Named Entity Annotation Projection Applied to Classical Languages

Tariq Yousef * **Chiara Palladino** † **Gerhard Heyer** * **Stefan Jänicke**‡

*Leipzig University †Furman University ‡University of Southern Denmark

<tariq.yousef@uni-leipzig.de>

Abstract

In this study, we demonstrate how to apply cross-lingual annotation projection to transfer named-entity annotations to classical languages for which limited or no resources and annotated texts are available, aiming to enrich their NER training datasets and train a model to perform NER tagging. Our approach employs sentence-level aligned corpora of ancient texts and the translation in a modern language, for which high-quality off-the-shelf NER systems are available. We automatically annotate the text of the modern language and employ a state-of-the-art neural word alignment system to find translation equivalents. Finally, we transfer the annotations to the corresponding tokens in the ancient texts using a direct projection heuristic. We applied our method to ancient Greek and Latin using the Bible with the English translation as a parallel corpus. We used the resulting annotations to enhance the performance of an existing NER model for ancient Greek.

1 Introduction

Named Entity Recognition (NER), like other NLP tasks, has benefited from the advances in language modeling and the availability of large annotated corpora. Numerous high-quality NER models are available for modern languages. However, classical and ancient languages lack adequate annotated data and language models essential for training NER models. Therefore, annotation projection can be employed over parallel text corpora to overcome this problem and transfer the annotation from modern languages for which accurate off-the-shelf NER systems are available.

The core concept of annotation projection is to perform automatic or manual linguistics annotation on a text and project the annotation to its translation using mapping heuristics that link the entities with their correspondences. Translation alignment has been used for this purpose to transfer various linguistic annotations, such as Semantic Role labels (Padó and Lapata, 2009; Kozhevnikov and Titov, 2013), Part-of-Speech (Huck et al., 2019; Tiedemann, 2014; Wisniewski et al., 2014), Named Entities tags (David et al., 2001; Ni et al., 2017; Jain et al., 2019), Relations and Arguments (Kim et al., 2010, 2014; Faruqui and Kumar, 2015; Lou et al., 2022), Semantic Parsing (Shao et al., 2020; Hinrichs et al., 2022), Syntactic and Dependency parsing (Xiao and Guo, 2015; Guo et al., 2015; Tiedemann, 2015). Recently, neural translation alignment models were able to produce accurate alignments for a variety of modern and classical languages, even with no or a small amount of training data profiting from contextualized multilingual language models.

In this paper, we present a processing pipeline to transfer NE annotations from a text in modern languages to parallel texts in classical or low-resourced languages. We use accurate NER models for modern languages and employ state-of-the-art neural alignment models at the word level to find the translation equivalents. Further, we propose a direct projection heuristic that maps the annotations from source to target tokens considering various alignment types. We used the obtained entities to improve the accuracy of existing NER models for ancient Greek. The proposed approach works for any language pair provided the parallel corpora are available and aligned at the sentence or paragraph level.

2 Related Work

Cross-lingual annotation projection of named entities in a parallel corpus has two main scenarios: The first scenario incorporates machine translation to translate the detected named entities of the source text and tries to look up the translated entities in the corresponding parallel sentence using various matching heuristics based on orthographic and phonetic similarity and edit distance text sim-

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023), pages 175–182
May 5, 2023 ©2023 Association for Computational Linguistics

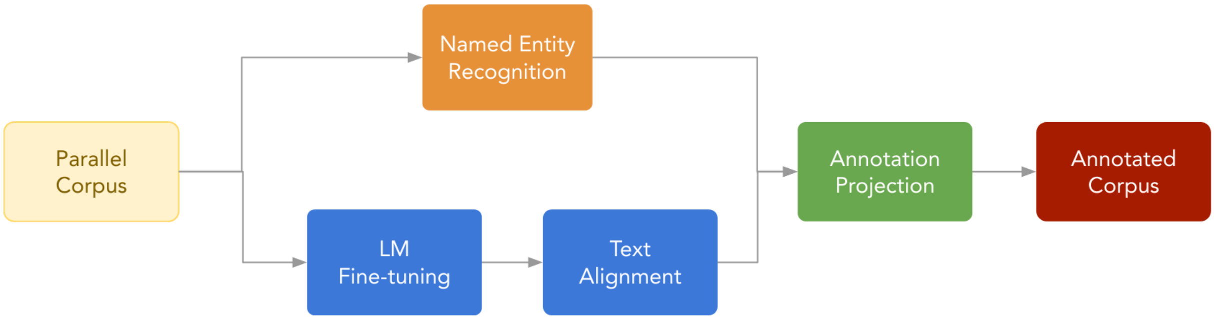

Figure 1: An overview of the proposed pipeline.

ilarity (Ehrmann et al., 2011; Jain et al., 2019). The second scenario employs automatic word alignment to find the translation equivalents of the detected entities in the parallel sentence and project the annotation (Ni et al., 2017; Agerri et al., 2018).
On NER in Digital Classics: The Classical Language Toolkit (CLTK) is the largest Python library to perform NLP tasks on ancient languages, including NER (Johnson et al., 2021). However, the lack of adequately annotated datasets for most classical languages is a fundamental hindrance to the high performance of this task. Other efforts have been made, starting from large annotated datasets of specific sources, using semantic annotation platforms and Machine Learning (Berti, 2019). Yousef et al. (2023b) trained a transformer-based NER for ancient Greek; however, the model performed poorly on multi-token entities since the training data used in the training process is composed of single-token entities.

3 Methodology

Figure 1 illustrates the 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] (Gardner et al., 2017), or *flairNLP*[3] (Akbik et al., 2019) 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 equivalents. An unsupervised fine-tuning using the parallel corpus can be employed using different training objectives to improve the word alignment accuracy. Subsequently, we find the corresponding translations 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, and English.

3.1 Corpus Collection

The Bible is an ideal source of parallel texts and is available in several modern and ancient languages. 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[4] to build our parallel corpus. The repository includes translations of the Bible in over 100 languages (Christodouloupoulos and Steedman, 2015). For our experiment, we used ancient Greek and Latinwith the English translation. Every verse has a unique ID that encapsulates the information of the book, chapter, and verse; 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*[5]. We followed the same naming convention to assign verse IDs.

[1] https://spacy.io/
[2] https://allenai.org/allennlp
[3] https://github.com/flairNLP/flair
[4] https://github.com/christos-c/bible-corpus
[5] https://scaife.perseus.org/library/urn:cts:greekLit:tlg0031/

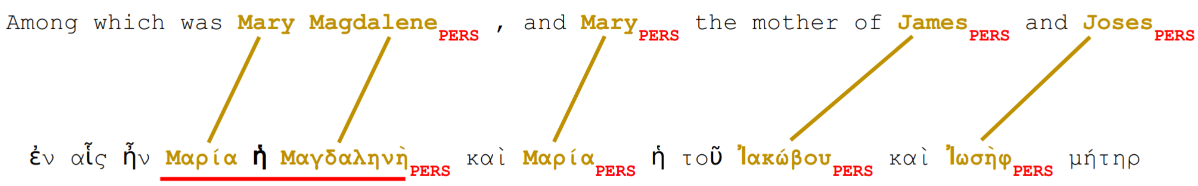

Figure 2: Annotation projection example.

3.2 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 classical 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[6]. The comparison revealed that *AllenNLP* and *flairNLP* significantly outperformed *spaCy*, and their performance was so close (Figure 3). In this study, we used *AllenNLP* NER tagger with four entity classes (PERS, LOC, ORG, MISC).

3.3 Automatic Translation Alignment

Translation alignment aims to link words/tokens in the source text with their correspondences in the translation. With the recent advances in multilingual transformer models and neural machine translation, a new era of alignment models has begun. Neural models, which significantly outperformed the statistical models, achieved state-of-the-art performance on a variety of language pairs (Zenkel et al., 2020; Jalili Sabet et al., 2020; Dou and Neubig, 2021; Garg et al., 2019; Chen et al., 2021), including ancient languages (Yousef et al., 2022a,c).

In our experiment, we employed an automatic alignment workflow that utilizes cross-lingual semantic similarity among tokens based on contextualized embeddings derived from multilingual language models such as mBERT (Devlin et al., 2018) or XLM-RoBERTa (Conneau et al., 2019) and derive the word-level alignments from the obtained similarity matrix using various heuristics auch as ARGMAX, ITERMAX (Jalili Sabet et al., 2020), SOFTMAX, ENTMAX (Dou and Neubig, 2021), and OPTIMAL TRANSPORT (OS) (Chi et al., 2021).

Various training objectives can be employed to fine-tune the language model supervised and unsupervised in order to enhance the cross-lingual transfer of the word embeddings and improve the alignment accuracy consequently. For instance, Translation Language Modeling (TLM) (Conneau and Lample, 2019), Self-training (SO) and Parallel Sentence Identification (PSI) objectives (Dou and Neubig, 2021),and Denoising Word Alignment (DWA) (Chi et al., 2021).

We trained a multilingual language model[7] that performed well on ancient Greek, Latin, and English language pairs (Yousef et al., 2022a). We fine-tuned XLM-RoBERTa unsupervised with a large corpus of parallel sentences in ancient languages and supervised with manually aligned translation pairs extracted from Ugarit database (Yousef et al., 2022b). We employed this language model in our experiment to derive word embeddings, COSINE SIMILARITY as a similarity measure, and ITERMAX as an alignment extraction heuristic since it achieved the best Phrase Alignment Accuracy (PAC) with large margin (Yousef et al., 2023a) and the second lowest Alignment Error Rate (AER) (Yousef et al., 2022a) on the ancient Greek-English dataset.

3.4 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. Suppose we have a sentence pair $S = \{s_1, s_2, \cdots, s_n\}$ and its translation $T = \{t_1, t_2, \cdots, t_m\}$. S is already NER annotated and $E = \{(s_k, Loc), (\{s_j, s_{j+1}\}, Pers) \cdots\}$ is the set of detected entities, and S, T are already aligned at word level and $A = \{(s_i, t_j), \cdots, (s_n, t_m)\}$ is the set of translation pairs. In order to project the

[6] More information is available in the appendix

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

annotations from S to T, we followed a simple mapping heuristic that considers the different alignment types:

When the entity $e(s_i, Cat) \in E$, whether a single- or multi-token entity, is aligned to:

- a single token t_j $(s_i, t_j) \in T$ (one-to-one or many-to-one alignments). We assign t_j the same category as the source entity (t_j, Cat). For instance, *Mary*-Μαρία, *James*-Ἰακώβου, and *Joses*-Ἰωσῆφ in Figure 2.
- multiple tokens $\{(s_i, t_j), (s_i, t_k)\} \subset T$ (one-to-many or many-to-many alignments), in this case, if the corresponding tokens are consecutive $|j - k| = 1$, they will be considered as one multi-token entity $(\{t_j, t_k\}, Cat)$. Otherwise, we annotate the range of tokens from j to k as one entity $(\{t_j, \cdots, t_k\}, Cat)$. However, if $|j - k| > 2$, we create two separate entities (t_j, Cat) and (t_k, Cat). For instance, *Mary Magdalene* and Μαρία ἡ Μαγδαληνὴ in Figure 2.
- $NULL$, i.e. the entity has no correspondence in the target language (one-to-null or many-to-null alignments), then no projection is required.

4 Results

We employed the projection approach to the 7950 verses of the new testament and resulted in 6,567 ancient Greek entities (6,104 single-token and 463 multi-token entities) and 6481 Latin entities (5940 single-token an 541 multi-token).
We performed qualitative evaluation to estimate the quality of the produced annotations on two language pairs: English-Ancient Greek and English-Latin. Two domain experts manually assessed 100 random verses, which corresponded to about 550 extracted entities per language, and assigned a score to each detected entities. Table 1 summarizes the evaluation results: The performance on Ancient Greek achieved the highest accuracy (86.63%) followed by Latin (82.34%).

The automatic NER annotation of English text achieved over 94% accuracy and the entities alignment on Ancient Greek-English achieved the highest accuracy (91.9%), since the alignment model is optimized for this language pair. However, the entities classification errors were common for personal names classified as locations and vice versa. In some cases, a Greek or Latin 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)especially in multi-token entities such as "Jesus Christ", "Simon Zelotes", or "Pontius Pilate".

Further, we used the resulted annotations to extend the available NER training dataset for ancient Greek[8] and fine-tune the existing ancient Greek NER models proposed by Yousef et al. (2023b)[9]. The obtained model achieved a higher F1, and a better performance on multi-token entities as reported in Table 2.

5 Conclusion

In this experiment, we used translation alignment to project NER annotations 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. The proposed approach can be employed to 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 tool 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 more accurate than the English-Latin since the language model used in the experiment is mainly fine-tuned on Ancient Greek texts. This experiment is a proof of concept, and due to limited computational resources, we used a subset of the Bible corpus (New Testament only). Using the entire corpus with other parallel corpora will result in more named entities and accurate NER models.

[8] https://scaife.perseus.org/reader/urn:cts:greekLit:tlg0008.tlg001.perseus-grc4

[9] https://huggingface.co/UGARIT/flair_grc_bert_ner

Score	Ancient Greek	Latin
correct alignment / correct NER	86.63%	82.34%
incorrect alignment / correct NER	7.26%	12.87%
correct alignment / incorrect NER	5.28%	3.96%
incorrect alignment / incorrect NER	0.83%	0.83%

Table 1: Manual evaluation of 100 randomly selected verses.

6 Limitations

The proposed approach requires accurate parallel corpora to achieve good results. Further, it 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 automatic NER tagger because if it can not detect the entity, it will not be projected. Replacing one or both automatic components with manual annotation or alignment would significantly enhance performance. Another obstacle is that multilingual language models do not support all languages and alphabets. We tested the projection approach on Coptic and Syriac translations of the Bible, and the results were terrible. The alignment workflow failed to generate accurate alignments since the language model we used to derive the embeddings is fine-tuned XLM-RoBERTa model, whose vocabulary is limited and does not support Coptic and Syriac alphabets.

References

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Monica Berti. 2019. Named entity annotation for ancient greek with inception. In *CLARIN Annual Conference Proceedings*, pages 1–4, Leipzig. CLARIN.

Chi Chen, Maosong Sun, and Yang Liu. 2021. 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. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. 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, Online. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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

Yarowsky David, Ngai Grace, Wicentowski Richard, et al. 2001. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zi-Yi Dou and Graham Neubig. 2021. 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. Association for Computational Linguistics.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages

118–124, Hissar, Bulgaria. Association for Computational Linguistics.

Manaal Faruqui and Shankar Kumar. 2015. 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.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. 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. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. 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.

Nicolás Hinrichs, Maryam Foradi, Tariq Yousef, Elisa Hartmann, Susanne Triesch, Jan Kaßel, and Johannes Pein. 2022. Embodied metarepresentations. *Frontiers in neurorobotics*, 16:836799.

Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. 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. Association for Computational Linguistics.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. 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. Association for Computational Linguistics.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. 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. Association for Computational Linguistics.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2010. 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. Coling 2010 Organizing Committee.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2014. Cross-lingual annotation projection for weakly-supervised relation extraction. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1):1–26.

Mikhail Kozhevnikov and Ivan Titov. 2013. 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.

Chenwei Lou, Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, Weiwei Tu, and Ruifeng Xu. 2022. 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.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. 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.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Bo Shao, Yeyun Gong, Weizhen Qi, Nan Duan, and Xiaola Lin. 2020. Multi-level alignment pretraining for multi-lingual semantic parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3246–3256.

Jörg Tiedemann. 2014. 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. Dublin City University and Association for Computational Linguistics.

Jörg Tiedemann. 2015. 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. Linköping University Electronic Press, Sweden.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. 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. Association for Computational Linguistics.

Min Xiao and Yuhong Guo. 2015. Annotation projection-based representation learning for cross-lingual dependency parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 73–82.

Tariq Yousef, Gerhard Heyer, and Stefan Jänicke. 2023a. 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*.

Tariq Yousef, Chiara Palladino, and Stefan Jänicke. 2023b. Transformer-based named entity recognition for ancient greek. *The Book of Abstracts of DH2023*.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise D'Orange Ferreira, and Michel Ferreira dos Reis. 2022a. 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. European Language Resources Association.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, and Maryam Foradi. 2022b. Translation alignment with ugarit. *Information*, 13(2).

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022c. 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. European Language Resources Association.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. 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. Association for Computational Linguistics.

A Appendix

NER models benchmarking

For benchmarking, we used spaCy with *en_core_web_lg* model, flairNLP[10] with *ner-english-large* model, and AllenNLP with *tagging-elmo-crf-tagger*. We annotated 100 random verses of the Bible and evaluated the results manually. The precision of the three models was close, but regarding the Recall, spaCy underperformed the other models significantly. From 7937 verses of the new testament, spaCy detected 3,788 entities, AllenNLP 6,403 entities, and flairNLP 6,883 entities. This explains why spaCy achieved low Recall.

Automatic Text Alignment

***Embeddings*:** We used *UGARIT/grc-alignment*[11] language model as source of embedddings and Cosine similarity to create the similarity matrix.

Alignment Extraction: We used *Itermax* (Jalili Sabet et al., 2020) to extract the translation pairs from the similarity matrix. We used the code as it is provided by authors[12].

Fine-tuning: To fine-tune the language models we used the training objectives proposed by Dou and Neubig. The code for fine-tuning is available on the authors Github repository[13].

NER model Trainig

To train a NER model for ancient Greek with the results of the annotation projection process, we used Flair framework[14] (Akbik et al., 2019) and $pranaydeeps/Ancient-Greek-BERT$ language model using 75% of the data for training, 12.5% for testing, and 12.5% as development dataset. We trained the models 10 epochs and used Conditional Random Field (CRF) for prediction. The size of the training dataset is (18,276 PERS, 6,655 MISC, 3,415 LOC, and 61 ORG)

[10] https://github.com/flairNLP/flair

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

[12] https://github.com/cisnlp/simalign

[13] https://github.com/neulab/awesome-align

[14] https://github.com/flairNLP/

		Our Model			UGARIT/flair_grc_bert_ner		
		Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Traing	PER	92.87 %	94.31 %	93.59 %	91.24%	94.45%	92.82%
	MISC	84.49 %	82.32 %	83.39 %	80.92%	83.17%	82.03%
	LOC	82.99 %	82.99 %	82.99 %	86.86%	78.35%	82.38%

Table 2: Training results.

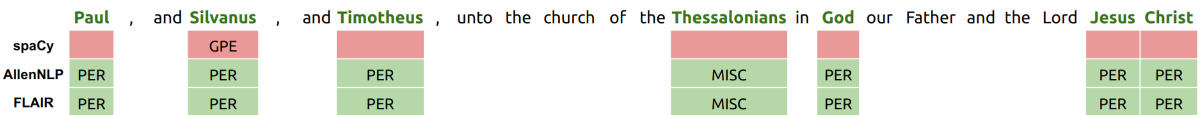

Figure 3: A performance comparison between three STOA NER models on biblical text (1 Thessalonians 1:1).