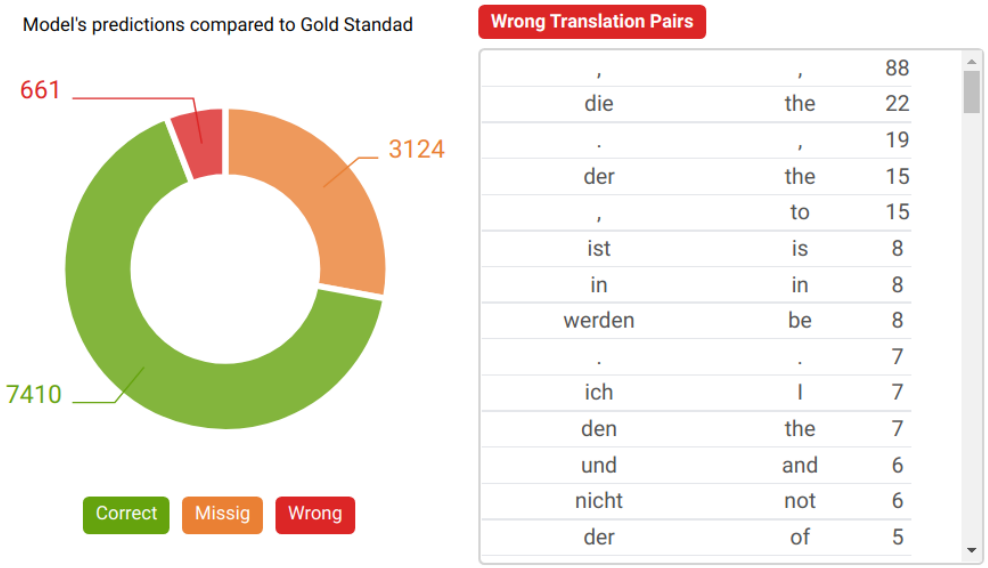


Figure 4.13: A Pie Chart shows the overlap between the model's predictions and the gold standard in three categories. The neighboring table shows the translation pairs with their frequency of a selected category.

4.3.4 Usage Scenarios

The framework offers a variety of usage scenarios that can be summarized as follows:

Gold standard quality control.

Visualizing the gold standard datasets using the *Grid View* allowed us to inspect their accuracy and assess their quality. The analysis of the English-French (EN-FR) dataset showed that the dataset contains several single or two token sentences, for which the alignment will always be correct (Figure 4.14C-D). Moreover, some sentences occur more than once in the dataset, and that would affect the evaluation process since they increase recall and precision and consequently reduce AER (Figure 4.14B,C,E). The inspection showed that there are several sentences with plenty of *Possible* alignments and few or no *Sure* alignments (Figure 4.14A,F).

Comparing datasets' characteristics.

Users can see all hosted datasets on the datasets page with different statistics on the number of sentences, tokens, sure and possible alignments and their percentages. For instance, the English-French (EN-FR) dataset has significantly more possible alignments than sure alignments (Figure 4.15). This explains why all alignment models have the lowest AER on this dataset compared to all other datasets. The same applies to the Romanian-English (RO-EN) dataset since it only has sure alignments, which explains why the AER is always higher than the other datasets.

Comparing model performance with different configurations.

As an example, we compare the performance of *Itermax* with two different embedding models, namely, mBERT and a fine-tuned mBERT, to estimate the improvement achieved with the fine-tuning process. In addition to comparing all quantitative metrics, the framework allows filtering sentences where model A outperforms model B. Figure 4.12 shows that the fine-tuning enhanced the overall alignment accuracy and allowed to predict four more correct *Sure* alignments and eliminate three incorrect ones.

Comparing quantitative metrics.

The framework provides different options to compare the models performance using different quantitative metrics at corpus and sentence levels using bar chart and table views. The aggregated results in the table view (V5) reveals that the fine-tuned mBert achieved the best results in all datasets regarding AER. While *Itermax* achieved the

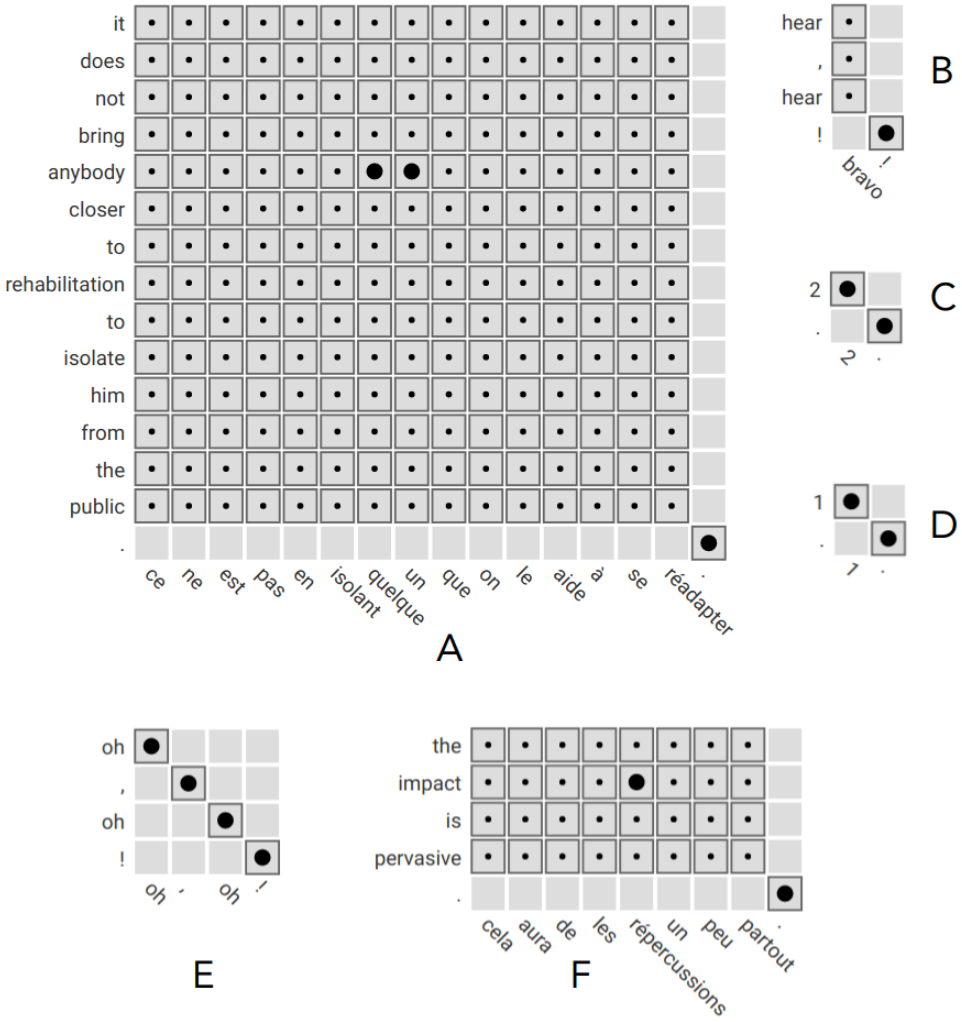


Figure 4.14: Examples from the EN-FR dataset. A) Sentence *0027* with too many possible links. B) This sentence is repeated 4 times in the datasets in sentences *0007*, *0008*, *0045*, and *0046*. C) This sentence is repeated twice in sentences *0001* and *0002*. D) Sentence *0011*, another example of short sentences with a number and a full stop. E) Short sentence with non-informative tokens repeated 3 times in sentences *0003*, *0004*, and *0005*. F) Sentence *0223*, another example of sentences with too many possible links.

best Recall on all datasets, *Argmax* with fine-tuned mBert embeddings achieved best precision on 7 datasets and second best precision on 2 datasets. Further, *Itermax* with the fine-tuned mBERT embeddings achieved the best Phrase Alignment Accuracy on all datasets. The *Match* algorithm generates more translation pairs than all other algorithms, and *Entmax* with XLM-RoBERTa embeddings generates always less translation

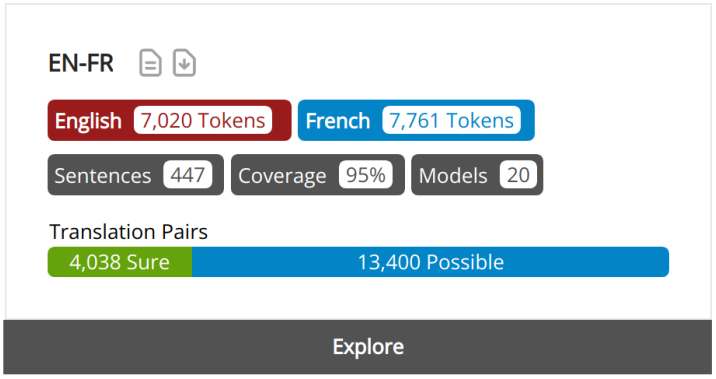


Figure 4.15: A dataset card contains all related information such as languages, number of tokens, sentences, Sure, and Possible pairs

pairs that all other algorithms.

Analyzing alignment errors.

From the pie chart provided for the *Single Dataset - Single Model* view (Figure 4.13), we can click on the red arc that represents the wrong alignment pairs to list all incorrect pairs produced by the model. Our analysis of *Itermax* with the fine-tuned mBert model on the German-English (DE-EN) dataset revealed the following:

- The most frequent wrong pairs involve a punctuation mark in one or both languages. However, such issues can be avoided by adding constraints that prevent aligning punctuation to a word.
- The majority of wrong pairs are function words, such as articles, pronouns, prepositions, and conjunctions, and most of them are semantically correct translations such as (*nicht - not*) (Figure 4.19C).
- Long sentences with repeated tokens are more likely to produce incorrect alignments even though the pairs are correct translations, but their positions in the two sentences do not correspond (Figure 4.16).
- The German-English (DE-EN) dataset contains incorrect alignments. For instance, in sentence *10*, the model generated the correct pair *präzise - precise*. However, it is classified as wrong because the gold standard aligns *präzise* with *very* and *sind* with *precise*, which is incorrect. Moreover, some sentences are not entirely aligned, and many tokens are left. For example, in sentence *40* (Figure 4.17), there are many *correct* translation pairs predicted by the model such as $Soziale - social$ and

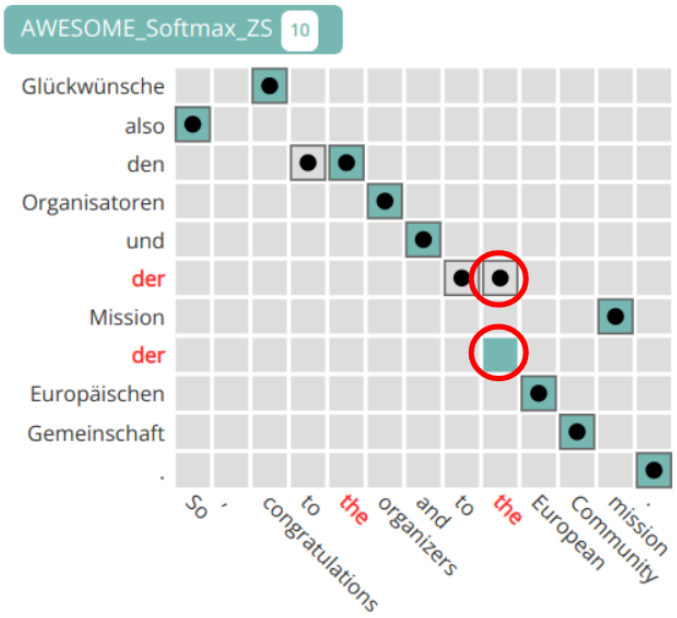


Figure 4.16: Sentence *127* from DE-EN dataset, incorrect alignment of repeated tokens.

Sicherheit – security, but they are not included in the gold standard. However, these errors are not model-specific but apply to different alignment models and datasets (Figure 4.18).

4.4 Summary and Future Work

Evaluating translation alignment models is a non-trivial task, and quantitative evaluation metrics are limited and can not reflect the actual quality of the alignment models. Therefore, qualitative evaluation is needed to understand the models' behaviour better and estimate their quality. For this purpose, we presented EvAlign to support quantitative and qualitative evaluation of alignment models by incorporating various visualization approaches and techniques.

EvAlign hosts several benchmark datasets and various alignment models. It offers different visualization views and filtering functions to help users to investigate alignment datasets and models and conduct various quality analyses. Moreover, we presented different usage scenarios that showcase the use and effectiveness of the tool.

Our analyses revealed that gold standard datasets, especially the German-English (DE-EN) and French-English (FR-EN), which have been used almost in all literature, contain plenty of errors and need to be revised and corrected by domain experts.

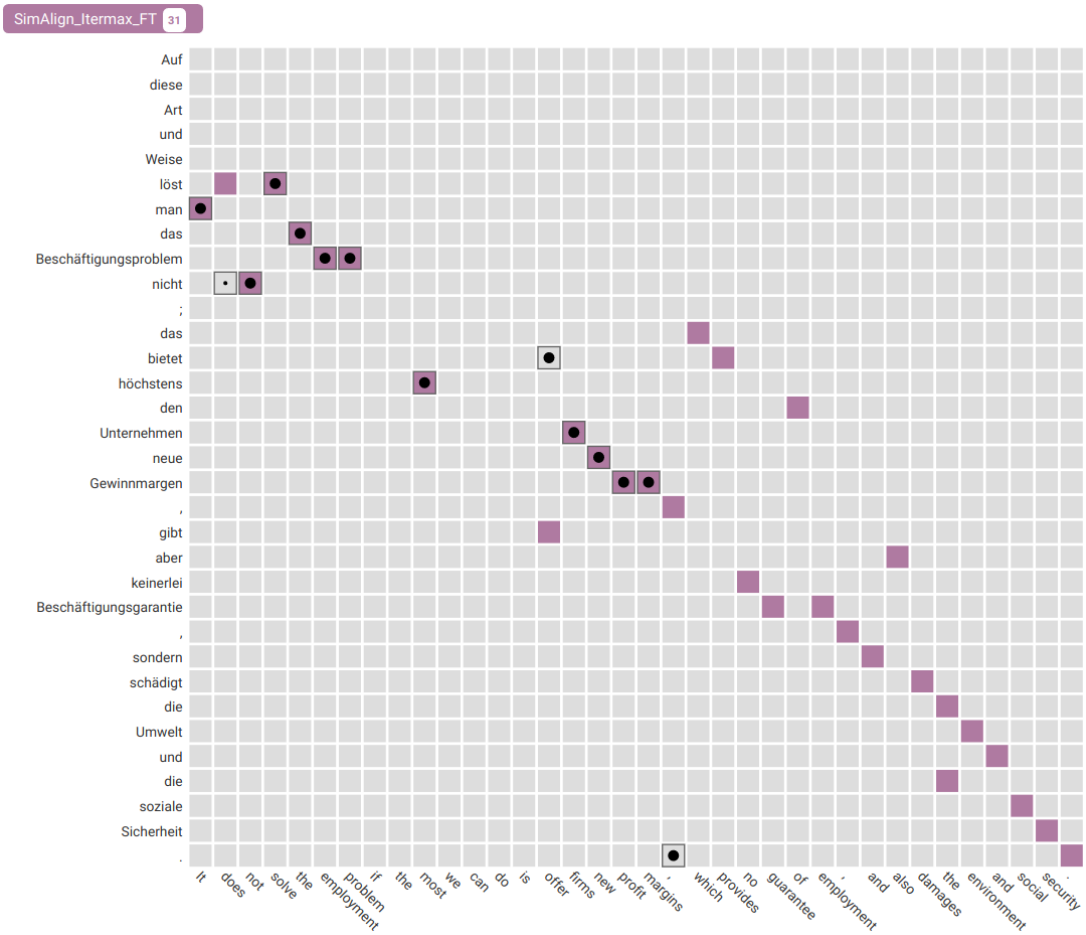


Figure 4.17: Sentence **40** from the DE-EN dataset, an example of incorrect/incomplete annotation of the gold standard sentence. The model predicts correct translation pairs but they are counted incorrect since they are not included in the gold standard.

During the development process, we encountered some difficulties that limited our efforts. For example, some papers that proposed new models do not share the source code or provide any information about the models' output. Therefore, we could not host such models on EvAlign. Similarly, not all datasets mentioned in the literature are accessible. Regarding the visualization views, the current implementation of the tool allows for comparing two alignment models at once. It is not possible now to compare more models. Also, the current implementation of the sentence-level side-by-side view visualizes only one model's alignments and does not allow for comparing two or more models. The grid view works very well with short sentences (less than 30 tokens); the readability, however, will decrease in the case of long sentences.

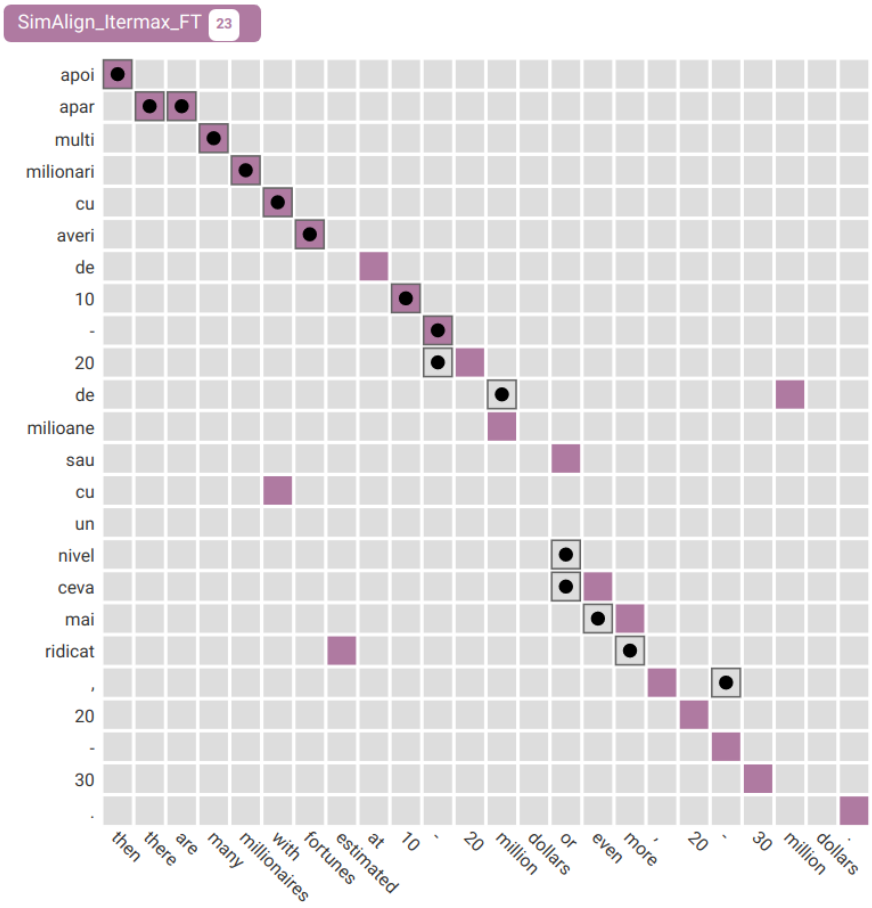


Figure 4.18: Sentence *101* in the RO-EN dataset; An example of incorrect/incomplete annotation of the gold standard sentence. The Romanian word *mollioane* is translated to *million* but the gold standard aligns the word *de* to *million* instead. Moreover, the sentence is not entirely aligned.

In future work, we aim to incorporate morphological features such as part-of-speech tags, lemma and named entities to assess model performances and classify alignment errors. Further, we will keep the tool updated by adding new datasets and models, and we encourage researchers to send us the output of their new models to publish them on EvAlign.

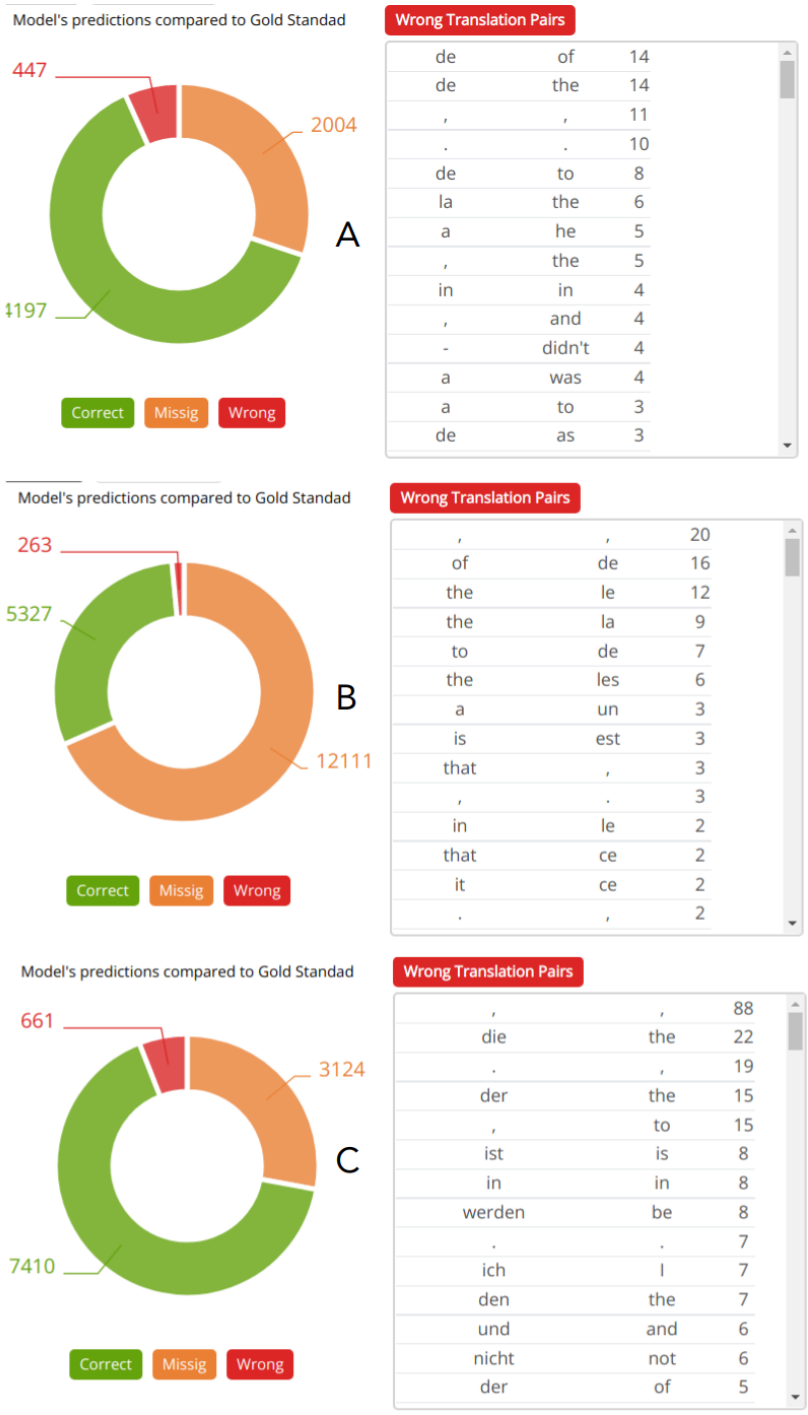


Figure 4.19: Frequent Alignment Errors, A) The alignment produced by *Argmax_XLMAlign* on RO-EN dataset. B) The alignment produced by *Argmax_XLMAlign* on EN-FR dataset. C) the side-by-side view. D) The alignment produced by *Argmax_XLMAlign* on DE-EN dataset.

Chapter 5

Applications of Translation Alignment

Translation alignment is an essential task in NLP and Digital Humanities and has been used in several applications and processing pipelines. In the following, we provide an overview of the applications where translation alignment plays a key role and report on our research on three cross-lingual applications. Annotation projection of named entity tags from English texts to the Ancient Greek parallel texts using the alignment model we developed in chapter 3. Then, we present our dynamic dictionary, a multilingual dictionary induced totally from parallel corpora of historical texts, using an accurate alignment model. Next, we show how we used translation alignment to explore and inspect the semantic transfer and semantic frameshifts in the German and English translations of two English semantic primes. Finally, we mention our plan for future work for developing a sentence-level alignment model using the word-level alignment model to support building high-quality parallel corpora aligned at sentence or paragraph levels. This chapter is based on the following publications:

- **Yousef T**, Palladino C, Jänicke S. *Transformer-Based Named Entity Recognition for Ancient Greek*. In DH 2023.

- **Yousef T**, Palladino C, Jänicke S. *Named Entity Annotation Projection Applied to Ancient Greek*. LaTeCHCLfL 2023.

- Hinrichs N, Foradi M, **Yousef T**, Hartmann E, Triesch S, Kaßel J, Pein J. *Embodied Metarepresentations*. Frontiers in neurorobotics. 2022;16.

5.1 Translation Alignment in Practice

Translation alignment has been successfully employed in a wide range of NLP tasks, of which we can mention:

Machine Translation. Translation alignment is essential in the word- and phrase-based Statistical Machine Translation (SMT) pipelines (42, 47, 54, 130, 131, 181). It has been used to extract translation equivalents from the training parallel corpus. Moreover, the quality of alignment models correlates with the translation model's performance (20, 85, 153, 233). Although Neural Machine Translation (NMT) models do not use word alignment in their pipelines, word alignment is used to analyze their outputs and assess their performance quality (28, 174, 225). Also, word alignment has been employed to filter and clean noisy parallel corpora (136, 276) and evaluate their quality.

Cross-Lingual Transfer. Translation alignment can be used to transfer linguistics annotation from one text to its translation. Various linguistics annotations can be transferred, such as Semantic Role labels (133, 183), Part-of-Speech (110, 240, 257), Named Entities tags (66, 113, 177), Relations and Arguments (81, 128, 129, 154), Semantic Parsing (217, 258), Syntactic and Dependency parsing (100, 241, 261). Further, Translation alignment has been employed to preserve and transfer structural annotations such as XML markups from the source documents into the machine-translated documents (104, 169, 237) and to conduct cross-lingual analysis and exploration of the linguistics phenomena (18, 213).

Other usage scenarios include Bilingual Lexicon Induction (BLI) from parallel texts (139, 161, 196), Word Sense Disambiguation using parallel corpora (69, 145, 155, 205, 244), creating multilingual embeddings (15, 75, 101), Grammar Induction (90, 224), Speech Recognition (80, 122), and Typological Analysis (112, 146).

5.2 Named-Entity Annotation Projection Applied to Ancient Greek

In this section, we demonstrate how we leveraged the alignment model we trained in chapter 3 to enhance the work on Named-Entity Recognition (NER) for ancient languages in general and Ancient Greek in particular. Our method uses sentence-level aligned parallel corpus of an ancient language and a modern language, for which accurate

off-the-shelf NER systems are available. We automatically annotate the text of the modern language using one of the available tools. Then we employ word alignment to find translation pairs. Finally, we project the annotations from the modern texts to the corresponding tokens in the ancient texts. Before describing our methodology, we briefly introduce the work on Named-Entity Recognition in Ancient Greek, its challenges, and available models.

5.2.1 NER & Ancient Texts

The identification and classification of names in ancient texts, especially classical ones like Homer's *Iliad*, is an essential task to support further text processing and facilitate the creation of reading environments for ancient languages (36). Typically, reader of Ancient Greek, or other ancient languages, faces the challenge of a complex and sometimes obscure language and very distant cultural references, which goes back to events and traditions that may not be immediately comprehensible to a non-specialist. Names are a substantial part of this challenge. Texts like the Bible or the *Iliad* contain hundreds, if not thousands, of names, many of which appear only once or twice. Even language specialists sometimes cannot immediately recognize certain places, people, or events. An automatic NER model for this category of texts can enormously facilitate the reader's task by extracting, classifying and linking ancient names to available resources (125), and by using those names to design applications that encourage different approaches to textual and historical exploration (26). Nevertheless, in the domain of ancient languages, NER is a challenging task. The lack of adequate infrastructure and annotated data is the main obstacle to developing reliable NER pipelines, as many of these languages have scarce (or none at all) services for annotation, lemmatization, morphosyntactic analysis, and named entity classification. A few resources of this kind have been developed, such as the Classical Language Toolkit (CLTK)[1]. However, the lack of adequately annotated texts in the original languages still makes Named Entity Recognition a challenge for many types of texts (79, 188).

5.2.2 Transformer-Based Models

The recent advances in neural networks and language modelling using transformers (245) and the availability of large pretrained language models allowed researchers to achieve state-of-the-art performance on various NLP tasks, including NER. We used transformers to train two automatic NER models for Ancient Greek. The models classify the entities into three categories, namely, Person, Location, and Miscellaneous and

[1]A Python library to perform NLP tasks on ancient languages, including NER (119)

achieved promising results on test and evaluation datasets. The models are available on Hugging Face[1].

Data

The data used for model training are collected from two resources, *The Deipnosophists* of Athenaeus of Naucratis[2], which have been annotated semi-automatically in the context of the Digital Athenaeus project[3] using INCEpTION annotation platform[4]. The entities are classified into ten categories. However, we reduced them to three main categories, PERson, LOCation, and MISCellaneous (Table 5.1). The reason behind this reduction is following the standards, where the majority of NER models use four categories (Persons, Locations, Organizations, and Misc), and since Organizations are not part of the training data, we used only three classes, However, the pipeline can be updated for more fine-grained classes such as Ethnic since it has a reasonable amount of examples in the training dataset.

Moreover, we used texts from *Odyssey* annotated manually by *Chiara Palladino*. The dataset contains 3.167 entities as PERsons and LOCations only. The main limitation of the dataset is that almost all entities are single-token entities which affects the model performance in detecting multiple-tokens entities. We also created an evaluation dataset to evaluate the models' performance. The dataset consists of 50 paragraphs randomly selected from the Book 5 of the *Histories* of Herodotus. The texts were annotated manually by Chiara Palladino using Recogito[5] and contain 351 entities.

Training & Results

In the experiments, we used two transformer-based ancient Greek language models, namely, *Ancient Greek Alignment*[6] (GRC_A) (271, 273), an XLM-RoBERTa-based multilingual model fine-tuned on the alignment task (as discussed in Chapter 3), and *Ancient Greek BERT*[7] (GRC_B) (222), a BERT-based monolingual model fine-tuned on POS tagging and morphological analysis. We used Flair framework[8] (9) to train and fine-tune the models using 75% of the data for training, 12.5% for testing, and 12.5% as

[1] https://huggingface.co/UGARIT/
[2] https://scaife.perseus.org/reader/urn:cts:greekLit:tlg0008.tlg001.perseus-grc4
[3] https://www.digitalathenaeus.org/
[4] https://inception-project.github.io/
[5] https://recogito.pelagios.org/
[6] https://huggingface.co/UGARIT/grc-alignment
[7] https://huggingface.co/pranaydeeps/Ancient-Greek-BERT
[8] https://github.com/flairNLP/

	Class	Frequency	3 Classes	Frequency
Odyssey	Person	2.469	PER	2.469
	Place	698	LOC	698
Deipnosophists	Person	12.424	PER	12.424
	Place	2.305	LOC	2.305
	Ethnic	3.548	MISC	6.735
	NoClass	2.263	MISC	
	Group	681	MISC	
	Title	206	MISC	
	Festival	20	MISC	
	Month	8	MISC	
	Language	7	MISC	
	Constellation	2	MISC	
Total				24.631

Table 5.1: An overview of the training dataset.

development dataset. We trained the models 10 epochs and used Conditional Random Field (CRF) for prediction.

The training results presented in table 5.2 show that GRC_A outperforms GRC_B regarding the F1-score on all classes. Overall, both models achieved over 82% F1 on all classes. However, their performance on PERson is significantly higher than LOCation and MISC because the training dataset contains more PERsons than other classes. The

		GRC_A			GRC_B		
		Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Training	PER	93.39%	96.33%	**94.84%**	91.24%	94.45%	92.82%
	MISC	84.69%	92.50%	**88.42%**	80.92%	83.17%	82.03%
	LOC	89.55%	77.32%	**82.99%**	86.86%	78.35%	82.38%
Evaluation	PER	90.48%	91.94%	91.20%	96.43%	87.10%	**91.53%**
	MISC	89.29%	94.34%	**91.74%**	92.00%	86.79%	89.32%
	LOC	82.69%	65.15%	72.88%	80.00%	84.85%	**82.35%**

Table 5.2: Training and evaluation results.

evaluation on *Histories*'s texts showed that GRC_B achieved higher precision and lower recall on PERsons and MISCs, but significantly higher recall and lower precision on LOCations. The qualitative evaluation revealed a lack of performance on multiple-token

entities. Nevertheless, there is still room for performance improvement by expanding the training dataset by manually annotating more texts or generating the annotations automatically with our model and then correcting them manually.

5.2.3 Cross-lingual Annotation Projection

One of the main limitations of our previous models is the nature of training data and the lack of multi-token entities. With the annotation projection we can overcome this issue and enrich training datasets with more examples. The core concept of annotation projection is to perform automatic or manual linguistics annotation on a text and project the annotation to a translation using mapping heuristics. Our proposed pipeline works for any language pair provided the parallel corpora is available and aligned at sentence or paragraph level.

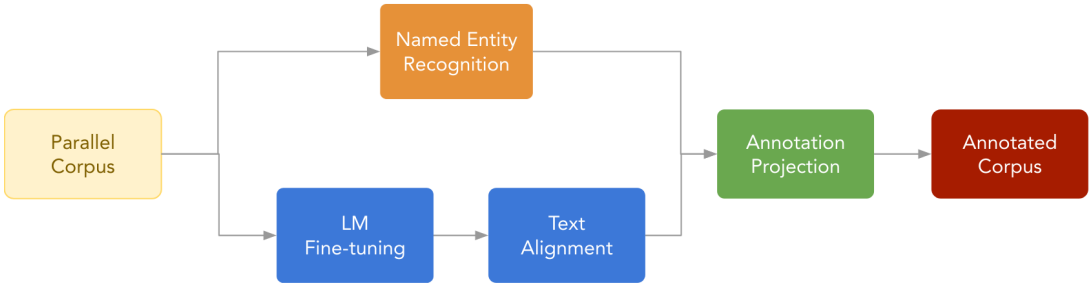


Figure 5.1: Named-Entity annotation projection pipeline.

Methodology

Figure 5.1 illustrates our proposed pipeline, it consists of three main components. We start with collecting and preparing a parallel corpus of ancient languages and at least one modern language, such as English, for which a high-quality off-the-shelf NER annotation tool is available. The corpus must be aligned at the sentences or paragraph level. Then we use a NER tool such as *spaCy*[1], *AllenNLP*[2], or *FLAIR*[3] to annotate the text of the modern language. In parallel, we use a state-of-the-art automatic alignment model to align the parallel sentences and extract the translation pairs. An unsupervised fine-tuning using the parallel corpus can be employed using different training objectives discussed in Chapter 3 in order to improve the word alignment

[1]https://spacy.io/
[2]https://allenai.org/allennlp
[3]https://github.com/flairNLP/flair

accuracy. Subsequently, we find the corresponding tokens of the detected named entities and project the annotation using a direct mapping heuristic. In our experiment, we used the Bible in Ancient Greek, Latin, Arabic, and English.
Figure 5.2 shows an example of the processing pipeline, the text used in the example is the English[1] and Ancient Greek translations of *John 1-6*, annotated with AllenNLP, and aligned with our alignment model.

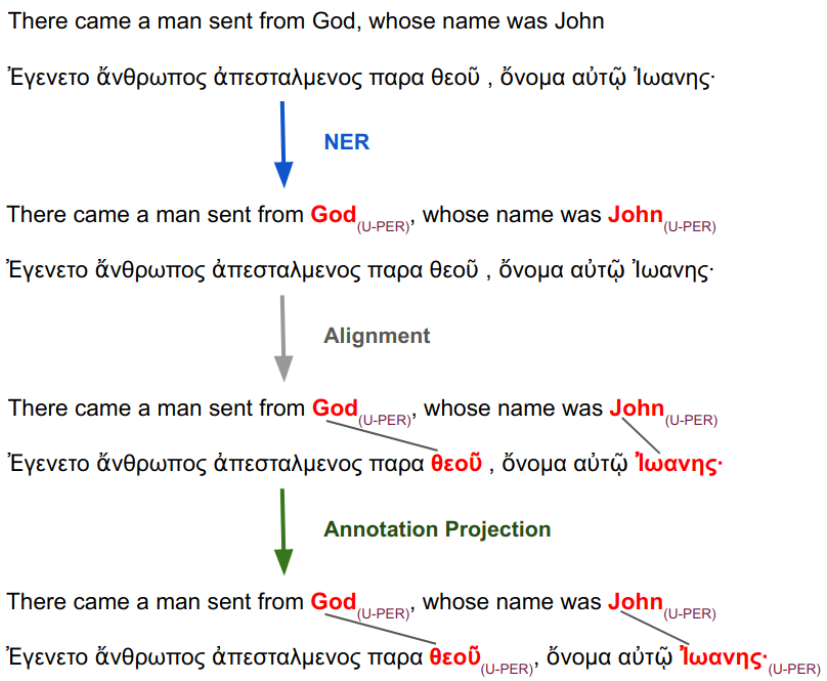


Figure 5.2: An example of the annotation projection using the proposed pipeline.

Corpus Collection The Bible is an ideal source of parallel texts and is available in several languages, including ancient ones. The corpus includes 31,102 verses (23,145 verses in the Old Testament and 7,957 in the New Testament). Further, the corpus is aligned at the verse level thanks to its hierarchical structure (Book/Chapter/Verse), which allows for producing accurate alignments at the word level. It is also rich in named entities, especially persons and locations. Nevertheless, the Bible corpus has its limitations regarding the language style and text diversity. We used the Bible Corpus repository[2] to build our parallel corpus. The repository includes translations of the Bible in over 100 languages (55).

[1] King James Version translation of the new testament.
[2] `https://github.com/christos-c/bible-corpus`

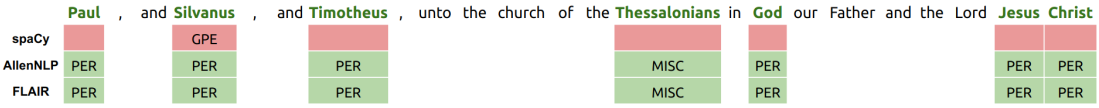


Figure 5.3: A performance comparison between three STOA NER models on biblical text (1 Thessalonians 1:1).

In our experiment, we used Ancient Greek, Latin, and Arabic with the KJV English translation. Every verse has a unique ID that encapsulates the book, chapter, and verse information. And this ID is unified among all translations. The ancient Greek translation was unavailable in the repository; therefore, we collected it from the Perseus Digital Library[1]. We followed the same naming convention to assign verse IDs.

Automatic NER Tagging Recently, tremendous progress has been made in the field of NER tagging with the advent of transformer models and the availability of training datasets of adequate size. Several NER Tagging systems were developed for many languages, especially modern European languages, and achieved high accuracy. However, these models are trained on modern texts, and their performance varies when annotating historical texts, such as the Bible. Therefore, we benchmarked three state-of-the-art NER tagging tools for English, *SpaCy*, *AllenNLP*, and *flairNLP*, to select the best model that delivers the highest accuracy on the biblical text. The comparison revealed that AllenNLP and flairNLP significantly outperformed spaCy, and their performance was so close (Figure 5.3). Therefore, we used AllenNLP NER tagger with four entity classes (PERS, LOC, ORG, MISC).

Automatic Translation Alignment. We used the alignment model discussed in Chapter 3, the model achieved good performance on Ancient Greek-English parallel texts. Supervised and unsupervised training can be employed to enhance the alignment accuracy using different training objectives discussed in Chapter 3. However, our alignment model is already trained on the Bible texts, therefore there is no need for further fine-tuning.

Annotation Projection The basic premise from which we start is that Named entities are informative components of any text and contribute to its meaning; therefore, a good translation should preserve the named entities of the original text and their relations. In order to project the annotations from the source to target text, we followed a simple mapping function:

[1]https://scaife.perseus.org/library/urn:cts:greekLit:tlg0031/

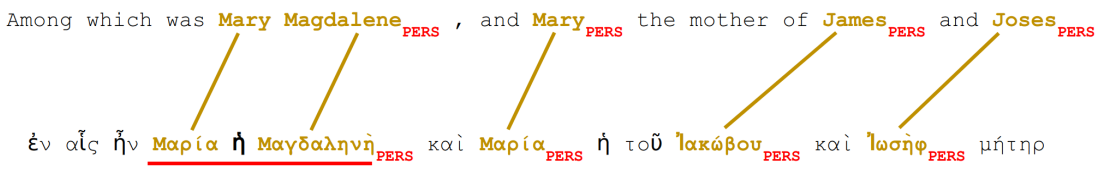


Figure 5.4: Annotation projection example.

- When the entity is a single- or multi-token entity and is aligned to a single token (one-to-one or many-to-one alignment). In this case, we assign the corresponding token in the target text the same category as the source entity. For instance, *Mary-Μαρία*, *James-Ἰακώβου*, and *Joses-Ἰωσὴφ* in Figure 5.4.
- When the entity in the source text is single or multi-token and is aligned to multiple tokens (one-to-many or many-to-many alignment), in this case, if the corresponding tokens are consecutive, they will be considered as one multi-token entity and get the same category. Otherwise, we annotate the range of tokens from the first to the last corresponding tokens as one entity and annotate it with the same category as in the source text. For instance, *Mary Magdalene* and *Μαρία ἡ Μαγδαληνὴ* in Figure 5.4.
- If the entity has no correspondence in the target language (one-to-null, many-to-null), then no projection is required.

Evaluation

We performed qualitative evaluation to estimate the quality of the produced annotations on three language pairs: English-Ancient Greek, English-Latin, and English-Arabic. Two domain experts[1] manually assessed 100 random verses, which corresponded to about 550 extracted entities per language, and assigned a score to each detected named entities as follows:

(1) - correct alignment and correct English NER annotation.
(2) - correct alignment but incorrect English NER annotation.
(3) - correct English NER annotation but incorrect alignment.
(4) - incorrect English NER annotation and incorrect alignment.

Table 5.3 summarizes the evaluation results: the automatic NER annotation achieved over 94% accuracy and the alignment on Ancient Greek-English achieved the highest

[1]Chiara Palladino inspected the English-Ancient Greek and -Latin datasets and Tariq Yousef inspected the English-Arabic dataset.

Score	Ancient Greek	Latin	Arabic
Correct alignment / Correct NER	86.63%	82.34%	75.54%
Incorrect alignment / Correct NER	7.26%	12.87%	21.16%
Correct alignment / Incorrect NER	5.28%	3.96%	2.98%
Incorrect alignment / Incorrect NER	0.83%	0.83%	0.33%

Table 5.3: Manual evaluation of 100 randomly selected verses.

accuracy, since the alignment model is optimized for this language pair. However, the errors in misclassification were common in the three language pairs for personal names classified as Locations and vice versa. Further, Sometimes a Greek, Latin, or Arabic noun would be misclassified as a consequence of the English translation, which adopted a different type of entity: many ethnonyms, which would be classified as MISC in our dataset, were translated in English as location names, and therefore classified as LOC. Additionally, incomplete or partial alignments were frequent in the dataset (9 cases in Ancient Greek, 28 in Latin, and 33 in Arabic) especially in multi-token entities such as "Jesus Christ", "Simon Zelotes", or "Pontius Pilate".

Summary

In this experiment, we used translation alignment to project named entities annotation from texts in modern languages to texts in ancient languages in order to create NER datasets for such languages, enrich the available datasets, or annotate texts where the existing NER models fail to create accurate annotations.

It is worth noting that the proposed method can work with any parallel corpus, not only the bible. However, many factors might affect the annotation performance, such as the translation quality, text genre, and performance of the NER system of the modern language used in the parallel corpus. Also, the proposed method can be applied to low-resourced modern languages to enrich the annotated NER training dataset. The automatic alignment accuracy varies between language pairs; It is not surprising that the English-Ancient Greek alignments are better than English-Arabic or English-Latin since the language model is mainly fine-tuned on Ancient Greek texts.

Finally, The pipeline employs two automatic components, and getting accurate results is subject to the performance of the two components and their success in annotating and aligning the texts. However, the workflow depends, in the first place, on the accuracy of the English or modern language's NER tagger because if it does not detect the entity, it will not be projected. Replacing one or both automatic components with manual annotation or alignment would significantly enhance performance.

5.3 Bilingual Lexica Induction from Parallel Corpora

In the field of bilingual dictionaries induction, we distinguish between two main paradigms: (1) Mapping two monolingual embedding spaces using unsupervised (60, 149) or weakly-supervised (161, 196) approaches with seed lexicon and linear projection for instance. (2) Using word alignment to extract translation pairs from parallel corpora aligned at sentence or paragraph level (112, 218, 229).

In this experiment, we use the alignment workflow and the trained language model discussed in Chapter 3 to align a parallel corpus of historical languages. The tool can be a valuable resource for scholars and new language learners; it aids them in exploring new vocabularies and discovering the meaning of words in various languages. The proposed workflow is a proof of concept, and the limited computation resources prevented us from extending the corpus with more parallel texts. However, integrating more corpora into the existing database is straightforward. Also, integrating bilingual dictionary entries or accurate manual alignments is possible thanks to the underlying graph data structure, which facilitates modelling different relation types among tokens. Our contributions can be summarized in the following points:

- We apply the automatic dictionary induction to historical low-resourced languages and produce high-quality translation pairs even for languages that do not share direct dictionaries or parallel texts, thanks to the alignment workflow and the trained language model.

- We model the obtained translation pairs in a graph data structure, which allows us to integrate manual alignments and dictionary entries and facilitate performing clustering or pivoting to generate translation pairs of languages with no direct connections.

Corpora. Table 5.4 provides an overview of the parallel corpora we used. A total of 371,000 sentences in 6 languages, namely, Ancient Greek, Arabic, English, Hebrew, Latin, and Persian. The main sources of the parallel texts is the Bible, Perseus Digital Library, and the DFHG corpus.

Alignment. we used the alignment workflow presented in Chapter 3 and the fine-tuned language model. We aimed to get accurate translation pairs than getting more pairs with less precision. Therefore, we used *Entmax* for extracting translation pairs since it achieved high precision over the three gold standard datasets.

Corpus	Languages	# Sentences	# TPs (Unique)
DFHG	Ancient Greek-Latin	8,000	221,165 (113,849)
Bible	Ancient Greek-Latin	7,941	106,303 (31,185)
	Ancient Greek-English	7,955	116,554 (31,281)
	Ancient Greek-Arabic	7,955	90,121 (40,107)
	Ancient Greek-Hebrew	7,955	87,992 (42,663)
	Ancient Greek-Persian	7,922	99,154 (41,317)
	Arabic-Latin	30,288	348,106 (139,480)
	Arabic-English	31,101	386,889 (112,927)
	Arabic-Hebrew	31,102	347,289 (127,899)
	Arabic-Persian	31,102	364,536 (144,933)
	Hebrew-Latin	30,288	334,509 (142,091)
	Hebrew-English	31,101	364,768 (118,011)
	Hebrew-Persian	31,020	340,905 (150,892)
	Latin-English	30,287	434,218 (99,875)
	Persian-English	31,019	479,145 (102,870)
	Persian-Latin	30,214	399,361 (139,863)
Odysses Xenophon Iliad	Ancient Greek-English	16,500	196,909 (67,944)
Total		398,268	5,058,829 (1,619,752)

Table 5.4: Overview of the processed parallel corpora.

Graph Generation Figure 5.5 illustrates the proposed graph structure. We model every translation pair as two nodes connected with an edge. Additional relations can be added to indicate different linguistic features if they are available. For example, connecting a phrase with its constituent words or linking a word with its lemma. These relations can be beneficial for running sophisticated queries. Shi et al. (218) proposed a matching ratio that considers the alignment frequency and how frequently the two words co-occurred in the corpus. However, this ratio works only with one-to-one alignments. Therefore we proposed an alignment score that considers phrases as well:

$$score(s,t) = \frac{2 * aligned(s,t)}{aligned(s|L_t) + aligned(t|L_s)} \tag{5.1}$$

Where $aligned(s,t)$ indicates how many times the two words/phrases are aligned together, $aligned(s|L_t)$ indicates how many times s is aligned in total to words/phrases in the same language as t, and $aligned(t|L_s)$ indicates how many times t is aligned in total to words/phrases in the same language as s.

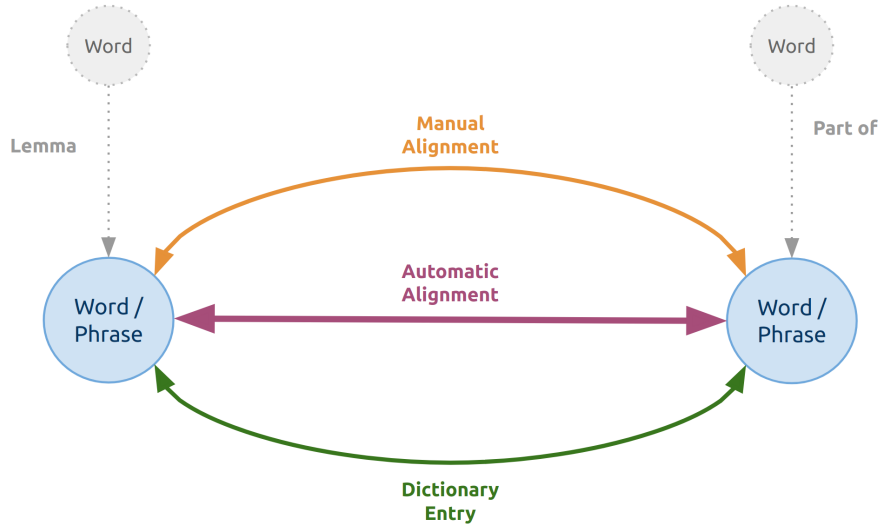


Figure 5.5: The graph structure of the induced translation pairs.

The resulting graph contains over 614k nodes and 1,620k edges from the type "*Automatic Alignment*". Additional 193k translation pairs collected from Ugarit can be added as *Manual Alignment* (see Chapter 2, section 2.5.3). Moreover, Graph clustering algorithms such as CHINESE WHISPER (35), a hard partitioning and flat clustering algorithm, can be applied to cluster graph entries into sets containing words/phrases that are semantically related or share the same meaning.

Figure 5.6 shows a cluster of aligned words/phrases in various languages. This cluster is one of 7300 clusters obtained after filtering out the relations with frequency less than 5, alignment score less than 0.25, and running *Chinese Whisper* clustering algorithm for 20 iterations. Figure 5.7 shows another cluster with an extended visualization, in which manual alignments and PART_OF relations are displayed.

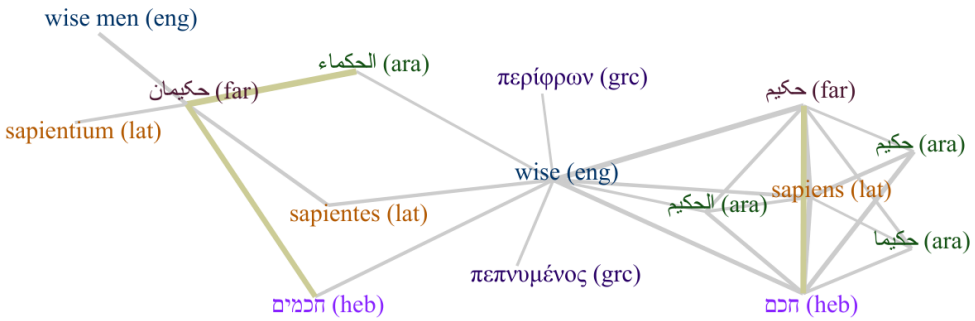


Figure 5.6: An example from graph clustering results.

The induced bilingual dictionaries are published on Github [1] in addition to the results of clustering algorithms with different parameterizations.

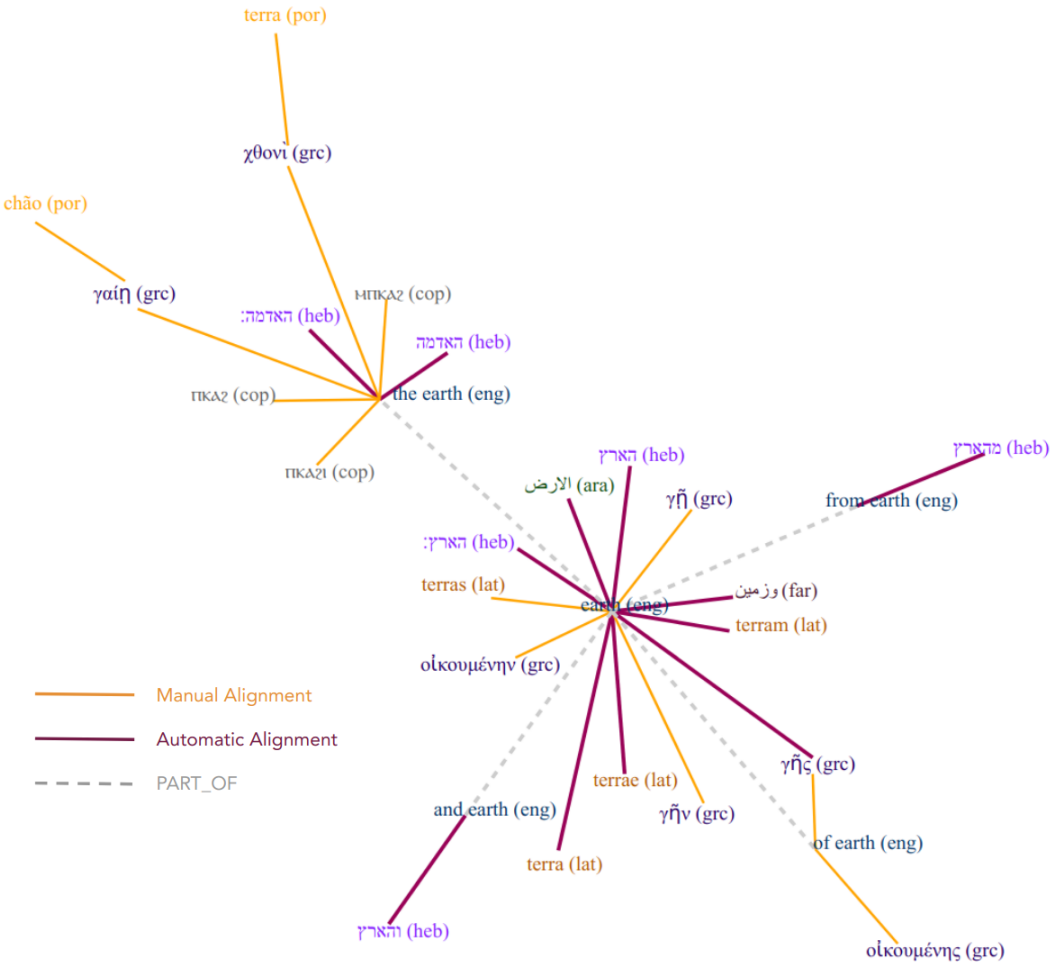


Figure 5.7: An example from graph clustering results with extended relations.

5.4 Evaluating Cross-Lingual Semantic Transfer

In this section, we address another use case where we employed translation alignment to investigate a cross-lingual phenomenon. This study is performed on modern languages, but the proposed pipeline can be employed for historical languages as well.

Our exploratory study proposes a novel approach to investigate the properties of the

[1] https://github.com/TariqYousef/Bilingual-Lexica-Induction-from-Parallel-Corpora

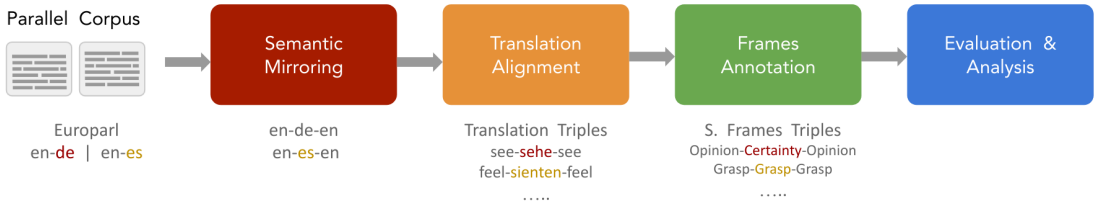


Figure 5.8: The proposed pipeline for exploring cross-lingual semantic frameshifts.

cross-lingual transfer of semantic frames based on semantic primes[1]. Our study focuses on two language pairs, English-German and English-Spanish. The study employs *Semantic Mirroring*, by translating the translations back to the original language, to consider the variations of frames evoked by original, translated, and back-translated sentences as clusters of meaning distinctions. This study has been conducted in collaboration with experts in Translation Studies, Citizen Science, and Cognitive Neuroscience. My contributions can be summarized as follows: (1) co-designing the proposed pipeline and implementing it, (2) performing the alignment, fine-tuning the language model to enhance the alignment accuracy, (3) aligning the semantic frames across translations, prepare the results for quantitative evaluation, and design the visualization views to facilitate the manual analysis and qualitative evaluation.

Methodology

Figure 5.8 illustrates our proposed pipeline. The pipeline employs three state-of-the-art automatic neural models: machine translation, automatic word-level alignment, and automatic semantic frames annotation. The parallel texts we used in this study originate from *EuroParl* Corpus (130). The corpus contains accurate translations performed by professional human translators and is aligned at sentence level, facilitating the alignment at word level. In this study, we focused on two semantic primes, namely the verbs *See* and *Feel* (and their inflected forms) since they are used to express sensations and opinions, in addition to the physical act of seeing through the eyes.

We filtered the corpus for sentences with English as the speaker's language (source language) and contains one of the semantic primes *See* and *Feel* or their inflected forms. Further, we collected the German and Spanish translations of the filtered subset. Then,

[1]According to Wikipedia, Semantic primes are a set of semantic concepts that are argued to be innately understood by all people but impossible to express in simpler terms. They are proposed by Wierzbicka (255), who suggested a list of primitives consists of 63 semantic primes, of which 13 are verbs.

we used *DeepL*[1], a state-of-the-art neural machine translation system, to re-translate the German and Spanish sentences to English to generate the translation mirrors. For word alignment, we used the same alignment workflow discussed in Chapter 3, and we fine-tuned XLM-RoBERTa on the the parallel corpus we collected and used it as embeddings source, and to ensure high precision, we used *Entmax* approach to extract the translation correspondences from the similarity matrix as proposed by (74).
In some sentences, the verbs, or their inflected forms, had no one-to-one correspondence or had no correspondence at all; Thus, the data had to be further sorted by filtering out these sentences. We evaluated the performance of the automatic alignment by randomly selecting 100 sentences for each language pair and examining the alignment manually. The evaluation indicated an alignment accuracy of 92% for English-German-English and 97% for English-Spanish-English.

Figure 5.9: Examples of word aligned and semantic frame tagged sentences.

Next, we performed an automatic semantic frame annotation for the verbs *See* and *Feel* and their translation correspondences in the German and Spanish translations and the English translation mirrors. For this purpose, we used the Large Ontology Multilingual Extraction (LOME)[2] (260), a state-of-the-art tool for accurate multilingual FrameNet parsing. We also evaluated LOME performance manually by inspecting the evoked frames in 100 parallel sentences, and the overall performance was 91%. Figure 5.9 illustrates the results of the word alignment and semantic frame annotation processes described in this section for the language pair English-German. As the Figure shows, the frames can stay consistent (A), or a frameshift after each translation iteration can be observed (B).

[1] `https://www.deepl.com`
[2] https://nlp.jhu.edu/demos/lome/

Frameshifts	Feel (DE)	See (DE)	Feel (ES)	See (ES)
x–x–x	39.42%	40.70%	47.69%	47.24%
x–x–y	7.31%	4.39%	6.25%	5.34%
x–y–x	21.92%	19.15%	12.96%	21.21%
x–y–y	17.88%	21.97%	27.08%	15.82%
x–y–z	13.46%	13.79%	6.02%	10.39%
x–x–*	46.73%	45.09%	53.94%	52.57%
*–x–x	57.31%	62.67%	74.77%	63.06%

Table 5.5: Experiment Results. x-x-y indicates the frames in source (x), translation (x), and translation mirror (y). An asterisk (*) means *any frame*.

Results

To facilitate detecting frameshifts and analyzing them, we visualized frames evoked by the verbs *Feel* and *See* by aggregating the aligned frames in each parallel sentence for each language pair. Then, we used *Sankey* diagrams to show the aggregated triples.

Figure 5.10 shows the most frequent patterns of frameshifts for the two verbs in the two language pairs. Table 5.5 also reports on the aggregated results of frameshifts. As we see, the semantic transfer stability in English-Spanish-English is higher than in English-German-English for both verbs. Another interesting observation is that machine translation (*-x-x) achieved higher semantic frame stability than human translators (x-x-*) for both verbs and in both languages. Again, the stability of semantic transfer between English and Spanish is significantly higher than between English and German in both human and artificial translations.

Nevertheless, it is worth noting that the reported results are acquired by an automatic pipeline that utilizes three automatic neural models with high accuracy; this accuracy will never reach the accuracy of human annotation, as the manual evaluation revealed. Further, there is no measure to compute the amount of semantic shift. Such a metric to quantify the semantic shift would be beneficial since some semantic frames are similar or share a similar meaning, and can almost be evoked by the verbs in the same context, which explains some automatic annotation errors. A detailed qualitative analysis of the obtained results from the perspective of cognitive neuroscience and translatology have been performed by domain experts and is reported in our published paper (105).

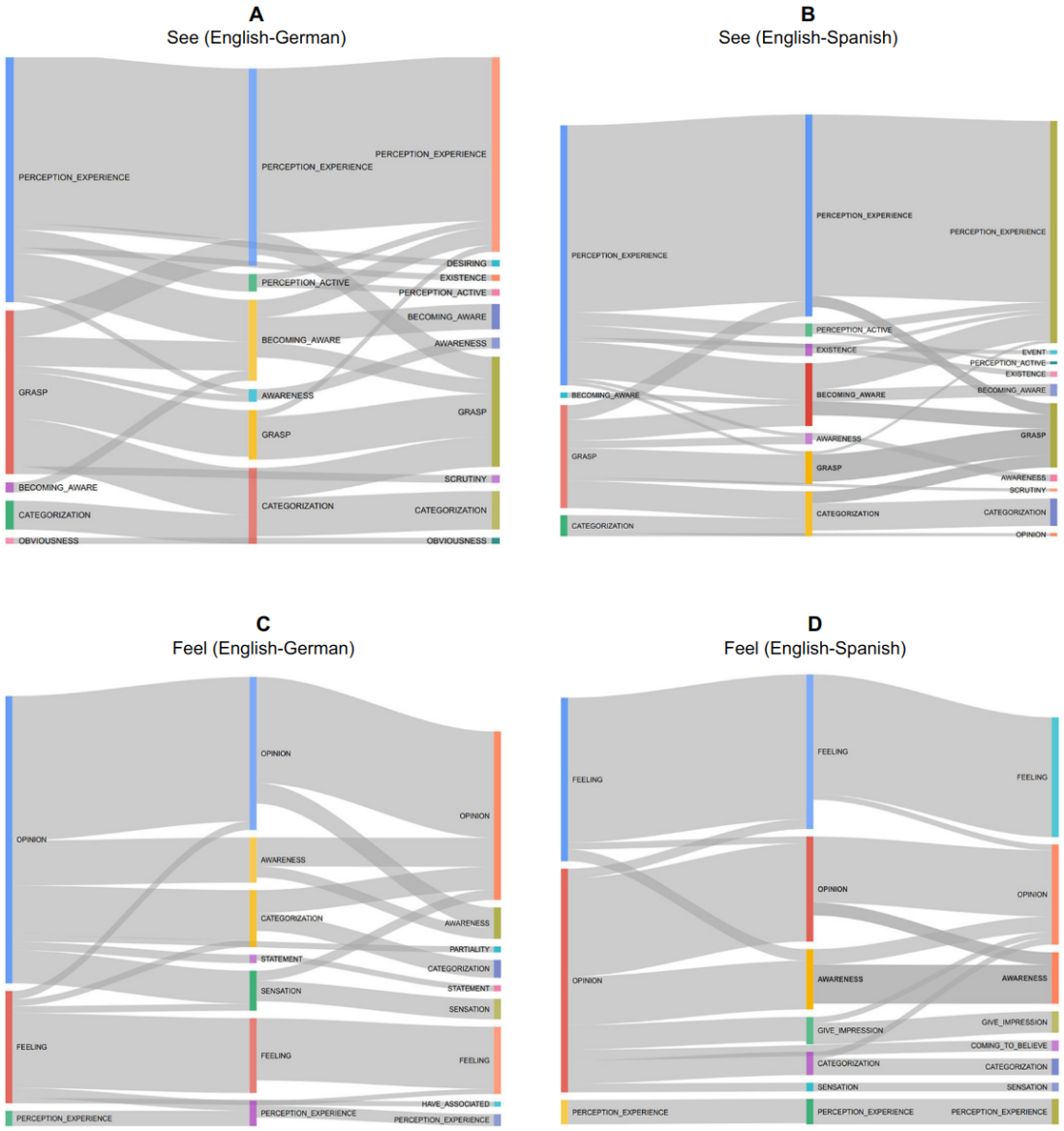


Figure 5.10: Frameshifts of the semantic primes *See* and *Feel*.

5.5 Future Work

The availability of parallel corpora and accurate alignment models facilitates research in different cross-lingual and multilingual experiments. In addition to the three cross-lingual applications we discussed, we intend to leverage the word-level alignment model for historical language to develop an accurate sentence-level alignment model. This model can be employed to segment long parallel documents into smaller aligned units

such as paragraphs or sentences. With the availability of vast amounts of OCR-ed books and manuscripts and their translations, the suggested aligner will support the efforts to create high-quality parallel corpora for historical and low-resourced languages. And subsequently enhance the development of various cross-lingual applications for these languages.

Our starting premise is that *two parallel sentences would have more translation pairs in common than non-parallel sentences*. The planned pipeline will segment the two documents into smaller parts using a sentence segmentation heuristic and then calculate the similarity between every two sentences in the two documents by aligning the sentences and counting the number of translation equivalents they share. Possible enhancement can be achieved by filtering out stop words since they appear in all sentences and are the most frequent source of alignment errors, as we discovered in chapter 4. Additional similarity criteria would be to compute the semantic similarity between the two sentences using sentence embeddings and cosine similarity. The similarity can also be normalized, and the sentence length can also be considered. Finally, there are two possibilities to extract the alignments, training a binary classifier to predict whether the two sentences are translations to each other or not. Or employing an alignment extraction heuristic to extract the sentence pairs from the two-dimensional similarity matrix. We intend to implement the described pipeline and evaluate different parameterizations and criteria in the near future.

Chapter 6

Text Alignment Visualization

Visualization plays an essential role in facilitating research in NLP and Digital Humanities. It assists developers and practitioners in analyzing the developed models and their behaviors. Text visual analytics have been used with various NLP tasks, such as text summarization (151, 235, 280), question answering (143), topic modeling (57, 221), sentiment analysis(141), text classification (207), machine translation (96, 227), and translation alignment (112, 226). Visualization can be applied at different development stages of NLP tasks: *exploring* training and benchmark datasets (235, 236), *comparing* and *analyzing* NLP models and *inspecting* their behaviours and parameters (231, 236), and visualizing the models output (96, 227) for *performance evaluation* and *error analysis*.

In the field of translation alignment, employing sophisticated visualization approaches to present alignments is essential as they help analyze translation equivalents, explore their usage contexts, and disambiguate word senses. The first visualization of aligned texts was introduced by the *Loeb Classical Library*,[1], which published important works of ancient Greek and Latin literature in bilingual format. The original texts were placed on the left-hand page, whereas the literal translations were placed on the right-hand page. In this chapter, we review visualization techniques used for monolingual and cross-lingual text alignment, introduce our work on the Ugarit, EvAlign, Parallel Corpus Visualizer, Summary Explorer, and present our tailored visual solutions for different alignment-related tasks. This chapter is based on the following publications:

- **Yousef T**, Janicke S. *A Survey of Text Alignment Visualization.* IEEE transactions on visualization and computer graphics. 2020;27(2):1149-59. (267)

[1] `https://www.loebclassics.com/`

- **Yousef T***[1], Syed S*, Al Khatib K, Jänicke S, Potthast M. Summary Explorer: Visualizing the State of the Art in Text Summarization. InProceedings of the 2021 Conference on EMNLP: System Demonstrations 2021 Nov (pp. 185-194).

6.1 Survey of Text Alignment Visualization

Alignment in its basic form is the process of finding correspondences between two data objects, and visualization supports the analysis of alignment patterns. In this survey, we consider three text alignment scenarios:

The *first scenario* for text alignment refers to the collation task in textual criticism aiming to investigate textual variation across different versions of a text (256).
The *second scenario* is concerned with the discovery and analysis of re-used text passages within a collection of texts (76). This oral or written reproduction of textual content is called *text re-use* (43). Deliberate text re-use appears in the form of direct quotes and phrases like winged words and wisdom sayings. A prominent application of text alignment in this context is plagiarism detection (84), for which a text with unacknowledged, re-used passages is compared with a reference documents database.

The *third scenario* is translation alignment. Text fragments are aligned with their translations at the word, sentence, or paragraph level.

Though the driving forces for developing solutions for the three scenarios are different, the methodological approaches to discover and analyze text alignments partially overlap. Text alignment visualizations are indispensable in all scenarios for targeted users who typically have a background in non-technical domains like the humanities, social and political sciences to make sense of automatically derived patterns. We reviewed over 40 visual interfaces that support text alignment analysis aiming to:

- discuss the state-of-the-art visualization strategies for text alignment.
- highlight the uniformity of them across several disciplines.
- reveal the current redundancy of solutions.

Drawing from the body of related works, we extract design implications to guide future developments for text alignment visualization and open challenges relating to insufficiently supported research tasks.

[1]* Equal contribution.

6.1.1 Text Alignment Visualizations

In the following, we review the different visualization techniques supporting the three text alignment scenarios, describe their visual representations, list their possible applications, and discuss their strengths and limitations. Table 6.1 provides a classification of visualization techniques embodied in the 40 alignment tools we surveyed.

6.1.1.1 Sequence-aligned Heat Maps

Sequence-aligned heat maps serve the purpose of conveying a summarized overview of alignment patterns among different source texts, and they are typically used to navigate to text units of interest.

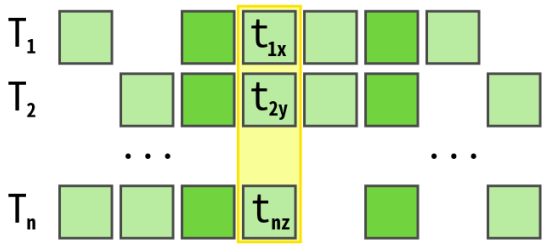


Figure 6.1: Visual representation of sequence-aligned heat maps.

Visual Representation Figure 6.1 illustrates the visual representation of sequence-aligned heat maps. Each text T_i from the corpus $T_1, \ldots, T_n$ is represented by a sequence of colored rectangles. A rectangle stands for a text unit, which is typically a sentence, paragraph, chapter, or even a larger text fragment. The text units are ordered as they appear in the text, and aligned text units are placed on the same vertical axis. Vertical sections with only one cell indicate that for the corresponding text unit, no match among the other texts exists. In contrast, a gap in the sequence i denotes that the corresponding text T_i does not contain a variant to matching text units of other texts. Sequence-aligned heat maps are usually combined with other text alignment visualization to show details on demand, i.e., that allows for close reading of the compared texts. The grid size plays an important role; the more text units and the higher the variation, the less space can be reserved for a vertical section. Therefore, they are usually accompanied by zooming and scrolling functions to enhance readability and facilitate navigation.

Task	Tool	Sequence Heat Map	Grid Heat Map	Aligned Barcodes	Text Heat Map	Side-by -side View	Tabular View	Variant Graph
Collation	TexTile (19)	✓						
	ViTA (6)		✓	✓				
	NMerge (25)							✓
	TRAVIZ (118)							✓
	DRVG (115)	✓						✓
	Storylines (7)							✓
	WordGraph (210)							✓
	Word Tree (250)							✓
	Stemmaweb (16)						✓	✓
	CollateX (67)						✓	✓
	PyCoviz (178)						✓	
	eComparatio (251)					✓	✓	
	iAligner (269)	✓				✓	✓	
	LERA (215)	✓				✓		
	CATView (203)	✓				✓		
	JuxtaCommon (121)				✓	✓		
	EVT (211)				✓	✓		
	Versioning Machine (214)					✓		
	ShakerVis (93)					✓		
	Variance Viewer(5)					✓		
	Iteal (116)			✓		✓		✓
	Baumann et al. (29)			✓		✓		
	Alignment maps (50)			✓		✓		
Text Reuse	PicaPica (22)			✓		✓		
	Text Re-Use Browser (265)		✓	✓		✓		
	Kitab Project (2)			✓		✓		
	GuttenPlag (21)			✓	✓			
	Chinese Text (232)				✓			
	Text Re-Use Grid (265)		✓					
	PatternGramm (208)	✓						
	News Auditor (30)					✓		
Translation	Ugarit (65)					✓		
	Alpheios (13)					✓	✓	
	Biblehub (13)					✓	✓	
	Japanese-Chinese (56)		✓					
	Linguee (3)					✓		
	Glosbe (1)					✓		
	Reverso Context (4)					✓		
	Divan Hafez(266)					✓		
	Ducat (223)					✓		
	Translation Arrays(49)					✓		

Table 6.1: Surveyed Tools & Visualization Techniques

Applications Sequence-aligned heat maps are often applied to navigate through different text variants to support the *collation* task (88, 203). The framework LERA (215) applies a coloring scheme to indicate different degrees of variation; the lighter the color

of a rectangle, the more similar the corresponding segment is to the other parallel segments (Figure 6.11). Two works employ the sequence-aligned heat map to different text hierarchy levels (19, 115). For example, TexTile allows users to dynamically inspect variation among sequences generated for the entire text-, page-, line- and word-level, as shown in Figure 6.2. A five-level univariate color scheme is applied to reflect the similarity or distance between a variant and its parallel segment in the reference text. Next to collation, sequence-aligned heat maps have also been used to visualize detected plagiarised text passages among a base text and a set of documents using the Composite Categorical Patterngram (CCP) (208). Different colors are used to distinguish between base file (blue dots), frequent patterns (green), and infrequent matched patterns (red).

Strengths & Weaknesses Being a textual abstraction, sequence-aligned heat maps make alignment patterns on word, sentence, and paragraph levels perceivable. However, they need to be combined with a text-based representation to read aligned fragments. The visual representation works well for less-varying texts; however, transposed fragments increase the number of columns and white space.

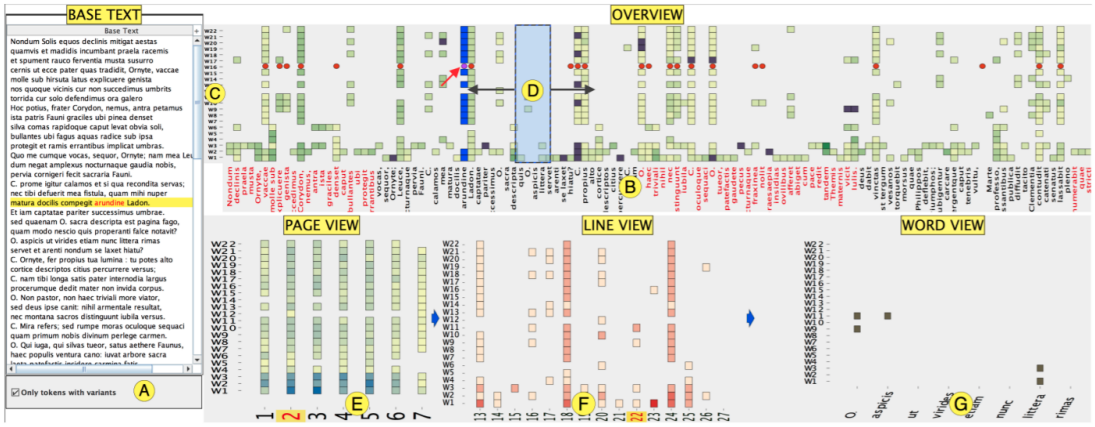


Figure 6.2: TexTile variation visualization at different text hierarchy levels (19).

6.1.1.2 Grid-based Heat Maps

Grid-based heat maps illustrate alignment patterns among two texts of a corpus, which allows drawing conclusions on the type of textual variation among them.

Visual Representation Each cell of the two-dimensional grid represents a juxtaposition of two text units t_{ix} and t_{jy} of two arbitrary texts T_i and T_j of the text corpus

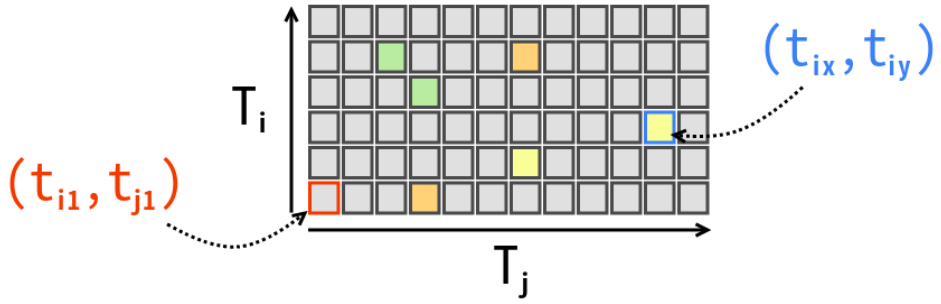


Figure 6.3: Visual representation of grid-based heat maps.

$T_1, \ldots, T_n$ (Figure 6.3). Cells might be evenly sized, but they can also visually reflect the lengths of the corresponding text units, e.g., by rectangular cells. Typically, color is used to inform on particular features of the relationship between t_{ix} and t_{jy}, e.g., alignment types or matching scores. Typically, clicking a cell opens a closer look at the juxtaposed text units.

Applications Grid-based heat maps are applied to different text hierarchy levels. In two applications, a cell, or a dot, represents a match among two text variants on word level, so that diagonal patterns illustrate similarity in text flow (6, 56). The Text Re-use Browser (265) uses a dot plot matrix similarly to highlight repetitive or systematic text re-use patterns. A dot represents a pair of similar sentences, and the dot's color reflects the matching or similarity score. The Text Re-use Grid (265) provides alignment information for an entire document collection. A cell stands for a pair of processed documents, and the color indicates the frequency and type of occurring text re-use patterns. Figure 6.4 provides examples for both Text Re-use Browser and Grid. EvAlign and Ugarit use the grid view to visualize manual and automatic translation alignment at word level.

Strengths & Weaknesses A limitation of this technique is that only two text sets can be chosen for comparison. However, the design is applicable to all text hierarchy levels—even alignments on document level are supported. The resultant matrix typically contains a large amount of white space and redundant information. However, the arrangement of glyphs makes alignment patterns salient.

6.1.1.3 Aligned Barcodes

GuttenPlag (21) proposed a *Barcode* visualization to highlight the amount of plagiarised text per page using colors (Figure 6.6). In contrast to such a singular barcode that

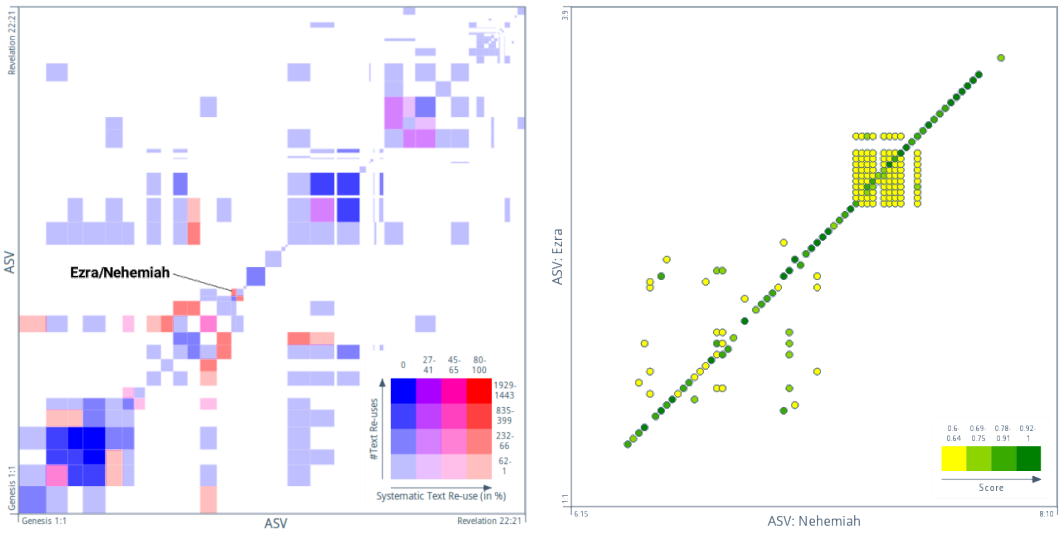


Figure 6.4: Text re-use Grid and text re-use browser illustrate text re-use patterns at different text scales (265).

only visualizes how passages of one text are reused, aligned barcode visualizations communicate the relations to reference documents by displaying alignment patterns.

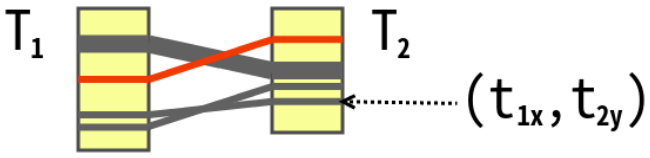


Figure 6.5: Visual representation of Aligned Barcodes.

Visual Representation Aligned barcodes are typically restricted to illustrating the alignment patterns between two texts. The sizes of the text boxes representing T_1 and T_2 represent the lengths of the texts or the number of text units. Each alignment tuple (t_{1x}, t_{2y}), typically on the level of a sentence, is drawn in the form of a line string connecting the relative positions of the related text fragments in T_1 and T_2 (Figure 6.5). Displaying all alignment tuples that way results in alignment patterns being visible. Next to line strings, rectangular areas can be used to illustrate the alignment of entire paragraphs (**?**). Also, color can be used to highlight different types of alignments (116).

Applications Aligned barcodes often appear as overview visualizations in addition to side-by-side views in collation (6, 116, 121) or translation scenarios (**?**). They direct

the user to interesting patterns of consecutive alignment units or transposed passages, and the aligned barcode can be used to navigate through the texts; an example of this scenario is shown in Figure 6.12. Aligned barcodes are also used to analyze text re-use patterns (2). The Text Re-use Browser uses color to indicate the matching score of an aligned text fragment, and the occurring pattern informs on repetitive or systematic text re-use. Picapica (22) applies an aligned barcode to visualize plagiarism using T_1 for the observed text and the box area for T_2 for the entire corpus of reference texts, which makes all plagiarised fragments explorable.

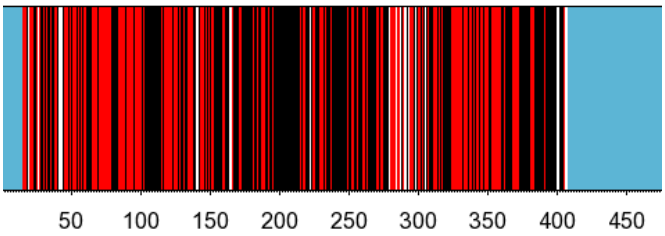


Figure 6.6: GuttenPlag Barcode Visualization.

Strengths & Weaknesses Aligned barcodes are text alignment abstractions. That makes them applicable to long texts, but they always require a linked text-based representation to expedite knowledge discovery. They are limited to comparing two text variants, typically on line, sentence or paragraph level. This makes patterns of subsequently aligned text fragments, as well as transpositions, easy to spot. Nevertheless, aligned barcodes can be hard to read with an increasing number of transpositions that occlude due to crossing lines.

6.1.1.4 Text-oriented Heat Maps

Yet having numerous source texts in the underlying corpus, text-oriented heat maps focus on a singular text overlaid with alignment information.

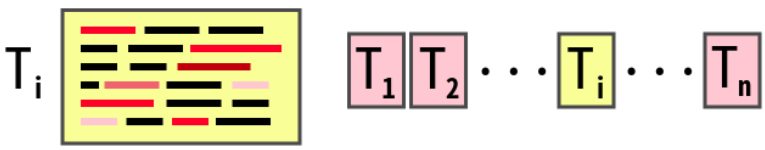


Figure 6.7: Visual representation of Text-oriented Heat Maps.

Visual Representation The focused text T_i is selected from the corpus $T_1, \ldots, T_n$ and the text (flow) of T_i gets magnified (Figure 6.7). Typically, words or phrases receive a colored background indicating how often they appear in the other texts of the corpus, or how strongly they vary. Single hue color maps (103) are used to differentiate between low and high frequency.

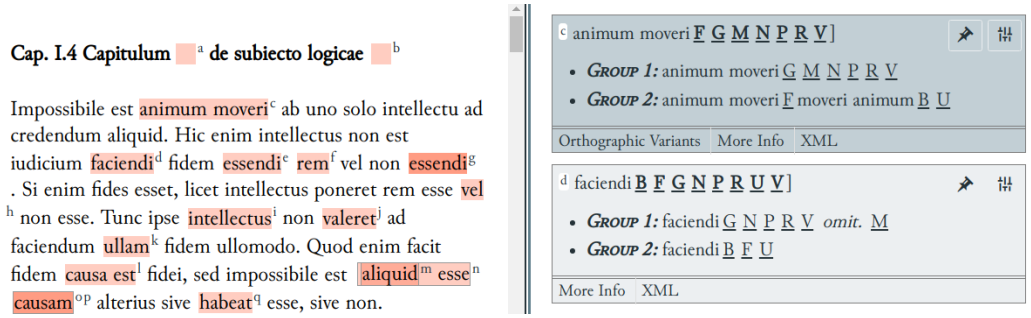


Figure 6.8: Edition Visualization Technology (EVT) text-oriented heat map visualization. Variance positions are highlighted and the critical apparatus is placed on the right side (211).

Applications JuXtacommons (121) and Edition Visualization Technology (211) use this technique to illustrate textual variations among multiple witnesses. They both utilize different degrees of saturation to display different levels of variance from the base witness T_i, the more saturated, the more witnesses differ (see Figure 6.8). Similarly, text-oriented heat maps can display how often a text fragment has been re-used (232). GuttenPlag (21) applies this technique to highlight the page-wise amount of plagiarised text passages; pages colored in saturated red contain a large amount of re-used text fragments, as shown in Figure 6.9. Summary Explorer (235) employs text-oriented heat maps to visualize the agreement between a summary and the original article on lexical and semantic level (Figure 6.27).

Strengths & Weaknesses The visual output is always generated for an individual text in focus. Variances to the other text editions are summarized and projected on the focused text so that individual relations to other texts are only conceivable through detailed views. Moreover, relations among other text editions remain hidden. However, text-oriented heat maps produce an excellent picture of variation on word, phrase, and paragraph levels, and the focused text can be quickly followed.

Figure 6.9: GuttenPlag text-oriented heat map visualization. Plagiarism positions are highlighted with different degrees of red color reflecting the amount of plagiarised text (21); images licensed under CC-BY-SA.

6.1.1.5 Side-by-side Views

Side-by-side visualizations are the most commonly used technique to display aligned texts at different levels. It locates the compared texts alongside each other and aligned units, which can be paragraphs, lines, sentences, phrases, or words, are highlighted. A navigation bar or aligned barcodes are usually associated with the side-by-side model to facilitate navigation through the aligned units in case of long texts.

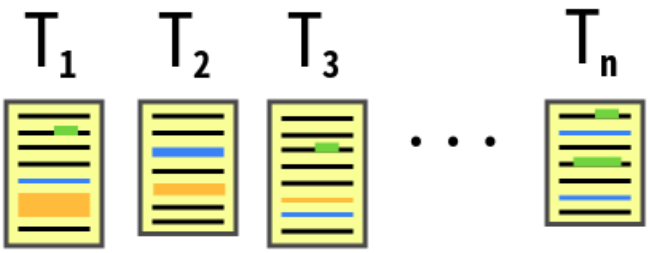


Figure 6.10: Visual representation of side-by-side views.

Visual Representation Side-by-side views reserve a rectangular space to display the text of different aligned editions $T_1, \ldots, T_n$ next to each other (Figure 6.10). Alignment tuples are not explicitly drawn in the form of connections. Typically, words, or their backgrounds, are colored to indicate—dependent on the use case—matches or variance among editions. Users need to scan the parallel texts to identify positions of variance. Interaction means like mouseover or mouse click are used to relate text units with their correspondences; thus, individual alignments can be shown on demand. Different colors are used to indicate different types of alignments.

Figure 6.11: LERA side-by-side view combined with a sequence-aligned heat map to facilitate navigation through the texts (215).

Applications Side-by-side views are the most commonly used technique to visualize *collation* results, very often limited to two text variants being compared (5, 88, 116, 121, 251). Versioning Machine (214) is one of the earliest tools proposed to visualize more than two TEI-encoded texts; the first version was launched in 2002. While it provides hover functionality to relate passages simultaneously, a coloring scheme to differentiate alignment types is not implemented. LERA is also capable of juxtaposing more than two variants (215). It colors inserted text fragments in green, substitutions in blue, and deletions in red. Figure 6.11 shows an example of seven juxtaposed variants of the German fairytale *Rumpelstilzchen*. Edition Visualization Technology also uses color to highlight distinct typology, but further apply color saturation to indicate the variability of a tradition (211). As for *collation*, side-by-side views also support analyzing text re-use patterns between two arbitrary texts of a corpus (2, 265). The color of text in those scenarios only indicates the re-used fragments. Side-by-side views are the means of choice to visualize translation alignments. Tools such as Ugarit (65), Linguee (3),

Glosbe (1) and Reverso Context (4) provide bilingual *KeyWord-in-Context (KWIC)* search functionality. The results of a keyword search are the translation and juxtaposed contexts in source and target language, in which the keyword appears; the word and its translations visually stick out (see Figure 6.13). Such word- or phrase-level alignments are also used in academic contexts (13, 223, 266). Ugarit uses color to discriminate aligned from unaligned tokens (65), and hover functionality is used to highlight related tokens. Cheesman et al. (50) apply side-by-side views to explore translation alignments among different translations of Shakespeare's *Othello*, by juxtaposing two texts, but also proposing a layout for base text and two translations.

Figure 6.12: Iteal combines aligned barcodes with a side-by-side view (116).

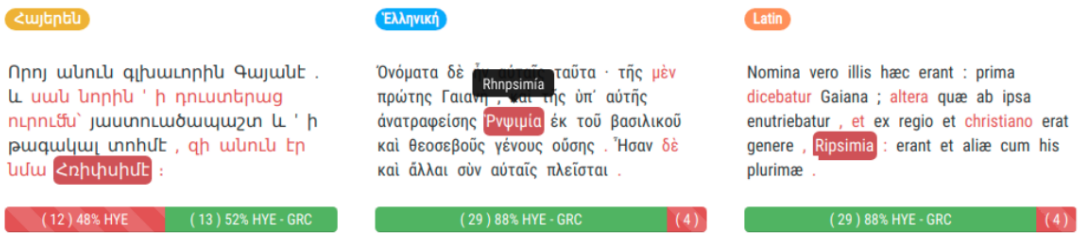


Figure 6.13: Ugarit side-by-side translation visualization, aligned words in Armenian, Ancient Greek and Latin text fragments are highlighted (65).

Strengths & Weaknesses A major constraint for side-by-side views is the limited screen width, which makes them applicable for a limited number of text variants. Users might also get distracted since they have to move the eyes between parallel texts to

identify positions of variance. However, in side-by-side views it is easy to read larger text fragments of a variant. Further, they allow inspecting variance on different text hierarchy levels, i.e., word, sentence, and paragraph.

6.1.1.6 Tabular Views

Tabular views visualize word-level variations for small text fragments such as sentences or short paragraphs.

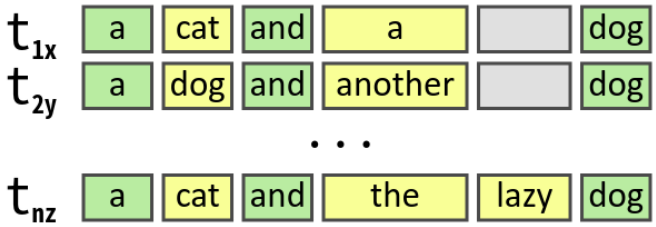


Figure 6.14: Visual representation of Tabular Views.

Visual Representation This views are represented in the form of a table, rows represent different aligned text units $t_{1x}, t_{2y}, \ldots, t_{nz}$. And columns stand for aligned tokens. Typically, columns are colored in order to show similarity or variation. Cells might be empty if a text unit does not contain a variant reading for a token (Figure 6.14).

Applications Tabular views are widely used to visualize text variants at word level (88, 178, 251). An example of turning six variants into a tabular view with CollateX (67) is shown in Figure 6.15. The *Interlinear Text* view offered by Alpheios (13) and Biblehub [1] employs the method to visualize translation alignments. Alpheios displays each word in the source text combined with its translation displayed under it while assigning a red color to unaligned tokens (Figures 6.16 and 6.17).

Strengths & Weaknesses Using tabular views, it is easy to spot aligned patterns on word and phrase level. In addition, a text variant, for which an entire line is reserved, can be easily followed. Though this representation works well for less-varying text fragments–although similar patterns induce increased space coverage through redundant information–, transposed passages quickly disperse the view as the number of columns and the amount of white space increase. In general, the screen size limits the number of

[1] https://biblehub.com/interlinear/matthew/1-1.htm

W1	At	the	first		God	made	the	heaven	and	the	earth	.
W2	In	the	beginning		God	created	the	heavens	and	the	earth	.
W3	In	the	beginning	, when	God	created	the	universe,				
W4	In	the	beginning	of God's preparing			the	heavens	and	the	earth	--
W5	In	the	beginning		God	created	the	sky	and	the	earth	.
W6	In	the	beginning		God	created		heaven	and		earth	.

Figure 6.15: CollateX tabular view visualization for six different English translations of the first Bible verse. Aligned tokens are placed at the same column green color reflects a total agreement among text variants (67).

	976 [e]	1078 [e]	2424 [e]	5547 [e]	5207 [e]	1138 [e]	5207 [e]	11 [e]
	Biblos	geneseōs	Iēsou	Christou	huiou	Dauid	huiou	Abraam
1	Βίβλος	γενέσεως	Ἰησοῦ	Χριστοῦ ,	υἱοῦ	Δαυὶδ ,	υἱοῦ	Ἀβραάμ :
	[The] book	of [the] genealogy	of Jesus	Christ	son	of David	son	of Abraham
	N-NFS	N-GFS	N-GMS	N-GMS	N-GMS	N-GMS	N-GMS	N-GMS

Figure 6.16: Bible interlinear visualization.

Τοῦ	δὲ	Ἰησοῦ	γεννηθέντος	ἐν	Βηθλέεμ	τῆς	Ἰουδαίας	ἐν	ἡμέραις	Ἡρῴδου	τοῦ	
	ergo	Iesus	natus	in	Bethleem		Iudaeae	in	diebus	Herodis		
βασιλέως	,	ἰδοὺ	μάγοι	ἀπὸ	ἀνατολῶν	παρεγένοντο	εἰς	Ἰεροσόλυμα				
regis		ecce	magi	ab	oriente	venerunt		Hierosolymam				

Figure 6.17: Alpheios interlinear visualization.

text variants that can be compared. Further, tabular views only work at word level and produce readable representations for short sentences.

6.1.1.7 Text Variant Graphs

Similar to tabular views, text variant graphs illustrate word-level variations for small text fragments, but, employing a graph structure to represent variation enables visualizing additional alignment features.

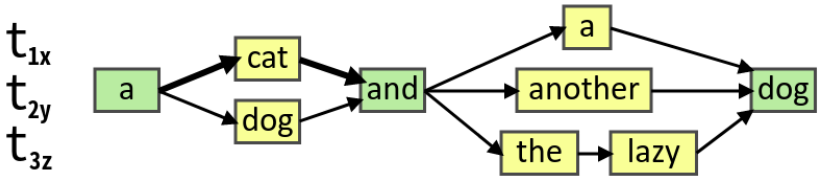


Figure 6.18: Visual Representation of Text Variant Graphs

Visual Representation Schmidt and Colomb (212) were the first who used a graph to visualize textual variance for multi-version texts. They proposed the *Variant Graph*, a directed acyclic graph (DAG) with nodes typically representing words or tokens of aligned text units. An edge in the graph stands for a pair of subsequent tokens in any of the textual variants. Thus, an edge represents a version, and each text variant contributing to the graph defines a path in the variant graph. The size of nodes may reflect how often a word appears among the editions. Likewise, the thickness of an edge can display the number of variants, including the corresponding word pair. Edges can be labelled or colored to inform on the corresponding variants.

Applications Among all visualization techniques discussed, the text variant graph is the only one specifically tailored to communicate a *collation* result of small text fragments. CollateX (67) adopted the concept to illustrate the results of their collation alignment algorithm. As edges are labeled, the graphs typically have a large width, even for short text fragments. Transpositions are shown with dashed lines. Stemmaweb (16) extended the idea proposing an interactive version of the graph to support stemma analysis. TRAViz (118) applies diverse visual features to communicate features typical for variant graphs aiming to make them easier to digest (see Figure 6.19). This includes sizing nodes according to frequency, coloring instead of labeling edges, and a sophisticated graph drawing algorithm that aligns related variants vertically. WordTree (250) and WordGraph (210) employ similar visual cues, but they are not particularly designed to

support the collation task. Silvia et al. (7) applied the storyline metaphor to visualize variations in classical texts. The model uses a force-directed layout algorithm to generate a topologically representative layout, which enables users to read and recognize text patterns with reasonable speed and accuracy; an example is shown in Figure 6.20.

Strengths & Weaknesses In contrast to tabular views, variant graphs reduce redundant information by merging equal or similar tokens. This strategy likewise makes it easy to spot divergent and similar sections among the variants. The visual output facilitates creating critical editions as text variant graphs can be seen as automated collation results. The variant graph design's backside is that individual variants are not salient, typically highlighted after user interaction. As tabular views, variant graphs currently only work on word level, at best with short, less-varying text fragments.

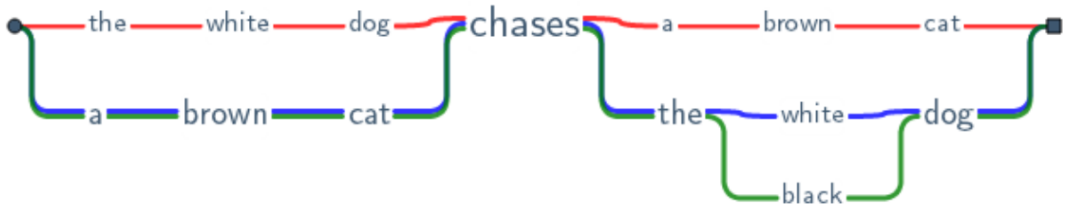


Figure 6.19: TRAViz graph showing variance among six English translations of the first Bible verse (118).

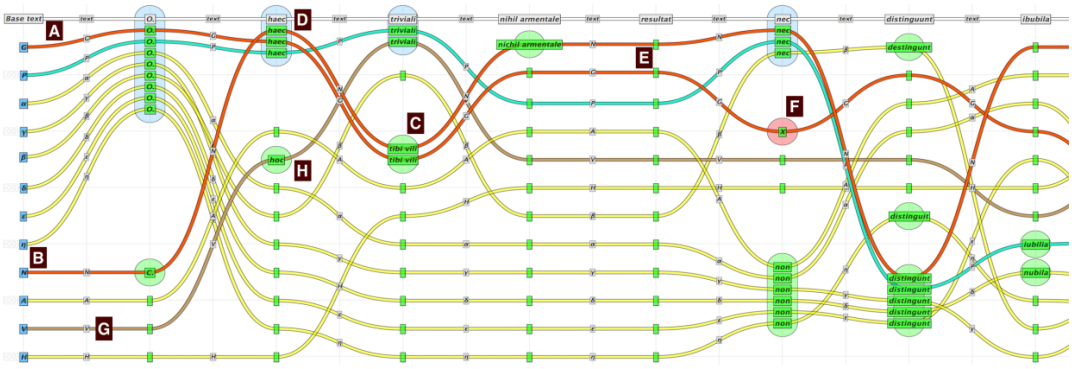


Figure 6.20: Storyline-based text variant graph (7).

6.1.1.8 Miscellaneous Techniques

Next to the techniques mentioned above, a few related works favor other methods to support the visual exploration of text alignment patterns. The Chinese Text Project (232)

uses a network graph to visualize the textual relationships in pre-Qin and Han texts. Each node represents a document in the corpus and edges indicate a similarity between documents. Edge thickness reflects the amount of shared (re-used) text fragments between two nodes. ShakerVis (93), developed to explore segment variations among German *Othello* translations, applies a parallel coordinates plot and linked scatter plots to illustrate how different translations relate to each other.

6.1.1.9 User Interactivity

Alignment tools typically provide means of interaction to configure the alignment computation and to manipulate and analyze the result's appearance. We distinguish pre- and post-alignment interaction.

Pre-alignment User Interactions While users can typically upload their own texts to be aligned, these interactions are rather configurations that influence how alignment algorithms will operate. For instance, users are enabled to adjust if the alignment algorithm will be sensitive to punctuation, capitalization, or diacritics or to ignore non-alphabetical tokens or line breaks (88, 121, 215). Other tools allow users to choose between exact or fuzzy word matching based on the Levenshtein distance (88, 116, 118). Further, some tools are equipped with various alignment algorithms the user can choose from (6, 67, 118).

Post-alignment User Interactions Such interactions relate to tasks users can perform once the alignment is computed and visualized. Post-alignment interactions span the entire range of Brehmer's task taxonomy (40).
The central task for which alignment visualizations are used is to ***compare*** different text versions. A variety of user interactions are offered for this purpose, e.g., ***search****ing* the alignment output or ***filter****ing* the results (65, 121, 215). Standard mouse interactions are typically reserved to highlight specific alignment patterns and to indicate similarity (13, 65, 118, 121, 214), or to demand detailed information or statistics (65, 121, 211, 215, 265). Some tools offer alignment visualization at different scales alongside with multifaceted means to ***browse*** and ***explore*** the alignment result. Typically, individual alignments can be ***select****ed*, and a fine-granular visual encoding is shown (13, 116, 121, 215, 251, 265). In other tools, visual components can be rearranged by the user, e.g., by moving adjacent panels triggering an update of the visualized alignment pattern (214), or by switching between horizontal and vertical tabular views (203). As alignment results might not meet users' expectations, some tools allow to ***manipulate*** the alignment interactively. For instance, in variant graph visualizations, it is allowed to split and

merge nodes (118, 204). In text-oriented heat map tools, users are able to ***change*** the base witness (121, 211). Juxtacommons (121) supports users to ***annotate*** notes to variance positions in the heat map.

Lastly, many tools support the ***output*** task by offering to export or download an alignment result for further usage, mostly in XML or JSON format (5, 13, 65, 121, 215). Some tools let users share their visualization via URL (121), or the visual output can be explicitly downloaded as image.

6.1.2 Design Implications

Visualization of textual alignment plays an important role in facilitating understanding and interpreting the relation between compared texts. We studied 40 tools and classified them according to the task and visualization techniques they offer to analyze occurring alignment patterns (see Table 6.1). Concerning different features of an alignment scenario, listed in Table 6.2, we classified the visualization techniques according to applicability (green fields), potential applicability (yellow fields), and inapplicability (red fields indicate future challenges). Respective fields are colored grey if the technique is by design not applicable to the alignment task. For adopting one of the discussed techniques or innovating a new visualization model for textual alignment, one should consider the following factors:

Alignment Task The selection of the appropriate visualization technique is affected mainly by the underlying task. In general, collations focus on varying text fragments and require highlighting variance patterns. For instance, variant graphs and text-oriented heat maps merge common parts of the compared texts easing to spot differences. However, text re-use and translation alignment rather focus on common patterns, which should be reflected in the visualization output, for example, by connecting related passages with lines. One can also consider switching between highlighting similarity or dissimilarity (115). In order to indicate an alignment relation, the best practice is to use the same color for related units (5, 121, 214). Another strategy is to merge common text units into one visual representation (118, 121, 211). Other techniques use connecting paths between related text passages (2, 6, 22, 50, 67, 116).

Number of Text Variants In most scenarios we observed, the number of compared text variants is below ten. Only heat map visualization techniques can handle a larger number of variants. Two main strategies apply. On the one hand, a single variant is focused, and text-oriented heat maps are used for visualization. On the other hand, a

heat map is used to generate an overview, and a subset of variants can be analyzed in more detail using a text-based visualization mode.

Text Hierarchy Levels Texts inhere different structures, in which fragments are organized, e.g., page and line or chapter and sentence. Before tailoring a new or choosing among existing solutions, one should clarify to what extent alignment patterns should be explorable at different scales. Appropriate visual encodings need to be selected in such a case. For instance, *iteal* uses aligned barcodes to inspect occurring patterns within the whole document, a side-by-side view for line-level alignments in sections, and variant graphs for analyzing word-level alignments for lines (116). For navigation purposes, Shneiderman's information-seeking mantra (219) *"Overview first, zoom and filter, then details-on-demand"* should be implemented, i.e., different visualization techniques should be linked as coordinated views to enable users exploring the corpus smoothly.

Language All discussed visualization techniques are applicable to a monolingual corpus. Although most scenarios deal with left-to-right Latin scripts, tools are also applicable to right-to-left scripts (13, 65, 117, 266) or writing systems without word boundaries (56). However, not all visualization techniques are applicable to multilingual scenarios.

Text Stability Less-varying texts give users more freedom to select an appropriate visualization technique. However, the more transpositions occur, the less meaningful the visualized results also get for the techniques showing green fields in Table 6.2. When dealing with unstable texts, we recommend applying various means of interaction, i.e., allowing to filter alignments according to a certain matching score or to hide all transposed alignments.

6.1.3 Future Challenges

Drawing from our own experience in developing text alignment visualizations and the open challenges highlighted as red fields in Table 6.2, we derived future research directions listed below:

The Need for a Modular System Drawing from the overview in Table 6.2, one can see that for each possible feature, at least one visualization exists that can be applied. Moreover, our observations when preparing the survey exposed a high number of redundant tools for solving the same issue. This indicates that research among the visualization, NLP, and digital humanities communities is not well known. However,

Corpus & Task-related Features		Sequence HM	Grid HM	Aligned Barcodes	Text HM	Side-by -side	Tabular View	Variant Graph
Tasks	Collation	✓	✓	✓	✓	✓	✓	✓
	Text Re-use	✓	✓	✓	✓	✓	(✓)	(✓)
	Translation Alig.	(✓)	✓	✓		✓	✓	
Languages	Monolingual	✓	✓	✓	✓	✓	✓	✓
	Multilingual	(✓)	✓	✓		✓	✓	
Alignment Levels	Word	✓	✓	✓	✓	✓	✓	✓
	Sentence/Passage	✓	✓	✓	✓	✓	(✓)	
	Document		✓					
Corpus	scalable to n texts	≈ 50	≈ 50	2	No Limit	≈ 10	≈ 20	≈ 20
	unstable texts		✓	✓	✓	✓		

Table 6.2: Applicability of Visualization Techniques to Corpus & Task-related Features

a modular system applicable to any related text alignment scenario would be an appropriate strategy to prevent this issue. A modular system could first allow choosing among different preprocessing means. Second, it could offer to change the alignment algorithm, as this is already done by CollateX (67). Finally, the user of such a system could toggle between different modes of visualization, gaining on-demand changes of perspective on the same matter.

The Need for Scalable Solutions Though visualizations seemingly serve any upcoming demand, Table 6.2 also indicates that only one scalable solution (text-oriented heat maps) exists. Other visualization techniques are limited to tens of texts to be compared. With the steady growth of textual data, the need to develop novel scalable text alignment visualization techniques represents a significant future challenge. For example, the Bible is known to exist in more than 450 different English translations. None of the listed approaches supporting collation is ready to visualize variation among them appropriately. Likewise, visualizing text re-use patterns in a large-scale text corpus in an efficient, clear, and user-friendly way would be desirable. Going back to the *Hamlet* example, it would be highly interesting to develop a technique capable of visually conveying the spread of Hamlet quotations offered ready-to-use by the HyperHamlet project (206).

The Need for Document-based Visualizations Table 6.2 shows that only one standard visualization technique, the grid-based heat map, is ready to convey alignment relations on the document level visually. All other techniques are, by default, incapable of visualizing such information. The Chinese Text Project (232) uses a standard tool to

generate a network graph also capable of showing relations among documents. However, such solutions require future extensions to convey alignment patterns across texts meaningfully. For example, graph drawings could be tailored to cluster text editions or include temporal dependencies inherent in text corpora.

Variant Graph Opportunities The frequent usage of the variant graph to support collation tasks indicates its value as opposed to other representations like the tabular view. However, Table 6.2 shows that its current design does not apply to the entire variety of textual features a corpus might enclose. For example, the structure of a variant graph might be valuable to support manual alignment scenarios. Further, variant graphs could also be applied to analyze sentence sequences across text editions. Especially for variant graphs, defined as being directed acyclic, an inclusion of loops to overcome the issue when dealing with unstable texts could be a beneficial extension. Lastly, variant graphs are, by default, not limited in how many text variants they can bear. However, we set the number in the table to twenty as almost no examples with a higher number of text editions exist. Developing smart methods of clustering recurring phrases among the editions that are not crucial for analyses and visualizing them could deliver insightful views of larger text bodies. All listed extensions of the variant graph require modifications in node and edge representation as well as a revision of the interaction scheme.

Exploring New Data Sets and Alignment Levels Alignment of *OCR output* has been employed to perform various tasks (39, 252, 264). Visualizing this kind of alignment can be challenging, especially if we consider relating output characters with their actual positions on the scanned pages. *Movies subtitles* also provide a great source of multilingual parallel texts. The alignment of movies subtitles at the sentence or word level has been the subject of numerous studies (10, 242, 243). Visualizing those alignments in a scalable, interactive form combined with search functionality could be a valuable resource for professional translators and translation studies researchers.

Technique Transfer to Related Domains Alignment is not limited to text only; it has been employed in other domains. The current visualization techniques could be transferred and used to visualize the alignment between other data types. Proven to be effective for text-based scenarios, a transfer to other domains could deliver new, interesting perspectives. For instance, variant graphs could be applied to visualize alignments among protein sequences, side-by-side views can be employed to alignments

among musical scores, or sequence-aligned heat maps can be applied to visualize audio signal alignments.

In the course of the last years, we have been ourselves involved in developing visualization techniques to support a diversity of text alignment scenarios. However, the overview of related tools and scenarios, in which related text alignment visualizations are applied, and what visual features they choose to convey alignment-related features, has been an interesting exercise. The resulting survey provides a comprehensive overview of how text alignment visualizations currently support domain-related tasks.

6.2 Visualizing Different Alignment Levels

In the context of the thesis, we have employed visualization in different aspects to accomplish various alignment-related tasks in light of the survey and the proposed design implications. Table 6.3 provides an overview of the implemented tools and the visualization techniques used to display the alignment at different levels. In the following, we present the tools and visualization views we implemented:

	Alignment Level		**Visualization Technique**					
	Word	**Sentence**	Side by Side	Grid View	Tabular View	Text H. Map	Parallel Coords.	Misc.
Ugarit	✓		✓					✓
EvAlign	✓		✓	✓				
Parallel Corpus Visualizer	✓		✓	✓	✓			
Summary Explorer	✓	✓	✓			✓		✓
Semantic Frames	✓		✓				✓	
Divan Hafez	✓	✓	✓				✓	
Information Flow	✓		✓					✓

Table 6.3: An overview of the implemented tools and the used visualization techniques.

6.2.1 Ugarit Translation Alignment Editor

In addition to its primary objective of generating word-level alignments manually, Ugarit offers a user-friendly reading environment for parallel texts. Ugarit shows the aligned texts side-by-side and provides transliteration to the Latin alphabet for many languages. Also, it allows users to search for a token and shows the results in different aggregated views. We discussed Ugarit and the various visualization approaches in Chapter 2.

6.2.2 EvAlign

EvAlign is a visual analytics tool that supports developers and practitioners in performing a qualitative and quantitative evaluation of automatic alignment models. We discussed EvAlign in detail in Chapter 5 and introduced various corpus- and sentence-level visualization approaches to facilitate inspecting and analyzing the performance and spotting frequent error patterns.

6.2.3 Parallel Corpus Visualizer

The Parallel Corpus Visualizer serves as basic reading environment for parallel texts to assist scholars and researchers in exploring an aligned corpus and looking closely at the aligned words and phrases and their aggregated statistics, which has the form of a dynamically extracted bilingual dictionary. The tool takes as input a manually or automatically aligned corpus in a specific format and parses the corpus to extract translation pairs with their frequency. Combining multiple aligned corpora is also possible. Additional annotation layers, such as part-of-speech tags and lemmas, can also be integrated.

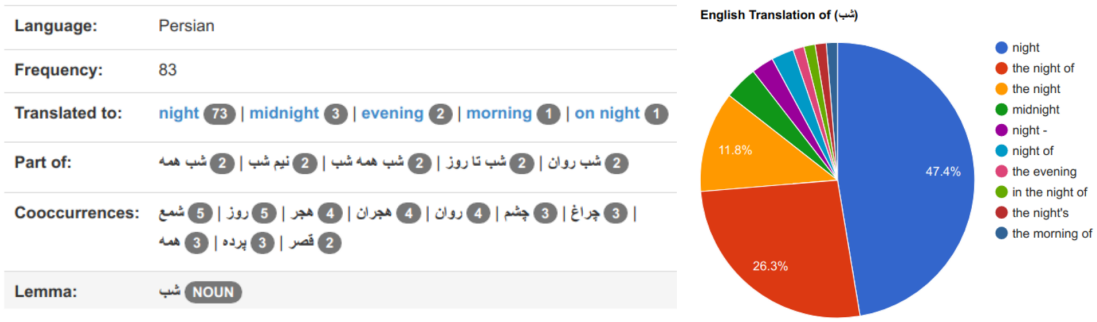


Figure 6.21: Parallel Corpus Visualizer aggregated search results.

The tool allows users to search for a specific word and visualizes the results in side-by-side views. Further, words are clickable, and when clicked, users will be forwarded to a page that displays all statistics about the word, such as frequency, possible translation equivalents and their frequencies, co-occurrences, and if the word is part of any frequent phrases in the corpus. It also shows the sentences where the selected word occurs in an interactive side-by-side view that shows the alignment on demand. The word will be already highlighted to facilitate capturing the contexts where the word occures.

We used this tool to visualize Divan Hafez[1], a famous Persian poetry consisting of 495 poems aligned manually at the word/phrase level with the English translation[2]. Further, we used it to visualize diverse automatically aligned texts from the Perseus Digital Library[3] (265).

6.2.4 Summary Explorer

The Summary Explorer[4] is an interactive visual analytics framework that supports quantitative and qualitative evaluation of automatic summarization models[5]. The framework offers various visualization views to facilitate navigation and exploration of the hosted datasets. It also helps developers and practitioners conduct a systematic quality assessment of the hosted models and evaluate their outputs regarding three important summary quality criteria: coverage, faithfulness, and position bias. The framework hosts the outputs of 55 automatic summarization models covering three benchmark datasets. The datasets hosted on the Summary Explorer include only single-document English summaries. However, the proposed visualization approaches and assessment scenarios can also be employed to visualize the outputs of cross-lingual summarization models and benchmark datasets, such as XWikis Corpus (198).

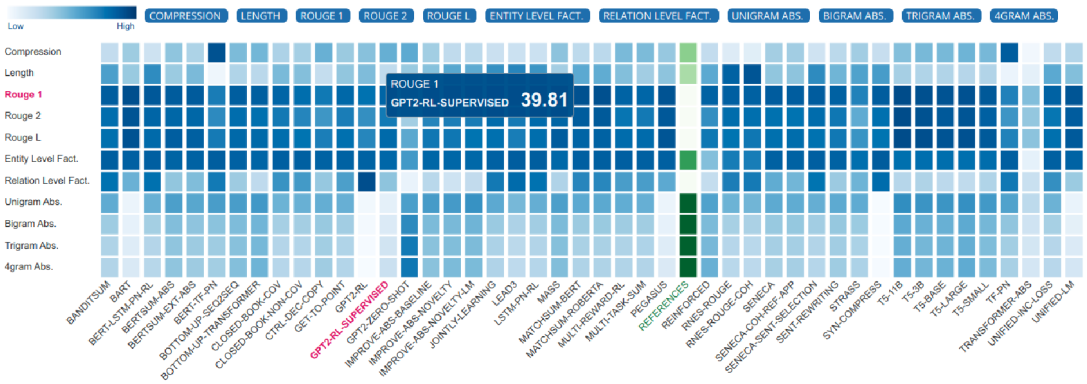


Figure 6.22: Summary Explorer heat map overview.

[1] `http://divan-hafez.com`

[2] The manual alignment has been done by Saman Rezaie and Maryam Foradi.

[3] `http://dynamiclexicon.com/`

[4] `https://tldr.webis.de/`

[5] This is joint work, and it has been done in collaboration with Shahbaz Syed, who was responsible for preprocessing the data and calculating the evaluation metrics, and Tariq Yousef was responsible for designing and implementing the visualization views.

Visual Design The conducted survey of text alignment visualization gave us a complete picture of the visualization technique used for text alignment and allowed us to tailor customized visualization views appropriate for the task. We applied Shneiderman's (219) Visual Information seeking Mantra ("overview first, zoom and filter, then details-on-demand") to implement the guided navigation through SUMMARY EXPLORER, which consists of four steps:

The navigation starts with selecting a benchmark dataset, then a list of available summary quality aspects will be displayed with a preview of their tailored visualizations and interactive use. According to the user's choice, a comprehensive overview of all available models will be displayed as a heat map showing the averages of various quantitative metrics (Figure 6.22), which enables a targeted filtering of the models based on their quantitative performance. The heat map shows only average values of quality metrics, which doesn't reflect the real distribution of the scores over the whole dataset, therefore upon model selection, histograms of each model's score distribution for each metric are displayed. After models selection is done, the user is forwarded to the corresponding quality aspect's view. The visualizations for the individual aspects of the three quality criteria share the property that two texts need to be visually aligned with one another.

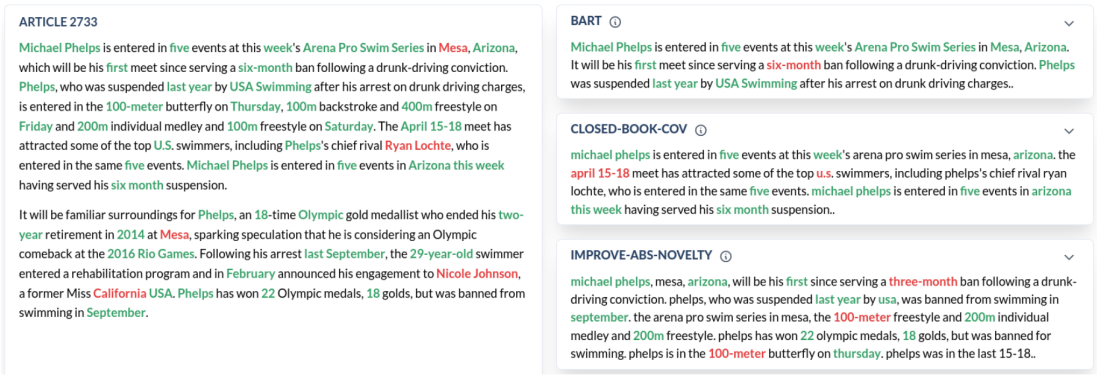


Figure 6.23: Summary Explorer entity coverage view.

Coverage View (Figure 6.24) Content coverage is visualized as alignment of summary sentences and document sentences at the semantic and lexical level in a side-by-side view. For entity coverage (relation coverage), a corresponding side-by-side view lists named entities (relations) in a summary and aligns them with named entities (relations) in its source document (Figure 6.23). For unaligned relations, corresponding document sentences can be retrieved.

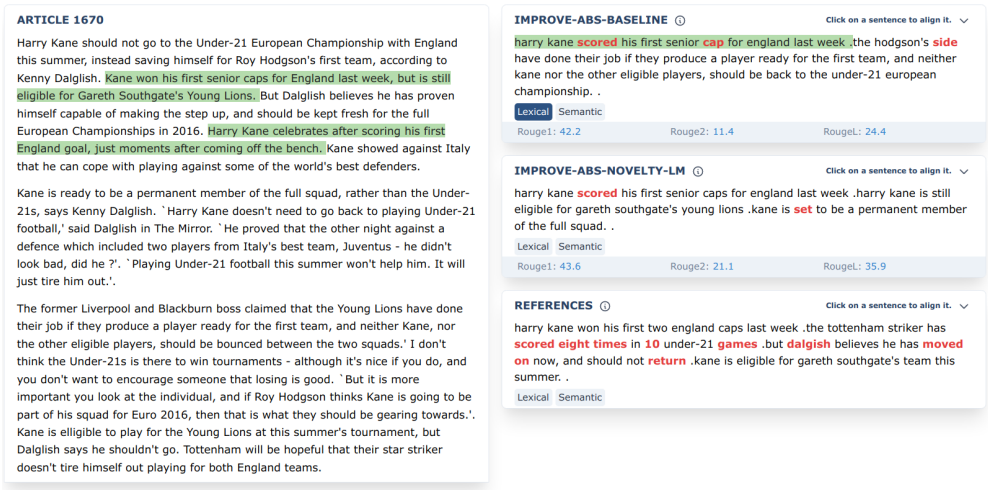


Figure 6.24: Summary Explorer content coverage ciew.

Figure 6.25: Summary Explorer hallucinations aggregated view.

Faithfulness View (Figure 6.24) Hallucinations are visualized by highlighting novel words in a summary. For each summary sentence with a hallucination, semantically and lexically similar document sentences are highlighted on demand. Since named entities and thus also entity relations form a subset of hallucinated words, the above coverage views do the same. Also, in an aggregated view (Figure 6.25), hallucinations found in multiple summaries are ordered by frequency, allowing to inspect a particular model with respect to types of hallucinations.

Position Bias View Position bias is visualized for all summaries given a source document (Figure 6.27) as a text heatmap, where a gradient color indicates for every sentence in a source document how many different summaries contain a semantically or lexically corresponding sentence. The darker the color, the higher the frequency. We also show the position bias for a specific model with respect to all its summaries in

Figure 6.26: Summary Explorer Relation Coverage View. Aligned relations are highlighted in green, whereas relations that have no correspondence in the original document are highlighted in red.

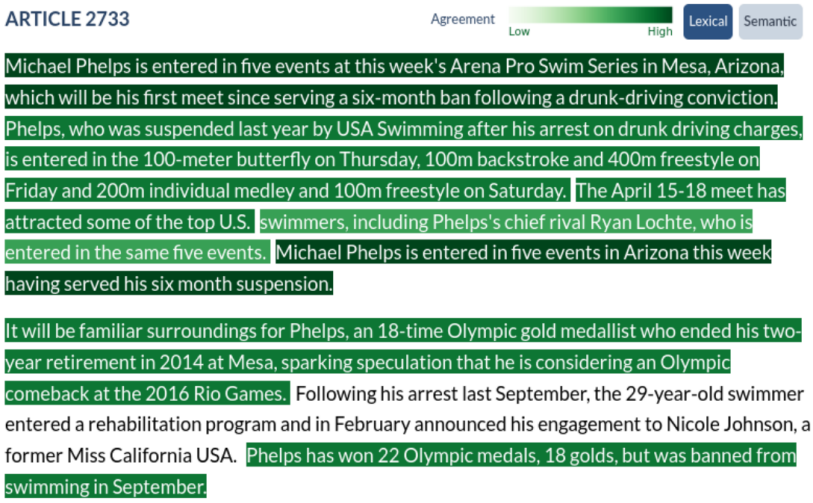


Figure 6.27: Summary Explorer position bias view.

a corpus (Figure 6.28). We visualize it as a heatmap for 50 randomly selected model summaries, where each summary is projected on a single horizontal bar representing the source document. Bar length reflects document length in sentences and aligned sentences are colored to reflect lexical or semantic alignment.

6.2.5 Miscellaneous Alignment-Related Visualizations

In the followings, we mention some showcases where we integrated additional annotation layers to the side-by-side views while using translation alignment as a bridge to accomplish various tasks:

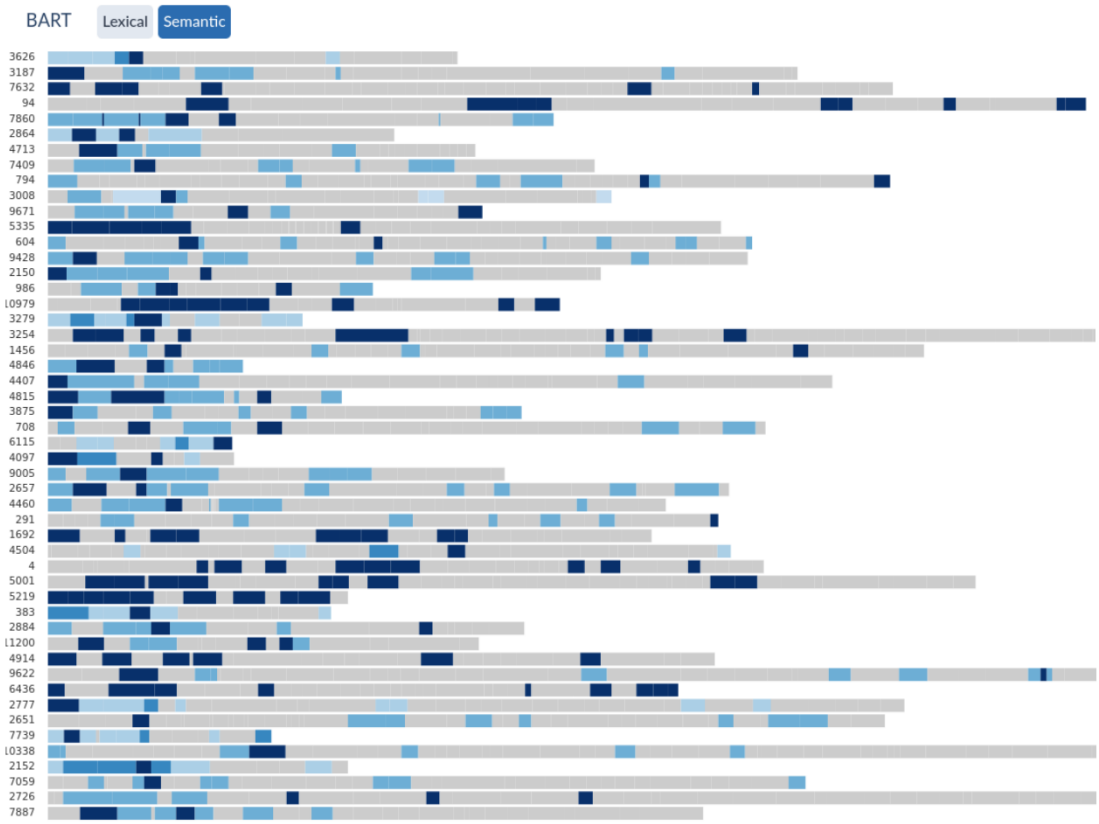


Figure 6.28: Summary Explorer positional bias aggregated view for 50 randomly selected documents.

6.2.5.1 Visualizing Alignment Inter-Annotator Agreement

Alignment gold standards are essential for assessing the performance of alignment models. However, developing gold-standard datasets is a non-trivial task. It requires manual annotation by domain experts, who will produce the annotations according to predefined guidelines. Even with the existence of clear and detailed guidelines, annotators can not reach an entire agreement on every single annotation. The experiment we presented in Chapter 2 (section 2.5.5.2) confirms this observation. While developing an automatic alignment model for Ancient Greek, we developed alignment gold standard datasets for three language datasets in collaboration with domain experts who annotated the parallel sentences and developed the alignment guidelines as well (see 4.1.1).

Comparing annotator behaviours and inspecting their alignment decisions for every sentence in the dataset would help to assess the robustness and clarity of the alignment guidelines and the quality of the resulting gold standard datasets. For this purpose,

we tailored a visualization view to show annotators' agreement and disagreement. The corpus-level view show aggregated statistics and includes common functionalities, such as searching, filtering, and sorting by IAA values. The view locates the sentences side-by-side and uses coloring schema to distinguish between different cases:

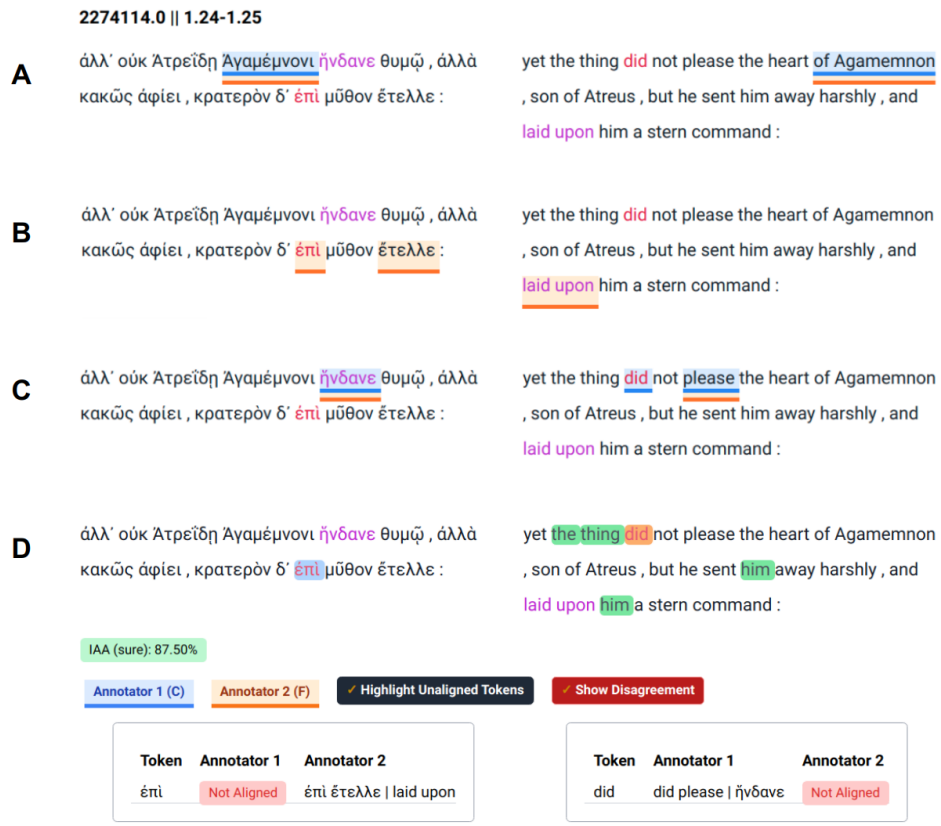


Figure 6.29: Annotators agreement visualization view.

Agreement. Tokens, where two annotators agree, are colored black. Agreement in this context means they took the same decision to align the token to the exact correspondences or left it unaligned. Hovering over a token shows how annotators linked it with other tokens (Figure 6.29A). Unaligned tokens will be highlighted on demand, and we use green for agreement and the annotator's individual color for disagreement (Figure 6.29C).

Disagreement. we distinguish between two cases: (1) tokens that are aligned by only one annotator is colored with red (Figure 6.29B). (2) tokens where annotators disagree

are colored with fuchsia (Figure 6.29C). On demand, disagreements can also be listed and visualized in a tabular view (Figure 6.29D).

The current implementation applies to visualize alignments of two annotators only. Although this is the most common case, there is still a need for a scalable solution.

6.2.5.2 Divan Hafez Visualization

An extended version of the Parallel Corpus Visualizer is under development, in which we provide translations of *Divan Hafez* in more languages and incorporate additional information and annotation layers such as transliteration to the Latin alphabet and the poem rhythm. Figure 6.30 shows the sentence level view. The different layers are aligned with the Persian tokens and will be highlighted according to the user's hovering. The alignments with English translation are created manually, whereas other languages are created automatically using the alignment model we discussed in Chapter 3.

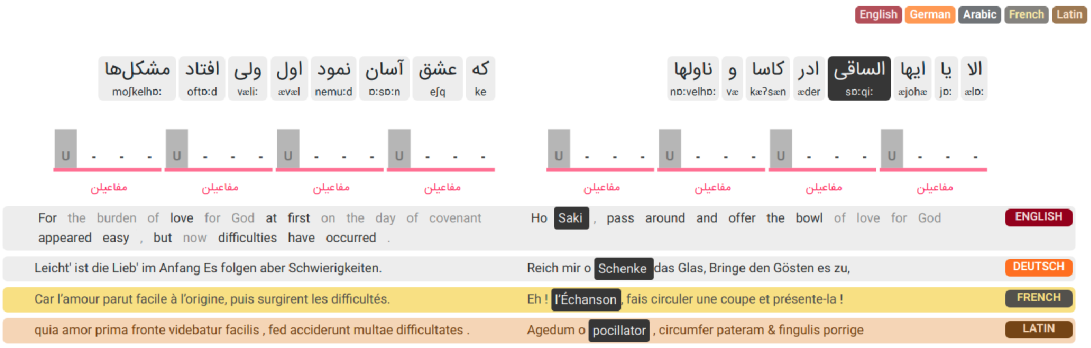


Figure 6.30: Divan Hafez sentence level view.

The corpus consists of 495 poems; every poem consists of multiple lines, and every line is composed of two parts. We provide alignments at the line (sentence) and word level. Figure 6.31 illustrates the alignment at line level among the original text and its English and German translations. Users can switch between the poem view and the detailed sentence-level views.

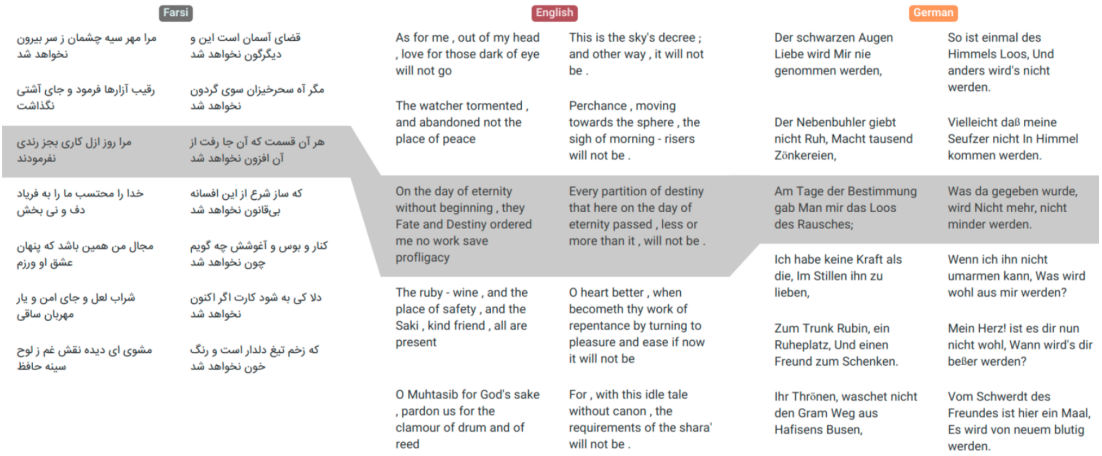


Figure 6.31: The poem-level view showing the original Persian poem aligned with the English and German translations at line level.

6.2.5.3 Visualizing Information Flow in Parallel Sentences

In the context of "Information Distribution and Language Structure"[1] project, we studied the linguistic information distribution between languages with strong noun/verb distinction, such as German, and languages with low noun/verb distinction, like Tagalog. For this purpose, we aligned the Bible corpus in German and Tagalog at word level and computed the *Information Content* (IC) for each word using our *Topic Content Model* (TCM) (132, 209).

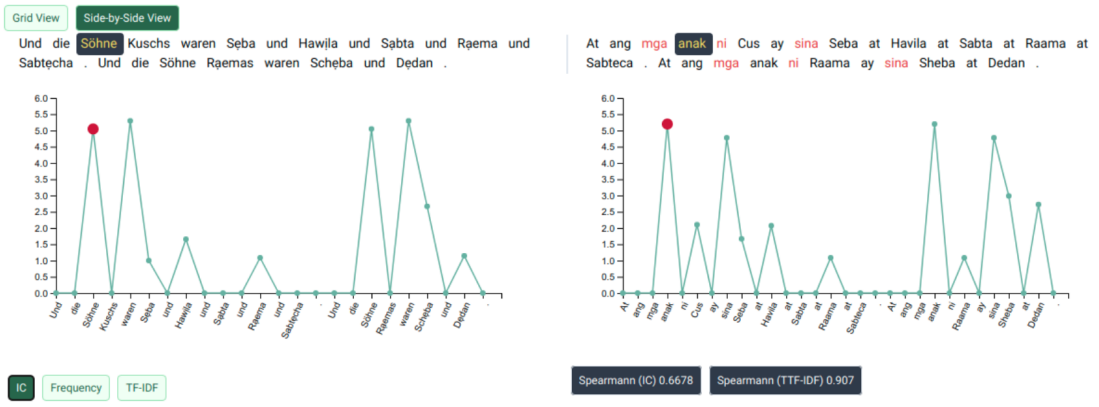


Figure 6.32: Information Flow.

[1]This project is funded by the Deutsche Forschungsgemeinschaft (DFG) - Project number 442315837. I have been working on this project between 2020 and 2022. `https://gepris.dfg.de/gepris/projekt/442315837`

Further, we calculated two other indicators, the word *frequency* in the corpus and the *tf-idf*. For the alignment, we used the embeddings-based alignment workflow explained in Chapter 3 (74, 114), and we fine-tuned the XLM-RoBERTa language model unsupervised with the Bible parallel sentences to enhance the word-level alignments.

The tailored visualization view allows users to compare the information distribution in the two sentences to assess if the information contents of two aligned words correlate or follow the same pattern. The view combines the side-by-side view to visualize the parallel sentences with two line charts to display the information flow for each sentence. Hovering any word will highlight it with the corresponding translation and information values in the line charts (Figure 6.32). Also, Users can switch among the calculated metrics through the buttons located at the bottom left side, and the line charts will be updated accordingly.

6.2.5.4 Visualizing Aligned Semantic Frames

In the context of our work on studying the cross-lingual semantic transfer of two semantic primes (see 5.4), we employed visualization to analyze and inspect the data at an aggregated level (corpus-level) and single sentence level. As we described, the alignment is done using fine-tuned multilingual language model and cross-lingual word-level semantic similarity, and the frame extraction is done using the *Large Ontology Multilingual Extraction* (LOME) framework (260).

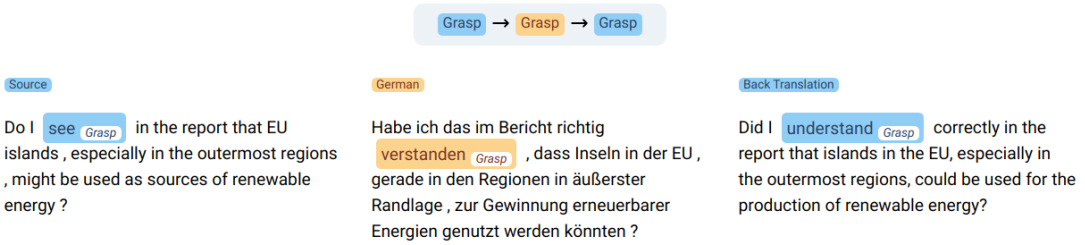


Figure 6.33: Sentence-Level view showing the cross-lingual semantic transfer.

In addition to the standard navigation and exploration functionalities, such as searching, filtering, and sorting, the corpus-level visualization includes an aggregated view of the semantic frames transfer (Figure 6.34). We used *Sankey Diagram* to model the semantic transfer across the three parallel sentences. The stream connecting two frames will be highlighted when hovered, and the corresponding sentences will be displayed when clicked.

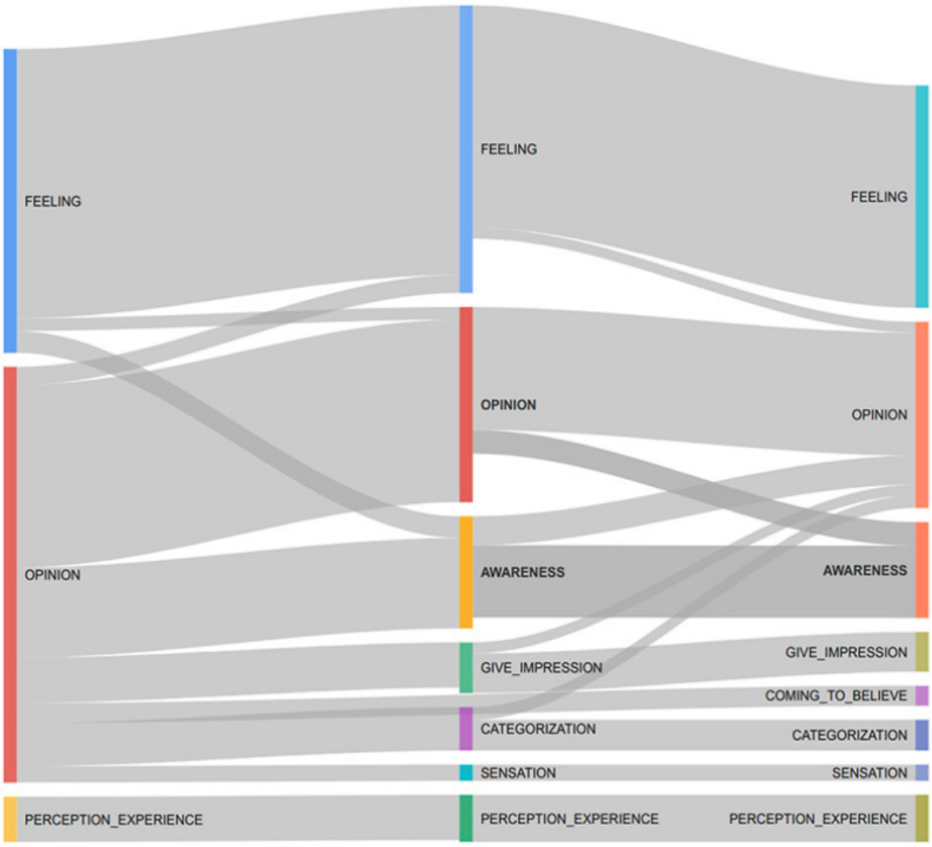


Figure 6.34: Semantic frames transfer aggregated view.

Further, the sentence-level view shows the three parallel sentences: the original English sentence, its German or Spanish translation, and the back translation to English produced by DeepL[1]. The sentences are located side-by-side, and the alignment is shown on demand. When users hover over a token, the corresponding tokens in the parallel sentences will be highlighted accordingly. In this view, an additional annotation layer is integrated to show the corresponding semantic frame. The semantic primes *See* and *Feel*, the subject of the study, and their correspondences will be highlighted by default, and their corresponding semantic frames will be displayed. A small legend summarizes the semantic frame transfer through the three sentences is displayed on the top of the sentence (Figure 6.33).

[1] `https://www.deepl.com/`

Chapter 7

Conclusion

This thesis aims to narrow the gap between NLP and DH by employing recent advances in language modelling and NLP models on historical texts. The automatic alignment model we trained serves as a bridge between modern and historical languages allowing students, scholars, and historical language learners to read and explore texts and languages they could not read before opening doors to explore the rich history and understand the past. The model also serves as a bridge to transfer linguistic annotations from modern languages for which off-the-shelf models are available. The output of this research can be summarized in the following:

Data

The main objective behind developing Ugarit was to collect training data to be used for training an automatic alignment model. Ugarit is user-friendly and intuitive, which facilitated its wide spread among students and literary researchers resulting in a large and diverse collection of aligned texts. From which we selected a subset of 190k translation pairs and used it to train the automatic alignment model. Further, we employed the alignment model to induce multilingual dictionaries by extracting the translation pairs from parallel corpora resulting in over 1.6M pairs. The translation pairs with their alignment scores are available on our Github repositories[1].

Models

For **automatic alignment**, we employed an alignment workflow based on multilingual embeddings, and cross-lingual semantic similarity (74, 114). The embeddings are generated by a multilingual transformer model. Since pretrained language models are

[1] https://github.com/UgaritAlignment/

not trained on Ancient Greek, we tested a training strategy that combines unsupervised training on monolingual Ancient Greek text, unsupervised training with parallel sentences, and supervised training with accurate manual alignments extracted from Ugarit. This strategy significantly enhanced the alignment workflow's performance and produced state-of-the-art performance on the various historical language pairs.

We also trained a **transformer-based NER model** for Ancient Greek, benefiting from the annotated data available on Perseus Digital Library and manual annotations created by Chiara Palladino. Due to the characteristics of the training data, which contains a majority of single-token entities, the model achieved great results on single-token entity detection and classification. In contrast, the performance on multi-token entities was significantly lower. This motivated us to explore efficient ways to enrich the training data, such as cross-lingual annotation projection in order to enhance the performance of the NER model. Table 7.1 lists the language models we trained and published on Hugging Face.

	Huggingface Name	Task	Type	Base Model
1	UGARIT/grc_alignment	Alignment	Multilingual	XLM-R
2	UGARIT/flair_grc_multi_ner	NER	Multilingual	XLM-R
3	UGARIT/flair_grc_bert_ner	NER	Monolingual	BERT

Table 7.1: A list of language models hosted on `https://huggingface.co`.

Tool	**Description**
ugarit.ialigner.com	Ugarit **Manual** Alignment Editor
ugarit-aligner.com	Ugarit **automatic** translation alignment tool.
evalign.info	EvAlign framework for visual evaluation of alignment models
tldr.webis.de	The Summary Explorer
dynamiclexicon.com	Reading environment for diverse aligned texts
divan-hafez.com	Reading environment of Divan Hafez poetery parallel corpus

Table 7.2: A list of online tools.

Applications

In the context of the thesis, we have employed the alignment model we trained in two alignment-based tasks: cross-lingual annotation projection of named entities from English to Ancient Greek, Latin, and Arabic texts and inducing bilingual dictionaries from parallel corpora. Further, we employed alignment in another cross-lingual experiment

on modern languages to study and analyze the cross-lingual semantic transfer and frameshifts of two semantic primes in English German and English Spanish parallel corpus.

Evaluation

In collaboration with domain experts, we developed alignment guidelines and alignment gold standard datasets for three language pairs. The work on the fourth language pair (Ancient Greek-Persian) is still in progress. Further, we reviewed the benchmark datasets and the current evaluation procedure and proposed two evaluation metrics, namely, Phrase Alingment Accuracy, and Alignment Coverage.

Visualization

We have also designed and implemented an interactive annotation tool for manual alignment and developed reading environments and visualization approaches for parallel texts aligned at different levels. Further, we created two visual analytics tools, the Summary Explorer and EvAlign, to aid developers and practitioners to assess the quality of text summarization models and automatic alignment models, respectively. Table 7.2 provides an overview of the online tools produced during the work on this thesis.

Future Research Directions

Data

Collecting more data on Ugarit: The experiments we conducted in Chapter 3 have proven the great importance of supervised training using the word-level aligned texts from Ugarit. Our future plan revolves around creating more accurate data in a crowd-sourcing fashion. That can be done systematically in collaboration with teachers and scholars and by unifying the alignment guidelines, which will increase the consistency and homogeneity of the alignments.

Targeting new languages: This thesis focused on Ancient Greek in the first place and Latin and classical Persian in the second place. In the future, we will expand our focus to other languages, such as Coptic and Syriac. Working with these languages is challenging due to the lack of resources and the alphabets they use, which are not part of the vocabularies of the pretrained language models such as mBERT and XLM-RoBERTa. Which means we need to train a multilingual model from scratch.

Enhancing the dynamic lexicon: We intend to expand the parallel corpora used to extract the translation pairs and include more languages, in addition to updating the lexicon with the new manual translation pairs created on Ugarit with regular bases. Incorporating the entries of bilingual dictionaries such as *Liddell-Scott-Jones* Greek-English Lexicon and *Lewis-Short* Latin-English lexicon is also on the top of our list.

Models

As we mentioned in Chapter 5, we plan to develop a sentence-level alignment model for historical languages. Such a model is of great importance because it will help build accurate parallel corpora aligned at the sentence or paragraph level, increasing the amount of training data for word-level alignment models and resulting in more accurate models. Further, we will work to improve the performance of current models by experimenting with different training objectives and different underlying pretrained language models, such as XLNI and XLM-Align(53).

Applications

Drawing from the success of integrating the alignment model in other development pipelines, we intend to tackle new cross-lingual tasks such as word-sense disambiguation and cross-lingual summarization for historical languages, also training sentence embeddings model for historical languages, which can be incorporated to improve cross-lingual semantic similarity. Our ultimate goal is to build a platform that offers different linguistic services for historical languages and provides reading environment and annotation tool for cross-lingual alignment at different granularity levels (manual and automatic), morphological annotation such as named entities and part-of-speech, and the Dynamic Lexicon.

References

[1] Glosbe online dictionary. URL https://glosbe.com. https://glosbe.com, Accessed: 2020-04-10. 114, 122

[2] Kitab project. URL http://kitab-project.org/. http://kitab-project.org/, Accessed: 2020-04-10. 114, 118, 121, 128

[3] Linguee. URL https://www.linguee.com/. https://www.linguee.com/, Accessed: 2020-04-10. 114, 121

[4] Revesrsocontext contextual dictionary. URL https://context.reverso.net/. https://context.reverso.net/, Accessed: 2020-04-10. 114, 122

[5] Variance viewer. URL http://variance-viewer.informatik.uni-wuerzburg.de/. http://variance-viewer.informatik.uni-wuerzburg.de/, Accessed: 2020-04-10. 114, 121, 128

[6] A. Abdul-Rahman, G. Roe, M. Olsen, C. Gladstone, R. Whaling, N. Cronk, R. Morrissey, and M. Chen. Constructive visual analytics for text similarity detection. *Computer Graphics Forum*, 36(1):237–248, 2017. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12798. 114, 116, 117, 127, 128

[7] Judit Ács. Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1938–1942, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/864_Paper.pdf. ix, 22, 114, 126

[8] Lars Ahrenberg, Magnus Merkel, Anna Sågvall Hein, and Jörg Tiedemann. Evaluation of word alignment systems. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May 2000. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2000/pdf/137.pdf. 13, 63, 69

[9] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019. 94

[10] Fahad Al-Obaidli, Stephen Cox, and Preslav Nakov. Bi-text alignment of movie subtitles for spoken english-arabic statistical machine translation. *CoRR*, abs/1609.01188, 2016. URL http://arxiv.org/abs/1609.01188. 131

[11] Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6318. URL https://aclanthology.org/W18-6318. 46

[12] Bridget Almas and Marie-Claire Beaulieu. Developing a New Integrated Editing Platform for Source Documents in Classics. *Literary and Linguistic Computing*, 28(4):493–503, 07 2013. ISSN 0268-1145. doi: 10.1093/llc/fqt046. URL https://doi.org/10.1093/llc/fqt046. 13

[13] Bridget Almas and Monica Berti. Perseids collaborative platform for annotating text re-uses of fragmentary authors. 09 2013. 13, 114, 122, 123, 127, 128, 129

[14] Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3904–3919, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.329. URL https://aclanthology.org/2021.findings-emnlp.329. 46, 48

[15] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*, 2016. 92

[16] Tara L. Andrews and Caroline Macé. Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *LLC*, 28:504–521, 2013. 114, 125

[17] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL https://aclanthology.org/P18-1073. 47

[18] Ehsaneddin Asgari and Hinrich Schütze. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1011. URL https://aclanthology.org/D17-1011. 92

[19] Bharathi Asokarajan, Ronak Etemadpour, June Abbas, Samuel J. Huskey, and Chris Weaver. Textile: A pixel-based focus+context tool for analyzing variants across multiple text scales. In *Eurographics Conference on Visualization, Short Papers, Barcelona, Spain, 12-16 June 2017*, pages 49–53, 2017. viii, 114, 115

[20] Necip Fazil Ayan and Bonnie J. Dorr. Going beyond AER: An extensive analysis of word alignments and their impact on MT. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Sydney, Australia, July 2006. Association

REFERENCES

for Computational Linguistics. doi: 10.3115/1220175.1220177. URL https://aclanthology.org/P06-1002. 63, 69, 92

[21] J. Beel B. Gipp, N. Meuschke. Comparative evaluation of text- and citation-based plagiarism detection approaches using guttenplag. In *11th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2011. 114, 116, 119, 120

[22] Alison Babeu. *The Perseus Catalog: of FRBR, Finding Aids, Linked Data, and Open Greek and Latin*, pages 53–72. 08 2019. ISBN 9783110599572. doi: 10.1515/9783110599572-005. 16, 114, 118, 128

[23] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 46

[24] Mona Baker. Corpus linguistics and translation studies*: Implications and applications. In *Researching Translation in the Age of Technology and Global Conflict*, pages 9–24. Routledge, 2019. 1

[25] David Bamman. Building a dynamic lexicon from a digital library. pages 11–20, 01 2008. 114

[26] Elton Barker and Melissa Terras. Greek Literature, the Digital Humanities, and the Shifting Technologies of Reading. In *Oxford Handbook Topics in Classical Studies*. Oxford University Press. ISBN 9780199935390. doi: 10.1093/oxfordhb/9780199935390.013.45. URL https://doi.org/10.1093/oxfordhb/9780199935390.013.45. 93

[27] Anabela Barreiro, Francisco Raposo, and Tiago Luís. Clue-aligner: An alignment tool to annotate pairs of paraphrastic and translation units. In *10th Language Resources and Evaluation Conference (LREC)*, pages 7–13, 2016. 13, 64

[28] Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. *arXiv preprint arXiv:1811.01157*, 2018. 92

[29] Martin Baumann, Markus John, Hermann Pflüger, Cornelia Herberichs, Gabriel Viehhauser, Wolfgang Knopki, and Thomas Ertl. An Interactive Visualization for the Analysis of Annotated Text Variance in the Legendary Der Heiligen Leben, Redaktion. In *LEVIA 2019: Leipzig Symposium on Visualization in Applications*, 2019. 114

[30] Michael Behrisch, Miloš Krstajić, Tobias Schreck, and Daniel Keim. The news auditor: Visual exploration of clusters of stories. pages 61–65, 01 2012. 114

[31] Drayton Benner. A tool for a high-carat gold-standard word alignment. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 80–85, 2014. 12, 13

[32] M Berti. Digital fragmenta historicorum graecorum (dfhg), 2017. 54

[33] Monica Berti. Historical fragmentary texts in the digital age. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, pages 257–276. De Gruyter, Berlin, 2019. 67

[34] Monica Berti. *Digital Editions of Historical Fragmentary Texts*. Digital Classics Books 5. Propylaeum, Heidelberg, 2021. 67

[35] Chris Biemann. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, New York City, June 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-3812. 103

[36] Christopher Blackwell and Gregory Crane. *CONCLUSION: CYBERINFRASTRUCTURE, THE SCAIFE DIGITAL LIBRARY AND CLASSICS IN A DIGITAL AGE*, pages 375–446. Gorgias Press, Piscataway, NJ, USA, 2010. ISBN 9781463219222. doi: doi: 10.31826/9781463219222-017. URL https://doi.org/10.31826/9781463219222-017. 93

[37] Phil Blunsom and Trevor Cohn. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 65–72, Sydney, Australia, July 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220184. URL https://aclanthology.org/P06-1009. 44

[38] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017. 47

[39] Federico Boschetti, Matteo Romanello, Alison Babeu, and David Bamman. Improving ocr accuracy for classical critical editions. pages 156–167, 09 2009. 131

[40] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013. 127

[41] Peter F. Brown, John Cocke, Stephen A. Della-Pietra, Vincent J. Della-Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul Rossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85, 1990. 44

[42] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. URL https://aclanthology.org/J93-2003. 44, 48, 92

[43] Marco Büchler, Philip R. Burns, Martin Müller, Emily Franzini, and Greta Franzini. *Towards a Historical Text Re-use Detection*, pages 221–238. Springer International Publishing, Cham, 2014. ISBN 978-3-319-12655-5. URL https://doi.org/10.1007/978-3-319-12655-5_11. 112

[44] Manuel Burghardt. Usability recommendations for annotation tools. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 104–112, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-3613. 14

[45] Manuel Burghardt. *Engineering annotation usability-Toward usability patterns for linguistic annotation tools*. PhD thesis, 2014. 14

[46] John Burnet. *Platonis opera. T. 3: Tetralogias V - VII continens*. Scriptorum classicorum bibliotheca Oxoniensis. Clarendon, Oxonii, nachdr. edition, 1903. ISBN 9780198145424. URL http://data.perseus.org/texts/urn:cts:greekLit:tlg0059.tlg003. 67

[47] Chris Callison-Burch, David Talbot, and Miles Osborne. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 175–182, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1218978. URL https://aclanthology.org/P04-1023. 13, 44, 45, 92

[48] H. M. Caseli, V. D. Feltrim, and M. G. V. Nunes. Tagalign: Uma ferramenta de pré-processamento de textos (nilc-tr-02-09). Technical Report 169, 2002. URL http://www2.dc.ufscar.br/helenacaseli/pdf/2002/NILC-TR-02-09.pdf. 13

[49] T. Cheesman. Translation Sorting: Eddy and Viv in Translation Arrays. *B. Wiggin (ed.), Un/Translatables*, 19(1):121–142, 2012. 114

[50] Tom Cheesman, Stephan Thiel, Kevin Flanagan, Zhao Geng, Alison Ehrmann, Robert S. Laramee, Jonathan Hope, and David M. Berry. Translation arrays: Exploring cultural heritage texts across languages. In *Proceedings of the Digital Humanities 2012*, 2012. 114, 122, 128

[51] Chi Chen, Maosong Sun, and Yang Liu. Mask-align: Self-supervised neural word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4781–4791, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.369. URL https://aclanthology.org/2021.acl-long.369. 46, 67

[52] Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.42. URL https://aclanthology.org/2020.emnlp-main.42. 67

[53] Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, He-Yan Huang, and Furu Wei. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, 2021. 46, 48, 51, 52, 53, 54, 67, 75, 148

[54] David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pages 263–270, 2005. 92

[55] Christos Christodouloupoulos and Mark Steedman. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395, 2015. 97

[56] Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Japanese-chinese phrase alignment using common chinese characters information. In *In Proceedings of MT Summit XIII*. Citeseer, 2011. 114, 116, 129

[57] Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pages 612–620. PMLR, 2013. 111

[58] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20 (1):37–46, 1960. 64

[59] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019. 48, 53

[60] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017. 47, 48, 101

[61] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116. 48, 75

[62] Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1223. URL https://www.aclweb.org/anthology/D19-1223. 52

[63] Simon Corston-Oliver and Michael Gamon. Normalizing german and english inflectional morphology to improve statistical word alignment. In *Conference of the Association for Machine Translation in the Americas*, pages 48–57. Springer, 2004. 45

[64] Ido Dagan, Kenneth Church, and Willian Gale. Robust bilingual word alignment for machine aided translation. In *Natural Language Processing Using Very Large Corpora*, pages 209–224. Springer, 1999. 44

[65] Neo4j Graph Database. Graph modeling guidelines. http://neo4j.com/developer/guide-data-modeling, 2019. URL http://neo4j.com/developer/guide-data-modeling. Accessed 02 Nov. 2019. 26, 114, 121, 122, 127, 128, 129

[66] Yarowsky David, Ngai Grace, Wicentowski Richard, et al. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8, 2001. 92

[67] Ronald Dekker and Gregor Middell. Computer-supported collation with collatex. managing textual variance in an environment with varying requirements. In *Supporting Digital Humanities 2011*, November 2011. URL http://hnk.ffzg.hr/bibl/SDH-2011/submissions/sdh2011_submission_54.pdf. 114, 123, 124, 125, 127, 128, 130

REFERENCES

[68] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805. 47, 48, 53, 75

[69] Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, 2002. 92

[70] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 64

[71] Shuoyang Ding, Hainan Xu, and Philipp Koehn. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5201. URL https://aclanthology.org/W19-5201. 46, 67

[72] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. 47

[73] Garrett Dome. *Teaching Latin with Ugarit Alignment*. PhD thesis, Tufts University, 2020. 26

[74] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.181. 11, 46, 48, 52, 53, 67, 75, 106, 142, 145

[75] Philipp Dufter, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze. Embedding learning through multilingual concept induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1520–1530, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1141. URL https://aclanthology.org/P18-1141. 92

[76] Lucia Dwika and Klaus Schulz. Significant word-based text alignment for text reuse detection. 01 2018. 112

[77] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1073. 44

[78] Elias D Eells. *Multilevel Alignment of Iliadic Texts*. PhD thesis, Tufts University, 2021. 26

[79] Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. Challenges and solutions for latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan, dec 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/W16-4012. 93

[80] Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37, 2018. 92

[81] Manaal Faruqui and Shankar Kumar. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356, 2015. 92

[82] Anise Ferreira, Christopher Blackwell, and Chiara Palladino. Edições digitais nas clássicas: elementos de gênero na produção e leitura da língua grega e do latim. *Linha D'Água*, 33(2):113–135, 2020. 26

[83] Mark A Finlayson and Tomaž Erjavec. Overview of annotation creation: Processes and tools. In *Handbook of Linguistic Annotation*, pages 167–191. Springer, 2017. 14

[84] Pamela Forner, Jussi Karlgren, Christa Womser hacker (eds, Martin Potthast, Tim Gollub, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Alberto Barrón-cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. Overview of the 4th international competition on plagiarism detection, 2012. 112

[85] Alexander Fraser and Daniel Marcu. Squibs and discussions: Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303, 2007. doi: 10.1162/coli.2007.33.3.293. URL https://aclanthology.org/J07-3002. 63, 92

[86] Pascale Fung and Kenneth Ward Church. K-vec: A new approach for aligning parallel texts. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*, 1994. URL https://aclanthology.org/C94-2178. 44

[87] William A Gale and Kenneth Church. Identifying word correspondences in parallel texts. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991. 44

[88] William A. Gale and Kenneth Ward Church. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics (ACL)*, 1991. URL http://acl.ldc.upenn.edu/H/H91/H91-1026.pdf. 114, 121, 123, 127

[89] Zvi Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys (CSUR)*, 18(1):23–38, 1986. 53

[90] Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 369–377, 2009. 92

[91] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software engineering, testing, and quality assurance for natural language processing*, pages 49–57, 2008. 44

[92] Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1453. URL https://aclanthology.org/D19-1453. 46, 67

[93] Zhao Geng, Tom Cheesman, Robert S. Laramee, Kevin Flanagan, and Stephan Thiel. Shakervis: Visual analysis of segment variation of german translations of shakespeare's othello. *Information Visualization*, 2013. URL http://ivi.sagepub.com/content/early/2013/07/22/1473871613495845.abstract. 114, 127

[94] Ulrich Germann. Yawat: yet another word alignment tool. In *Proceedings of the ACL-08: HLT demo session*, pages 20–23, 2008. 13, 64

[95] Timur Gilmanov, Olga Scrivner, and Sandra Kübler. Swift aligner, a multifunctional tool for parallel corpora: Visualization, word alignment, and (morpho)-syntactic cross-language transfer. In *LREC*, pages 2913–2919, 2014. 11, 13, 64

[96] Meritxell Gonzàlez, Jesús Giménez, and Lluís Màrquez. A graphical interface for MT evaluation and error analysis. In *Proceedings of the ACL 2012 System Demonstrations*, pages 139–144, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-3024. 111

[97] João Graça, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. Building a golden collection of parallel multi-language word alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/250_paper.pdf. 11, 65

[98] Joao Graca, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. Building a golden collection of parallel multi-language word alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008. 66, 75

[99] Stephen Grimes, Xuansong Li, Ann Bies, S. Kulick, Xiaoyi Ma, and Stephanie Strassel. Creating arabic-english parallel word-aligned treebank corpora at ldc. 2010. 12, 13

[100] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, 2015. 92

[101] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. A representation learning framework for multi-source transfer parsing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2734–2740. AAAI Press, 2016. 92

[102] Nway Nway Han and Aye Thida. Annotated guidelines and building reference corpus for myanmar-english word alignment. *arXiv preprint arXiv:1909.11288*, 2019. 66

[103] Mark Harrower and Cynthia A Brewer. Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. 119

[104] Kazuma Hashimoto, Raffaella Buschiazzo, James Bradbury, Teresa Marshall, Richard Socher, and Caiming Xiong. A high-quality multilingual dataset for structured documentation translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 116–127, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5212. URL https://aclanthology.org/W19-5212. 92

[105] Nicolás Hinrichs, Maryam Foradi, Tariq Yousef, Elisa Hartmann, Susanne Triesch, Jan Kaßel, and Johannes Pein. Embodied metarepresentations. *Frontiers in neurorobotics*, 16, 2022. 7, 10, 107

[106] Anh Khoa Ngo Ho and François Yvon. Generative latent neural models for automatic word alignment. *arXiv preprint arXiv:2009.13117*, 2020. 67

[107] Maria Holmqvist and Lars Ahrenberg. A gold standard for English-Swedish word alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 106–113, Riga, Latvia, May 2011. Northern European Association for Language Technology (NEALT). URL https://aclanthology.org/W11-4615. 66

[108] Juliane House. *Translation: the basics*. Routledge, 2017. 1

[109] Jin-Xia Huang and Key-Sun Choi. Chinese-korean word alignment based on linguistic comparison. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 392–399, 2000. 45

[110] Matthias Huck, Diana Dutka, and Alexander Fraser. Cross-lingual annotation projection is effective for neural part-of-speech tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 223–233, Ann Arbor, Michigan, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1425. URL https://aclanthology.org/W19-1425. 92

[111] Nancy Ide. *Introduction: The Handbook of Linguistic Annotation*, pages 1–18. Springer Netherlands, Dordrecht, 2017. ISBN 978-94-024-0881-2. doi: 10.1007/978-94-024-0881-2_1. URL https://doi.org/10.1007/978-94-024-0881-2_1. 14

[112] Ayyoob ImaniGooghari, Masoud Jalili Sabet, Philipp Dufter, Michael Cysou, and Hinrich Schütze. ParCourE: A parallel corpus explorer for a massively multilingual corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and*

the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 63–72, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.8. URL https://aclanthology.org/2021.acl-demo.8. 92, 101, 111

[113] Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*, 2019. 92

[114] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.147. URL https://aclanthology.org/2020.findings-emnlp.147. 11, 46, 48, 52, 53, 67, 75, 142, 145

[115] S. Jänicke and A. Geßner. A Distant Reading Visualization for Variant Graphs. In *Conference Abstracts of the Digital Humanities 2015*, 2015. 114, 115, 128

[116] S. Jänicke and D. J. Wrisley. Interactive visual alignment of medieval text versions. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 127–138, 2017. viii, 114, 117, 121, 122, 127, 128, 129

[117] Stefan Jänicke, Thomas Efer, Marco Büchler, and Gerik Scheuermann. Designing Close and Distant Reading Visualizations for Text Re-use. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pages 153–171. Springer, 2014. 129

[118] Stefan Jänicke, Annette Geßner, Greta Franzini, Melissa Terras, Simon Mahony, and Gerik Scheuermann. TRAViz: A Visualization for Variant Graphs. *Digital Scholarship in the Humanities*, 30(suppl_1):i83–i99, 10 2015. 114, 125, 126, 127, 128

[119] Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.3. URL https://aclanthology.org/2021.acl-demo.3. 93

[120] Marcin Junczys-Dowmunt and Arkadiusz Szał. Symgiza++: symmetrized word alignment models for statistical machine translation. In *International Joint Conferences on Security and Intelligent Information Systems*, pages 379–390. Springer, 2012. 44

[121] Juxtacommons. A user guide to juxta commons, 2019. URL http://juxtacommons.org/guide. http://juxtacommons.org/guide, Accessed: 2019-09-10. 114, 117, 119, 121, 127, 128

[122] Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and Shrikanth Narayanan. Sailalign: Robust long speech-text alignment. In *Proc. of workshop on new tools and methods for very-large scale phonetics research*, 2011. 92

[123] Martin Kay and Martin Röscheisen. Text-translation Alignment. *Computational linguistics*, 19(1):121–142, 1993. 44

[124] Mir jalaleddin Kazzazi. *Iliad*. Markaz, 1998. 40

[125] Josh Kemp. Beyond translation: Building better greek scholars, Sep 2021. URL https://medium.com/pelagios/beyond-translation-building-better-greek-scholars-561ab331a1bc. 93

[126] Sue J Ker and Jason S Chang. A class-based approach to word alignment. *Computational linguistics*, 23(2): 313–343, 1997. 45

[127] Hisham A. Kholidy and N. Chatterjee. Towards developing an arabic word alignment annotation tool with some arabic alignment guidelines. In *2010 10th International Conference on Intelligent Systems Design and Applications*, pages 778–783, 2010. doi: 10.1109/ISDA.2010.5687168. 18

[128] Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 564–571, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1064. 92

[129] Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. Cross-lingual annotation projection for weakly-supervised relation extraction. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1):1–26, 2014. 92

[130] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005. 92, 105

[131] Philipp Koehn, Franz J Och, and Daniel Marcu. Statistical phrase-based translation. Technical report, University of Southern California Marina Del Rey Information Sciences Inst, 2003. 92

[132] Max Kölbl, Yuki Kyogoku, J Nathanael Philipp, Michael Richter, Clemens Rietdorf, and Tariq Yousef. The semantic level of shannon information: Are highly informative words good keywords? a study on german. In *Natural Language Processing in Artificial Intelligence—NLPinAI 2020*, pages 139–161. Springer, 2021. 141

[133] Mikhail Kozhevnikov and Ivan Titov. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, 2013. 92

[134] Ivana Kruijff-Korbayová, Klára Chvátalová, and Oana Postolache. Annotation guidelines for Czech-English word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/575_pdf.pdf. 66

[135] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. 51

[136] Murathan Kurfalı and Robert Östling. Noisy parallel corpus filtering through projected word embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 277–281, 2019. 92

[137] Patrik Lambert, Adria DE GISPERT, Rafael BANCHS, and José B MARINO. Guidelines for word alignment evaluation and manual alignment. *Language resources and evaluation*, 39(4):267–285, 2005. 66, 67, 69

[138] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017. 48

[139] Adrien Lardilleux, Julien Gosme, and Yves Lepage. Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *The Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 252–256. European Language Resources Association (ELRA), 2010. 70, 92

[140] Adrien Lardilleux, Yves Lepage, and François Yvon. The contribution of low frequencies to multilingual subsentential alignment: a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189–217, 2011. 44

[141] Brandon Laughlin, Christopher Collins, Karthik Sankaranarayanan, and Khalil El-Khatib. A visual analytics framework for adversarial text generation. In *2019 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 1–10. IEEE, 2019. 111

[142] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1027. URL https://aclanthology.org/P15-1027. 47

[143] Gyeongbok Lee, Sungdong Kim, and Seung-won Hwang. Qadiver: Interactive framework for diagnosing qa models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9861–9862, 2019. 111

[144] Jae-Hee Lee, Seung-Wook Lee, Gumwon Hong, Young-Sook Hwang, Sang-Bum Kim, and Hae-Chang Rim. A post-processing approach to statistical word alignment reflecting alignment tendency between part-of-speeches. In *Coling 2010: Posters*, pages 623–629, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-2071. 45

[145] Els Lefever, Véronique Hoste, and Martine De Cock. ParaSense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-2055. 92

[146] William D. Lewis and Fei Xia. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008. URL https://aclanthology.org/I08-2093. 92

[147] Jin-Ji Li, Dong-Il Kim, and Jong-Hyeok Lee. Annotation guidelines for Chinese-Korean word alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/137_paper.pdf. 18

[148] Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1124. URL https://aclanthology.org/P19-1124. 46

[149] Yanyang Li, Yingfeng Luo, Ye Lin, Quan Du, Huizhen Wang, Shujian Huang, Tong Xiao, and Jingbo Zhu. A simple and effective approach to robust unsupervised bilingual dictionary induction. *arXiv preprint arXiv:2011.14874*, 2020. 101

[150] Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June 2006. Association for Computational Linguistics. URL https://aclanthology.org/N06-1014. 44

[151] Shixia Liu, Michelle X Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):1–28, 2012. 111

[152] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692. 47

[153] Adam Lopez and Philip Resnik. Word-based alignment, phrase-based translation: What's the link? In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 90–99, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.11. 92

[154] Chenwei Lou, Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, Weiwei Tu, and Ruifeng Xu. Translation-based implicit annotation projection for zero-shot cross-lingual event argument extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2076–2081, 2022. 92

[155] Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065, 2020. 92

REFERENCES

[156] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://aclanthology.org/D15-1166. 46

[157] Yanjun Ma, Sylwia Ozdowska, Yanli Sun, and Andy Way. Improving word alignment using syntactic dependencies. Association for Computational Linguistics, 2008. 45

[158] Lieve Macken. An annotation scheme and gold standard for Dutch-English word alignment. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/100_Paper.pdf. 66

[159] E. C Marchant. *Xenophontis Opera omnia*. Clarendon Press, 1910. ISBN 9780198145523 9780198145530 9780198145547 9780198145561 9780198145554. URL http://www.perseus.tufts.edu/hopper/text?doc=urn:cts:greekLit:tlg0032.tlg007. OCLC: 802674413. 67

[160] Kelly Marchisio, Conghao Xiong, and Philipp Koehn. Embedding-enhanced giza++: Improving alignment in low-and high-resource scenarios using embedding space geometry. *arXiv preprint arXiv:2104.08721*, 2021. 46, 67

[161] Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. Bilingual lexicon induction for low-resource languages using graph matching via optimal transport. *arXiv preprint arXiv:2210.14378*, 2022. 92, 101

[162] David Mareček. Automatic alignment of tectogrammatical trees from czech-english parallel corpus. Master's thesis, Charles University, MFF UK, 2008. 18

[163] I. Dan Melamed. Manual annotation of translational equivalence: The blinker project. *CoRR*, cmp-lg/9805005, 1998. URL http://arxiv.org/abs/cmp-lg/9805005. 12, 63

[164] Antonio Valerio Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1614. URL https://aclanthology.org/W16-1614. 47

[165] Rada Mihalcea and Ted Pedersen. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, pages 1–10, 2003. 65, 66

[166] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 47

[167] Robert C. Moore. Improving ibm word alignment model 1. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 518–525, Barcelona, Spain, July 2004. URL http://acl.ldc.upenn.edu/acl2004/main/pdf/352_pdf_2-col.pdf. 45

[168] Robert C. Moore. A discriminative framework for bilingual word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 81–88, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1011. 44

[169] Mathias Müller. Treatment of markup in statistical machine translation. Association of Computational Linguistics, 2017. 92

[170] A.T. Murray. *The Iliad. With an English translation by A.T. Murray.* William Heinemann ; G.P. Putnam's Sons London (England) : New York (New York), 1924. URL http://data.perseus.org/texts/urn:cts:greekLit:tlg0012.tlg001. 67

[171] Saeed Nafisi. *Iliad*. Elmi Farhangi, 1958. 40

[172] Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.41. URL https://aclanthology.org/2020.emnlp-main.41. 45, 46, 48, 67

[173] Shahrzad Naseri, Jeffrey Dalton, Andrew Yates, and James Allan. Ceqe: Contextualized embeddings for query expansion. In *European Conference on Information Retrieval*, pages 467–482. Springer, 2021. 48

[174] Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4007. URL https://aclanthology.org/N19-4007. 92

[175] Mariana Neves and Ulf Leser. A survey on annotation tools for the biomedical literature. *Briefings in bioinformatics*, 15(2):327–340, 2014. 14

[176] Anh Khoa Ngo Ho and François Yvon. Optimizing word alignments with better subword tokenization. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 256–269, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-research.21. 67

[177] Jian Ni, Georgiana Dinu, and Radu Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, 2017. 92

[178] Elisa Nury. Visualizing collation results. *Variants*, pages 75–94, 03 2019. 114, 123

[179] Franz Josef Och and Hermann Ney. A comparison of aligment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-ACL 2000)*, 2000. 66

[180] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. doi: 10.1162/089120103321337421. URL https://aclanthology.org/J03-1002. 44, 56, 65, 68, 69

[181] Franz Josef Och, Christoph Tillmann, and Hermann Ney. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999. URL https://aclanthology.org/W99-0604. 92

[182] Robert Östling and Jörg Tiedemann. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146, October 2016. URL http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf. 44

[183] Sebastian Padó and Mirella Lapata. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340, 2009. 92

[184] Chiara Palladino. Reading texts in digital environments: Applications of translation alignment for classical language learning. *The Journal of Interactive Technology and Pedagogy*. 23

[185] Chiara Palladino. Reading texts in digital environments: Applications of translation alignment for classical language learning. *J. Interact. Technol. Pedagog*, 18:724–731, 2020. 26

[186] Chiara Palladino and Tariq Yousef. We want to learn all languages! digital classicist seminar london, 2021. URL https://youtu.be/R2Ms6yAMZss. Accessed 27 Oct. 2021. 24

[187] Chiara Palladino and Tariq Yousef. To say almost the same thing? A study on cross-linguistic variation in ancient texts and their translations. *Digital Scholarship in the Humanities*, 01 2023. ISSN 2055-7671. doi: 10.1093/llc/fqac086. URL https://doi.org/10.1093/llc/fqac086. fqac086. 5, 10, 12, 32

[188] Chiara Palladino, Farimah Karimi, and Brigitte Mathiak. Ner on ancient greek with minimal annotation. In *https://dh2020. adho. org/*. DH2020, 2020. 93

[189] Chiara Palladino, Anna Zhang, Maryam Foradi, and Tariq Yousef. How to read all languages: Translation alignment with ugarit. In *DH*, 2020. 5, 10, 12, 25

[190] Chiara Palladino, Maryam Foradi, and Tariq Yousef. Translation alignment for historical language learning: a case study. *DHQ*, 15(3), 2021. URL http://www.digitalhumanities.org/dhq/vol/15/3/000563/000563.html. 2, 5, 10, 11, 12, 23, 26

[191] Chiara Palladino, Farnoosh Shamsian, and Tariq Yousef. Using parallel corpora to evaluate translations of ancient greek literary texts. 06 2022. 5, 10

[192] Chiara Palladino, Farnoosh Shamsian, and Tariq Yousef. Using parallel corpora to evaluate translations of ancient greek literary texts. an application of text alignment for digital philology research. *Journal of Computational Literary Studies*, 1(1), 2022. 12, 42

[193] Chiara Palladino, Tariq Yousef, Farnoosh Shamsian, and Nadia Kanagawa. Ugarit: Translation alignment technologies for under-resourced languages. In *DIGITAL HUMANITIES 2022 Book of Abstracts*, pages 55–57, Tokyo, Japan, 07 2022. 5, 10, 12

[194] Chiara Palladino, Tariq Yousef, Farnoosh Shamsian, and Nadia Kanagawa. How to read all languages: Translation alignment with ugarit. In *DH*, 2022. 25

[195] Richard B Parkinson, Whitfield Diffie, Mary Fischer, and RS Simpson. *Cracking codes: the Rosetta Stone and decipherment*. Univ of California Press, 1999. 1

[196] Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. *arXiv preprint arXiv:1908.06625*, 2019. 92, 101

[197] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 47

[198] Laura Perez-Beltrachini and Mirella Lapata. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.742. URL https://aclanthology.org/2021.emnlp-main.742. 134

[199] Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, 2019. 52

[200] ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12, 1802. 47

[201] Matthew Petrillo and Jessica Baycroft. Introduction to manual annotation. *Fairview research*, 2010. 16

[202] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019. 53

[203] M. Pöckelmann, A. Medek, P. Molitor, and J. Ritter. Catview - supporting the investigation of text genesis of large manuscripts by an overall interactive visualization tool. In *Proceedings of the Digital Humanities 2015*, 2015. 114, 127

[204] M. Pöckelmann, A. Medek, P. Molitor, and J. Ritter. Catview - supporting the investigation of text genesis of large manuscripts by an overall interactive visualization tool. In *Proceedings of the Digital Humanities 2015*, 2015. 128

REFERENCES

[205] Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. Multimirror: Neural cross-lingual word alignment for multilingual word sense disambiguation. In *IJCAI*, pages 3915–3921, 2021. 92

[206] Sixta Quassdorf. Hyperhamlet: A database of quotations from and allusions to shakespeare's most famous tragedy. October 2019. 130

[207] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics*, 23(1):61–70, 2016. 111

[208] Randy L. Ribler and Marc Abrams. Using visualization to detect plagiarism in computer science classes. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 173–, Washington, DC, USA, 2000. IEEE Computer Society. ISBN 0-7695-0804-9. URL http://dl.acm.org/citation.cfm?id=857190.857699. 114, 115

[209] Michael Richter, Maria Bardají I Farré, Max Kölbl, Yuki Kyogoku, J Nathanael Philipp, Tariq Yousef, Gerhard Heyer, and Nikolaus P Himmelmann. Uniform density in linguistic information derived from dependency structures. In *ICAART (1)*, pages 496–503, 2022. 141

[210] Patrick Riehmann, Henning Gruendl, Martin Potthast, Martin Trenkmann, Benno Stein, and Bernd Froehlich. Wordgraph: Keyword-in-context visualization for netspeak's wildcard search. *IEEE transactions on visualization and computer graphics*, 18, 03 2012. 114, 125

[211] Roberto Rosselli Del Turco and Chiara Di Pietro. Between innovation and conservation: The narrow path of user interface design for digital scholarly editions. pages 133–163, 12 2018. 114, 119, 121, 127, 128

[212] Desmond Schmidt and Robert Colomb. A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, 67:497–514, 06 2009. 125

[213] Bettina Schrader. How does morphological complexity translate? a cross-linguistic case study for word alignment. In *Proceedings of Linguistic Evidence Conference*, pages 189–191, 2006. 92

[214] Susan Schreibman. The versioning machine. *Literary and Linguistic Computing*, 18:101–107, 04 2003. 114, 121, 127, 128

[215] S. Schütz and M. Pöckelmann. Lera - explorative analyse komplexer textvarianten in editionsphilologie und diskursanalyse. In *Proceedings of the Digital Humanities im deutschsprachigen Raum, DHd 2016*, 2016. 114, 121, 127, 128

[216] Farnoosh Shamsian. Digital classics and learning greek in iran, sunoikisis digital classics, thursday may 14, 2020. URL https://youtu.be/R2Ms6yAMZss. Accessed 27 Oct. 2021. 24

[217] Bo Shao, Yeyun Gong, Weizhen Qi, Nan Duan, and Xiaola Lin. Multi-level alignment pretraining for multilingual semantic parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3246–3256, 2020. 92

[218] Haoyue Shi, Luke Zettlemoyer, and Sida I Wang. Bilingual lexicon induction via unsupervised bitext construction and word alignment. *arXiv preprint arXiv:2101.00148*, 2021. 101, 102

[219] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*, pages 364–371. Elsevier, 2003. 79, 129, 135

[220] Maia Shukhoshvili. Methodology of translation alignment of georgian text of plato's "theaetetus". *Int. J. Lang. Linguist*, 4:63–69, 2017. 26

[221] Carson Sievert and Kenneth Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3110. URL https://aclanthology.org/W14-3110. 111

[222] Pranaydeep Singh, Gorik Rutten, and Els Lefever. A pilot study for bert language modelling and morphological analysis for ancient and medieval greek. In *The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021)*, 2021. 94

[223] Neel Smith and Christopher Blackwell. Ducat citation alignment 1.3.0. URL https://eumaeus.github.io/uva_cex_ducat/. https://eumaeus.github.io/uva_cex_ducat/, Accessed: 2020-04-10. 114, 122

[224] Benjamin Snyder, Tahira Naseem, and Regina Barzilay. Unsupervised multilingual grammar induction. Association for Computational Linguistics, 2009. 92

[225] Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-4014. 92

[226] David Steele and Lucia Specia. Wa-continuum: Visualising word alignments across multiple parallel sentences simultaneously. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 121–126, 2015. 111

[227] David Steele and Lucia Specia. Vis-eval metric viewer: A visualisation tool for inspecting and evaluating metric scores of machine translation output. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 71–75, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-5015. URL https://aclanthology.org/N18-5015. 111

[228] Steinór Steingrímsson, Hrafn Loftsson, and Andy Way. CombAlign: a tool for obtaining high-quality word alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 64–73, Reykjavik, Iceland (Online), May 31–2 June 2021. Linköping University Electronic Press, Sweden. URL https://aclanthology.org/2021.nodalida-main.7. 46, 66, 67

[229] Steinór Steingrímsson, Hrafn Loftsson, and Andy Way. Pivotalign: Leveraging high-precision word alignments for bilingual dictionary inference. *Proc. of LDK*, 2021. 101

[230] Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. A discriminative neural model for cross-lingual word alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 910–920, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1084. URL https://aclanthology.org/D19-1084. 45, 46, 48, 67

[231] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. S eq 2s eq-v is: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363, 2018. 111

[232] Donald Sturgeon. Unsupervised identification of text reuse in early Chinese literature. *Digital Scholarship in the Humanities*, 33(3):670–684, 11 2017. ISSN 2055-7671. URL https://doi.org/10.1093/llc/fqx024. 114, 119, 126, 130

[233] Sara Stymne, Jörg Tiedemann, and Joakim Nivre. Estimating word alignment quality for SMT reordering tasks. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 275–286, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3334. URL https://aclanthology.org/W14-3334. 63, 92

[234] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 46

[235] Shahbaz Syed, Tariq Yousef, Khalid Al Khatib, Stefan Jänicke, and Martin Potthast. Summary explorer: Visualizing the state of the art in text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 185–194, 2021. 8, 9, 111, 119

[236] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. *arXiv preprint arXiv:2008.05122*, 2020. 111

[237] Arda Tezcan and Vincent Vandeghinste. SMT-CAT integration in a technical domain: Handling XML markup using pre & post-processing methods. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*, Leuven, Belgium, May 30–31 2011. European Association for Machine Translation. URL https://aclanthology.org/2011.eamt-1.10. 92

[238] Maxat Tezekbayev, Vassilina Nikoulina, Matthias Gallé, and Zhenisbek Assylbekov. Speeding up entmax. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1142–1158, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.86. URL https://aclanthology.org/2022.findings-naacl.86. 52

[239] Brian Thompson and Philipp Koehn. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1136. URL https://aclanthology.org/D19-1136. 11

[240] Jörg Tiedemann. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1175. 92

[241] Jörg Tiedemann. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 191–199, Vilnius, Lithuania, May 2015. Linköping University Electronic Press, Sweden. URL https://aclanthology.org/W15-1824. 92

[242] Jörg Tiedemann. Improved sentence alignment for movie subtitles. In *In Proceedings of RANLP, Borovets*, 2007. 131

[243] Andreas Tsiartas, Prasanta Kumar Ghosh, Panayiotis G. Georgiou, and Shrikanth Narayanan. Context-driven automatic bilingual movie subtitle alignment. In *Proceedings of Interspeech 2009*, Brighton, UK, September 2009. 131

[244] Dan Tufiş, Radu Ion, and Nancy Ide. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1312–1318, 2004. 92

[245] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 47, 93

[246] Jean Véronis and Philippe Langlais. Evaluation of parallel text alignment systems. In *Parallel text processing*, pages 369–388. Springer, 2000. 63

[247] Stephan Vogel. PESA: Phrase pair extraction as sentence splitting. In *Proceedings of Machine Translation Summit X: Papers*, pages 251–258, Phuket, Thailand, September 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.33. 44

[248] Stephan Vogel, Hermann Ney, and Christoph Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, 1996. URL http://acl.ldc.upenn.edu/C/C96/C96-2141.pdf. 44

[249] Yile Wang, Leyang Cui, and Yue Zhang. Improving skip-gram embeddings using bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1318–1328, 2021. doi: 10.1109/TASLP.2021.3065201. 48

REFERENCES

[250] M. Wattenberg and F. B. Viégas. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, Nov 2008. ISSN 1941-0506. 114, 125

[251] M. Wattenberg and F. B. Viégas. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, Nov 2008. ISSN 1941-0506. 114, 121, 123, 127

[252] D. Wemhoener, I. Z. Yalniz, and R. Manmatha. Creating an improved version using noisy ocr from multiple editions. In *2013 12th International Conference on Document Analysis and Recognition*, pages 160–164, 2013. 131

[253] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019. 48

[254] Christian Werner. *Homero, Ilíada*. Ubu Editora, 1ª edição edition, November 2018. ISBN 978-85-928869-2-9. 67

[255] Anna Wierzbicka. *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK, 1996. 105

[256] W.P. Williams and C.S. Abbott. *An Introduction to Bibliographical and Textual Studies*. Modern Language Association of America, 1985. ISBN 9780873521338. URL https://books.google.de/books?id=WQzhAAAAMAAJ. 112

[257] Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1779–1785, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1187. URL https://aclanthology.org/D14-1187. 92

[258] Yuk Wah Wong and Raymond Mooney. Learning for semantic parsing with statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 439–446, New York City, USA, June 2006. Association for Computational Linguistics. URL https://aclanthology.org/N06-1056. 92

[259] Di Wu, Liang Ding, Shuo Yang, and Mingyang Li. MirrorAlign: A super lightweight unsupervised word alignment model via cross-lingual contrastive learning. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 83–91, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.8. URL https://aclanthology.org/2022.iwslt-1.8. 46, 67

[260] Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. Lome: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, 2021. 106, 142

[261] Min Xiao and Yuhong Guo. Annotation projection-based representation learning for cross-lingual dependency parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 73–82, 2015. 92

[262] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104. URL https://aclanthology.org/N15-1104. 47

[263] Minhan Xu and Yu Hong. Improving word alignment with contextualized embedding and bilingual dictionary. In *CCF Conference on Big Data*, pages 180–194. Springer, 2020. 67

[264] Ismet Yalniz and R. Manmatha. A fast alignment scheme for automatic ocr evaluation of books. pages 754 – 758, 10 2011. 131

[265] Tariq Yousef and Monica Berti. The digital fragmenta historicorum graecorum and the ancient greek-latin dynamic lexicon. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH) 2015*, Warsaw, Poland, December 2015. Institute of Computer Science, Polish Academy of Science. 22, 114, 116, 117, 121, 127, 134

[266] Tariq Yousef and Maryam Foradi. Word alignment of divan-e-hafez. URL http://divan-hafez.com/. http://divan-hafez.com/, Accessed: 2020-04-10. 114, 122, 129

[267] Tariq Yousef and Stefan Janicke. A survey of text alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1149–1159, 2021. doi: 10.1109/TVCG.2020.3028975. 7, 9, 19, 20, 79, 111

[268] Tariq Yousef and Stefan Jänicke. Visual Evaluation of Translation Alignment Data. In Marco Agus, Wolfgang Aigner, and Thomas Hoellt, editors, *EuroVis 2022 - Short Papers*. The Eurographics Association, 2022. ISBN 978-3-03868-184-7. doi: 10.2312/evs.20221101. 9, 64

[269] Tariq Yousef, Chiara Palladino, and Gregory Crane. Intra-language text alignment using ialigner. In *Proceedings, 7th International Conference of Digital Archives and Digital Humanities*. National Taiwan University, 12 2016. 114

[270] Tariq Yousef, Chiara Palladino, and Stefan Jänicke. Transformer-based named entity recognition for ancient greek. 11 2022. doi: 10.13140/RG.2.2.34846.61761. 7, 9

[271] Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise dâ€™Orange Ferreira, and Michel Ferreira dos Reis. An automatic model and gold standard for translation alignment of ancient greek. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5894–5905, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.634. 6, 9, 11, 66, 94

[272] Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, and Maryam Foradi. Translation alignment with ugarit. *Information*, 13(2), 2022. ISSN 2078-2489. doi: 10.3390/info13020065. URL https://www.mdpi.com/2078-2489/13/2/65. 5, 9, 12, 64

[273] Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. Automatic translation alignment for ancient greek and latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lt4hala2022-1.14. 6, 9, 66, 94

[274] Tariq Yousef, Gerhard Heyer, and Stefan Jänicke. EVALIGN: Visual evaluation of translation alignment models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 277–297, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-demo.31. 9, 64

[275] Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. Named entity annotation projection applied to classical languages. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 175–182, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.latechclfl-1.19. 9

[276] Ieva Zariņa, Pēteris ikiforovs, and Raivis Skadiņš. Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 185–192, 2015. 92

[277] Thomas Zenkel, Joern Wuebker, and John DeNero. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*, 2019. 46

[278] Thomas Zenkel, Joern Wuebker, and John DeNero. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.146. URL https://aclanthology.org/2020.acl-main.146. 46, 67

[279] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179. URL https://aclanthology.org/P17-1179. 47

[280] Yi Zhang, Dingding Wang, and Tao Li. idvs: an interactive multi-document visual summarization system. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 569–584. Springer, 2011. 111

[281] Vilém Zouhar and Daria Pylypenko. Leveraging neural machine translation for word alignment. *arXiv preprint arXiv:2103.17250*, 2021. 67

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, 13.06.2023

(Unterschrift)