Findings of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages

Oksana Dereza[1]★, **Adrian Doyle**[1]★, **Priya Rani**[1]★,
Atul Kr. Ojha[1], **Pádraic Moran**[2], **John P. McCrae**[1]

[1] Insight SFI Centre for Data Analytics, Data Science Institute, University of Galway, Ireland
[2] Classics, University of Galway, Ireland

★firstname.lastname@insight-centre.org,
atulkumar.ojha@insight-centre.org,
padraic.moran@universityofgalway.ie,
john.mccrae@insight-centre.org

Abstract

This paper discusses the organisation and findings of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages. The shared task was split into the constrained and unconstrained tracks and involved solving either three or five problems for 12+ ancient and historical languages belonging to four language families and making use of six different scripts.

There were 14 registrations in total, of which three teams participated in each track. Out of these six submissions, two systems were successful in the constrained setting and another two in the unconstrained setting, and four system description papers were submitted by different teams.

The best average results for POS-tagging, lemmatisation and morphological feature prediction were 96.09%, 94.88% and 96.68% respectively. In the mask filling problem, the winning team could not achieve a higher average score across all 16 languages than 5.95% at the word level, which demonstrates the difficulty of this problem. At the character level, the best average result over 16 languages was 55.62%.

1 Introduction

The importance of NLP for studies in the classics is growing, as can be seen by the variety of technologies, digital text resources, and applications being developed to support research tasks in this field in recent years (Hawk et al., 2018; Neidorf et al., 2019; Stifter et al., 2021; Johnson et al., 2021). As the value of machine learning for historical linguistics is becoming more apparent, academic interest in word embedding models for use in these contexts is also increasing (Bamman and Burns, 2020; Singh et al., 2021; Hu et al., 2021; Riemenschneider and Frank, 2023; Dereza et al., 2023b).

Since the rise of word embeddings, their evaluation has been considered a challenging task that sparked considerable debate regarding the optimal approach. The two major strategies that researchers have developed over the years are intrinsic and extrinsic evaluation. The first amounts to solving specially designed problems like semantic proportions, or comparing the similarity of machine-generated words or sentences against human-generated examples. The second one focuses on solving downstream NLP tasks, such as sentiment analysis or question answering, probing word or sentence representations in real-world applications.

In recent years, sets of downstream tasks called benchmarks have become a very popular, if not default, method to evaluate general-purpose word and sentence embeddings. Despite the general trend towards multilinguality and ever-growing attention to under-resourced languages, ancient and historical languages remain under-served by embedding evaluation benchmarks, and the goal of this shared task is to bridge this gap. We argue that there is a need for a universal multilingual evaluation benchmark for embeddings learned from ancient and historical language data and view this shared task as a proving ground for it.

2 Related work

Starting with decaNLP (McCann et al., 2018) and SentEval (Conneau and Kiela, 2018), general-purpose multitask benchmarks for Natural Language Understanding (NLU) have become increasingly common in the literature, and new ones are reported regularly (Wang et al., 2019, 2020;

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 160–172

Shavrina et al., 2020; Xu et al., 2020; Kurihara et al., 2022; Urbizu et al., 2022; Berdicevskis et al., 2023). However, even the largest multilingual benchmarks, such as XGLUE, XTREME, XTREME-R or XTREME-UP (Hu et al., 2020; Liang et al., 2020; Ruder et al., 2021, 2023), only include modern languages.

The EvaLatin evaluation campaign (Sprugnoli et al., 2020, 2022) attracted some embedding-based solutions for POS-tagging, lemmatisation, and morphological feature prediction challenges (Wróbel and Nowak, 2022; Mercelis and Keersmaekers, 2022), but it did not specifically focus on embedding evaluation. Moreover, it was confined to Latin, which is the English of the ancient world in terms of language resources and technologies available. Individual scholars focusing on Latin and Ancient Greek mostly adopt Large Language Models (LLMs) together with their evaluation techniques through downstream tasks (Bamman and Burns, 2020; Singh et al., 2021; Yamshchikov et al., 2022; Riemenschneider and Frank, 2023; Krahn et al., 2023), while those working with less-resourced languages tend to translate intrinsic evaluation datasets from modern languages or create their own diagnostic tests (Tian et al., 2021; Hu et al., 2021; Dereza et al., 2023a). However, this is not a universal rule: the latest paper on distributional semantic models of Ancient Greek proposes a new dataset for intrinsic evaluation, AGREE (Stopponi et al., 2024), while some recent papers featuring medieval French and Spanish adopt transformer models and test them on Named Entity Recognition (Grobol et al., 2022; Torres Aguilar, 2022).

3 Setup and Schedule

For the purposes of our evaluation, languages are distinguished in accordance with ISO 639-3 codes[1] except for Latin, which was manually separated as discussed in Section 4. As a result, different historical stages of Irish and Latin are treated as distinct 'languages' in this paper. Such a distinction may be linguistically arbitrary, at least in the case of certain texts. However, as Universal Dependencies (UD) (Zeman et al., 2023) corpora are separated in accordance with ISO 639 codes, and the majority of data used in this evaluation was drawn from this resource, the same system for distinguishing languages was utilised here.

[1] https://iso639-3.sil.org

The Shared Task involved three problems (hereafter also referred as 'challenges' and 'downstream tasks') for 13 languages in the constrained setting and five problems for 16 languages in the unconstrained setting. These languages belong to four language families and use six different scripts (see Table 1 for detailed information).

3.1 Subtasks

A. Constrained
   1. POS-tagging
   2. Lemmatisation
   3. Morphological feature prediction

B. Unconstrained
   1. POS-tagging
   2. Lemmatisation
   3. Morphological feature prediction
   4. Filling the gaps (mask filling)
      a. Word-level
      b. Character-level

3.2 Timeline

The final timeline of the shared task is as follows.

05 Nov 2023: Release of training & validation data
02 Jan 2024: Release of test data
15 Jan 2024: System submission
22 Jan 2024: Paper submission
29 Jan 2024: Notification of acceptance
05 Feb 2024: Camera-ready submission

A tokenisation error was identified in the test data for problem 4a and in the Classical Chinese test data for problem 4b after the test data had been released. It was promptly corrected on 12 Jan 2024.

4 Data

For problems 1-3, data from Universal Dependencies v.2.12 (Zeman et al., 2023) was used for 11 ancient and historical languages, omitting corpora which contained fewer than 1,000 tokens or for which only a test set was available. Old Hungarian texts, annotated to the same standard as UD corpora, were added to the dataset from the MGTSZ website[2] (HAS Research Institute for Linguistics, 2018; Simon, 2014). Old Hungarian data was edited to simplify complex punctuation marks

[2] http://oldhungariancorpus.nytud.hu/en-codices.html

Language	Code	Script	Dating	Train-T	Valid-T	Test-T	Train-S	Valid-S	Test-S
Ancient Greek ♣	grc	Greek	800 BCE – 110 CE	334,043	41,905	41,046	24,800	3,100	3,101
Ancient Hebrew ♢	hbo	Hebrew	900 – 999 CE	40,244	4,862	4,801	1,263	158	158
Classical Chinese ♠	lzh	Hanzi	47 – 220 CE	346,778	43,067	43,323	68,991	8,624	8,624
Coptic ♢	cop	Coptic	0 – 199 CE	57,493	7,282	7,558	1,730	216	217
Gothic ♣	got	Latin	400 – 799 CE	44,044	5,724	5,568	4,320	540	541
Medieval Icelandic ♣	isl	Latin	1150 – 1680 CE	473,478	59,002	58,242	21,820	2,728	2,728
Classical & Late Latin ♣	lat	Latin	100 BCE – 399 CE	188,149	23,279	23,344	16,769	2,096	2,097
Medieval Latin ♣	latm	Latin	774 – early 1300s CE	599,255	75,079	74,351	30,176	3,772	3,773
Old Church Slavonic ♣	chu	Cyrillic	900 – 1099 CE	159,368	19,779	19,696	18,102	2,263	2,263
Old East Slavic ♣	orv	Cyrillic	1025 – 1700 CE	250,833	31,078	32,318	24,788	3,098	3,099
Old French ♣	fro	Latin	1180 CE	38,460	4,764	4,870	3,113	389	390
Vedic Sanskrit ♣	san	Latin (transcr.)	1500 – 600 BCE	21,786	2,729	2,602	3,197	400	400
Old Hungarian ♡	ohu	Latin	1440 – 1521 CE	129,454	16,138	16,116	21,346	2,668	2,669
Old Irish ♣	sga	Latin	600 – 900 CE	88,774	11,093	11,048	8,748	1,093	1,094
Middle Irish ♣	mga	Latin	900 – 1200 CE	251,684	31,748	31,292	14,308	1,789	1,789
Early Modern Irish ♣	ghc	Latin	1200 – 1700 CE	673,449	115,163	79,600	24,440	3,055	3,056

Table 1: Language families: ♣ – Indo-European, ♢ – Afro-Asiatic, ♠ – Sino-Tibetan, ♡ – Finno-Ugric. The 'Code' column refers to an ISO 639-3 code with the exception of Medieval Latin. The 'Script' column refers to the scripts used in the dataset rather than the script(s) typical for a particular language. The 'Dating' column describes the period when texts in the dataset were created, not when a particular language existed, cited according to the electronic editions/corpora these texts come from. Finally, we provide the size of each subset in sentences (S) and tokens (T).

masked	src
Cé [MASK] secht [MASK] im gin sóee suilgind, co bráth, mó cech delmaimm, issed ma do-ruirminn.	Cé betis secht tengtha im gin sóee suilgind, co bráth, mó cech delmaimm, issed ma do-ruirminn.

Table 2: An example of training data for word-level gap filling (problem 4a).

masked	src
Cé betis se[_]ht te[_]gtha im gin s[_]ee suilgind, co bráth, mó cech[_]delmaimm, isse[_] ma do-ruirminn.	Cé betis secht tengtha im gin sóee suilgind, co bráth, mó cech delmaimm, issed ma do-ruirminn.

Table 3: An example of training data for character-level gap filling (problem 4b).

used to approximate manuscript symbols. Tokens which were POS-tagged `PUNCT` were altered so that the form matched the lemma. Otherwise, no characters intended to approximate orthographic manuscript features were changed.

As the ISO 639-3 standard does not distinguish between historical stages of Latin, as it does between other languages like Irish, but it was desirable to approximate this distinction for Latin, we further split Latin data. This resulted in two Latin datasets; Classical and Late Latin, and Medieval Latin. This split was dictated by the composition of the Perseus (Celano et al., 2014) and PROIEL (Haug and Jøhndal, 2008) treebanks. As the Late Latin *Vulgata* is mixed with the work of Classical Latin authors in these treebanks, it was unfeasible to separate Classical Latin from Late Latin, though this may have been preferable. For the purposes of this evaluation we use the ISO 639-3 code `lat` for the Classical and Late Latin dataset, and we apply the faux-code, `latm`, to Medieval Latin.

Historical forms of Irish were only included in mask filling challenges, as the quantity of historical Irish text data which has been tokenised and annotated to a single standard to date is insufficient for the purpose of training models to perform morphological analysis tasks. The Irish texts for problem 4

were drawn from CELT[3] (Ó Corráin et al., 1997), Corpas Stairiúil na Gaeilge[4] (Acadamh Ríoga na hÉireann, 2017), and digital editions of the St. Gall glosses[5] (Bauer et al., 2017) and the Würzburg glosses[6] (Doyle, 2018). This provides a good case study of how performance may vary across different historical stages of the same language. Each Irish text taken from CELT is labelled 'Old', 'Middle' or 'Early Modern' in accordance with the language labels provided in CELT metadata. Because CELT metadata relating to language stages and text dating is reliant on information provided by a variety of different editors of earlier print editions, this metadata can be inconsistent across the corpus and on occasion inaccurate. To mitigate complications arising from this, texts drawn from CELT were included in the dataset only if they had a single Irish language label and if the dates provided in CELT metadata for the text match the expected dates for the given period in the history of the Irish language.

The upper temporal boundary was set at 1700 CE, and texts created later than this date were not included in the dataset. The choice of this date is driven by the fact that most of the historical language data used in word embedding research dates back to the 18^{th} century CE or later, and we would like to focus on the more challenging and yet unaddressed data. A detailed list of text sources for each language in the dataset is provided on our GitHub.[7]

The resulting datasets for each language were then shuffled at the sentence level and split into training, validation and test subsets at the ratio of $0.8 : 0.1 : 0.1$. Table 1 provides an overview of the data: language family, script, dating, and the size of each subset in sentences and tokens.

For word-level mask filling (problem 4a), 10% of tokens in each sentence were randomly replaced with a `[MASK]` token. Masked Language Models (MLMs) conventionally mask 15% of tokens, and Wettig et al. (2023) showed that an even higher masking rate could be beneficial for models the size of BERT-large.[8] However, our dataset is substantially smaller; moreover, sentences from historical texts are often much shorter than in modern language due to their genre or purpose (e.g. glosses, annals, charters etc.) For these reasons, it was unfeasible to set the masking rate higher than 10% for the benchmark presented in this paper, particularly for the smallest datasets.

For character-level gap filling (problem 4b), sentences were split into individual characters for languages with alphabetical writing systems. For Classical Chinese, each Hanzi character was decomposed into individual strokes with the help of `hanzipy`[9] package with the deepest decomposition level available, 'graphical'. Then, 5% of characters in each sentence were randomly replaced with a `[_]` token.

There were no restrictions on masked word/character position, and they could also be consecutive. Some sentences could have more than one masked word or character, and some (shorter) ones could have none.

For problems 1-3, participants received the data in CONLL-U format.[10] The data for tasks 4a and 4b was released in `tsv` format, as shown in Tables 2 and 3.

After the end of the competition an updated version of the dataset, including test labels, was published on Zenodo[11] (Dereza, 2024).

5 Evaluation

The shared task was hosted on CodaLab[12] and will remain available for post-competition submissions for anyone who would be interested in testing their approach on our data.

Our evaluation script calculates a score for each problem in the task (POS-tagging, lemmatisation etc.) per language with the metrics listed in Table 4. Following the authors of GLUE and SuperGLUE (Wang et al., 2019, 2020), we weigh each downstream task equally and provide a macro-average of per-problem scores as an overall score for a language. These scores are then averaged by CodaLab and displayed on the leaderboard as Rank.

[3] https://celt.ucc.ie/publishd.html

[4] http://corpas.ria.ie/index.php

[5] http://www.stgallpriscian.ie/

[6] https://wuerzburg.ie/

[7] https://github.com/sigtyp/ST2024/blob/main/list_of_text_sources.md

[8] BERT (Devlin et al., 2019) was pretrained on BookCorpus, a dataset consisting of 11,038 unpublished books, and English Wikipedia, which contains 6,780,526 articles as of February 2024: https://huggingface.co/bert-large-uncased

[9] https://github.com/Synkied/hanzipy

[10] https://universaldependencies.org/format.html

[11] https://doi.org/10.5281/zenodo.10655061

[12] Unconstrained track: https://codalab.lisn.upsaclay.fr/competitions/16818
Constrained track: https://codalab.lisn.upsaclay.fr/competitions/16822

As is common in evaluation benchmarks (Wang et al., 2020; Hu et al., 2020; Ruder et al., 2021), we use multiple metrics for every problem (e.g. F1 and Accuracy @1 for POS-tagging) except for morphological annotation. This helps to smooth out shortcomings that individual metrics may have and to make the evaluation scenario more forgiving for complicated problems (e.g. combining Accuracy @1 and Accuracy @3 for lemmatisation). Accuracy @1 is usually referred to as simply 'accuracy' and calculated as a ratio of correct predictions to all predictions. While Accuracy @1 verifies if the top prediction is correct or not, Accuracy @3 is a milder metric that checks if the correct answer is among top-3 predictions.

In the case of morphological annotation, we calculate a macro-average of Accuracy @1 per tag, and also introduce punishment for predicting incorrect features. For example, if a token should only have two morphological features, and a system predicts the correct value for one, but the incorrect value for the other, and then also suggests a feature that this token should not have at all, the score achieved for this token will be $1 + 0 - 1 = 0$.

The evaluation scripts for both constrained and unconstrained tracks are available on the Shared Task GitHub.[13]

Task	Metrics
POS-tagging	Acc@1, F1
Detailed morphological annotation	Macro-average of Acc@1 per tag
Lemmatisation	Acc@1, Acc@3
Filling the gaps (word-level)	Acc@1, Acc@3
Filling the gaps (character-level)	Acc@1, Acc@3

Table 4: Evaluation metrics.

6 Baseline Models

Baselines were provided for the three challenges which are shared by both the constrained and unconstrained tracks. As the aim of this Shared Task was to provide a benchmark for embedding models, multi-layer perceptron network models were developed to classify token data for each of the three challenges. For the sake of ensuring simplicity across the baseline models and results, model design and input data format was kept as similar as possible across all challenges. Slight variation was tolerated, however, depending on the requirements of each specific challenge.

Specific models were trained for each of the 13 languages for both the POS-tagging and lemmatisation challenges. By contrast, the approach taken for the morphological annotation challenge was to train a language-agnostic model for each of the 44 morphological features used across all languages in the dataset. This was found to produce better results than using language specific models, particularly for morphological features which were not common across all languages in the dataset. It also reduced model training time, as the alternative would have been to create a discrete model for each feature in use by each individual language, resulting in significantly more models.

Early stopping was applied during training of all models to avoid overfitting. Validation loss was used as a metric to determine when early stopping should be applied for POS-tagger and morphological feature analysis models. However, tracking validation accuracy instead was found to produce better results when training lemmatiser models. All POS-tagger and morphological feature analysis models used 64 neurons per hidden layer, as did lemmatiser models for smaller datasets, however, for languages with larger datasets this was found to be insufficient. To avoid hampering performance, lemmatiser models were created with up to 1024 neurons per hidden layer, depending on the size of the dataset.

Aside from the areas of divergence just mentioned, the design aspects common to all models are as follows:

- Hidden layers: 2
- Activation: ReLU
- Dropout: 20%
- Optimiser: Adam (Kingma and Ba, 2015)

6.1 Data Preparation

Text data was pre-processed before being used in model training for each of the three challenges. Feature engineering was carried out on the input data to ensure models would focus on the most valuable information to inform morphological analysis. For each token which would be used as input data across all three challenges, the following information was extracted:

1. The token itself (entirely in lower case letters)
2. The length of the sentence in which the token occurs (number of tokens)

[13] https://github.com/sigtyp/ST2024/

3. The length of the token itself (number of letter characters)
4. Whether the token occurred first in the sentence (Boolean: true or false)
5. Whether the token occurred last in the sentence (Boolean: true or false)
6. Whether the first letter of the token was capitalised (Boolean: true or false)
7. Whether the entire token was in all caps (Boolean: true or false)
8. Whether the entire token was in all lowercase (Boolean: true or false)
9. The first letter of the token
10. The second letter of the token
11. The third letter of the token
12. The last letter of the token
13. The second last letter of the token
14. The third last letter of the token
15. The previous token (entirely in lower case)
16. The following token (entirely in lower case)

In addition to the information listed above, language codes were also extracted for the morphological annotation challenge. This was necessary because the models themselves were not trained on individual languages for this particular challenge, but language information would nevertheless be useful in identifying morphological features. Once this information had been generated for each token, it was compiled and vectorised so that it could be used as input data in model training and validation.

POS-tags and lemmata associated with each token were extracted from the training and validation sets on a language-by-language basis. They were then encoded and set aside to be used as labels during model training. Generating label data for morphological annotation models was more complicated. First, the training data for all languages was combined, as was the validation data for all languages. Next, morphological features associated with each token across all of the combined languages were extracted. If a particular morphological feature was not used by a given token, a value of '_' was generated to indicate non-use. In the case of features common across many languages, this resulted in relatively balanced training and validation datasets. However, for uncommon features this could result in less than 1% of labels having values other than '_'. This would result models simply learning to classify every token as '_' for that feature. To overcome this issue, if more than 80% of labels in any training or validation set had the value '_', the size of the dataset was reduced by dropping random instances of '_' values until at least 20% of the dataset had labels with other values. Finally, these were encoded for model training.

7 Submitted Systems

There were 14 registrations in total, of which three teams submitted to each track. Out of these six submissions, two systems were successful in the constrained setting and another two in the unconstrained setting, and four system description papers were submitted by different teams.

We expected that participants would use the same pre-training technique for every problem, as is common in benchmarking, but the winning teams applied different pre-training approaches to different problems. At the same time, all participants leveraged various transformer architectures, with RoBERTa (Liu et al., 2019) and its modifications being the most popular one.

While all participants outperformed our baselines for morphological feature prediction with the best average result about 96% across 13 languages, only the winning teams beat the baselines for POS-tagging and lemmatisation, achieving average results of 95.25% and 93.67% respectively in the constrained setting, and 96.09% and 94.88% in the unconstrained setting. Baselines were not provided for the mask filling problems which formed a part of the unconstrained track only. At the word level, the winning team could not achieve a higher average accuracy across all 16 languages than 5.95%, with the best result for an individual language being 16.9% for Medieval Icelandic. This outlines the particular difficulty of this specific problem. At the character level, the best average result over 16 languages was 55.62% and the best result for an individual language was 74.59% for Gothic.

The combined results of the constrained and unconstrained settings for problems 1-3 are provided in Table 5. Table 6 shows results for problem 4 from the unconstrained track. Finally, average results across all problems for each track can be found in Table 7. These tables are provided in the Appendix A.

7.1 Constrained Setting

For the constrained subtask, participants were not allowed to use anything apart from the provided datasets, but they could reduce and balance them

if they saw fit. Our intention was to avoid any cross-lingual transfer in the constrained setting, including the transfer between the languages within the provided dataset. However, we seem to have failed to communicate this properly, and one of the systems submitted to this track made use of cross-lingual transfer within the dataset. Nevertheless, the system with embeddings pre-trained for each language individually achieved a better result.

7.1.1 Heidelberg-Boston

The winning team in the constrained track, representing Heidelberg University and Sattler College, submitted a system that uses a combination of contextual word and character embeddings pre-trained from scratch for each language in the dataset individually[14] (Riemenschneider and Krahn, 2024). Bringing together the hierarchical tokenisation method (Sun et al., 2023) and the DeBERTa-V3 architecture (He et al., 2023) for POS-tagging and morphological feature prediction, and using character-level nanoT5 models (Nawrot, 2023) for lemmatisation allowed the team to be on par with the winners of the unconstrained track, achieving the average score of 95.25%, 93.67% and 96.18% across 13 languages for POS-tagging, lemmatisation and morphological feature prediction respectively.

7.1.2 Team 21a

The team representing Allen Institute for Artificial Intelligence pretrained a multilingual transformer model, LiBERTus,[15] that follows RoBERTa's pretraining architecture (Liu et al., 2019) and takes inspiration from Conneau et al. (2020) regarding the scaling of BERT models to multiple languages. The authors point out that their model struggles with multiword expressions in Coptic and Ancient Hebrew (Miranda, 2024), which most likely refers to composite characters and vowel markings. Despite the use of cross-lingual transfer, the model's average score falls about 10% behind that of the winning team, reaching the average of 82.47%, 81.98% and 90.70% across 13 languages for POS-tagging, lemmatisation and morphological feature prediction respectively.

7.2 Unconstrained Setting

For the unconstrained subtask, participants could use any additional data in any language, including pre-trained embeddings and LLMs. Surprisingly, the winning team did not make use of embeddings at all in problem 4b, although this shared task was specifically dedicated to embedding evaluation. Still, we accepted this submission in full as the variety of approaches the team tried may be insightful for the reader.

7.2.1 UDParse

The winner of the unconstrained track is the UDParse team from Orange Innovation. To solve problems 1-3, the team trained their own UDParse parser[16] with the use of openly available contextualised embeddings: multilingual mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020) and GPT2 (Radford et al., 2019), and language-specific slavicBERT (Arkhipov et al., 2019) for Old Church Slavonic and Old East Slavic and heBERT (Chriqui and Yahav, 2022) for Ancient Hebrew. The team used distilBERT (Sanh et al., 2019) for word-level mask filling and an embedding-less n-gram based model for character-level mask filling (Heinecke, 2024). They achieve the average score of 96.09%, 86.47% and 96.68% across 13 languages for POS-tagging, lemmatisation and morphological feature prediction respectively. Their average results across 16 languages for word-level and character-level mask-filling are 3.77% and 55.62% respectively.

7.2.2 TartuNLP

The TartuNLP team from the University of Tartu submitted a system based on the adapters framework (Poth et al., 2023) that uses parameter-efficient fine-tuning (Dorkin and Sirts, 2024). They applied the same approach uniformly to all tasks and 16 languages by fine-tuning stacked language- and task-specific adapters for XLM-RoBERTa.[17] Although their system, achieving the average of 85.67% and 88.14% across 13 languages in POS-tagging and morphological feature prediction, is outperformed by UDParse, this is probably explained by the effectiveness of the UDParse morphological parser rather than by the quality of embeddings employed by either team. At the

[14] https://github.com/bowphs/SIGTYP-2024-hierarchical-transformers

[15] https://github.com/ljvmiranda921/LiBERTus

[16] https://github.com/Orange-OpenSource/udparse

[17] https://github.com/slowwavesleep/ancient-lang-adapters/tree/sigtyp2024

same time, TartuNLP outperforms UDParse in lemmatisation by 8.41%, achieving 94.88% on average across 13 languages, and in word-level mask filling, achieving 5.95% on average across 16 languages. The team's results for character-level mask filling generally concede 10-15% to the winner, which highlights an interesting observation: a very simple character-based n-gram model can be more effective in a low-resource setting than cutting edge approaches.

8 Discussion

Analysing results of the competition, we made a few interesting observations. First of all, data scarcity does have an effect on sequence labelling tasks, such as POS-tagging and morphological feature prediction, but this effect is not as dramatic as one might expect. Thus, the difference between the smallest corpus of 21K tokens (Vedic Sanskrit) and the biggest corpus of 599K tokens (Medieval Latin) is only 9.5% on average for POS-tagging and 11.3% for morphological feature prediction. The same is true for lemmatisation; however, models trained for this task seem to be more susceptible to orthographic variation and lexical variety in the data, as well as to the morphological complexity of a language. Thus, we see poorer results for lemmatisation across all languages despite the milder metrics.

Cross-lingual and cross-temporal (i.e. from modern languages to their ancestors) transfer could have played an important role in the systems that used XLM-RoBERTa. However, Riemenschneider and Krahn (2024) showed that similar results can be achieved with pre-training on modestly sized monolingual data without any transfer.

Mask filling tasks appeared to be much harder than we expected even for SOTA models. The problem could be attributable to the following reasons, or to some combination thereof:

- High lexical variety
- Orthographic variation
- Relatively short sentences
- Code-switching (e.g. Latin in historical Irish texts)
- Data scarcity (mask filling requires more training data than, for example, POS-tagging)
- Composite characters and vowel markings in Coptic and Ancient Hebrew
- Non-trivial character decomposition in Classical Chinese

9 Conclusion

The Shared Task on Word Embedding Evaluation for Ancient and Historical Languages attracted participants from five major research institutions and was an important step towards creating a universal multilingual evaluation benchmark for embeddings learned from ancient and historical language data. The best average results across 13 languages for POS-tagging, lemmatisation and morphological feature prediction were 96.09%, 94.88% and 96.68% respectively. However, participants only managed to achieve an average of 5.95% at word-level and 55.62% at character-level across 16 languages in more challenging mask filling tasks.

The dataset and evaluation scripts are available on our GitHub,[18] and the post-competition phase on CodaLab will remain open for anyone interested in testing their approach on our data. We are planning to further expand the dataset with more languages and add more downstream tasks in the next release of the benchmark. We would appreciate any suggestions and collaboration from both computer scientists and historical linguists.

Acknowledgements

This shared task was in part supported by the Irish Research Council under grant number IRCLA/2017/129 (CARDAMOM – Comparative Deep Models of Language for Minority and Historical Languages[19]) and co-funded by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight) and SFI/12/RC/2289_P2 (Insight_2). We would also like to thank Universal Dependencies, University College Cork, the Royal Irish Academy, and HAS Research Institute for Linguistics for providing the source data.

References

Acadamh Ríoga na hÉireann. 2017. Corpas Stairiúil na Gaeilge 1600-1926. Retrieved: June 10, 2022.

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

[18] https://github.com/sigtyp/ST2024

[19] https://www.cardamom-project.org/

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology.

Bernhard Bauer, Rijcklof Hofman, and Pádraic Moran. 2017. St. Gall Priscian Glosses, version 2.0. Accessed: February 14, 2023.

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Giuseppe G. A. Celano, Daniel Zeman, and Federica Gamba. 2014. The Ancient Greek and Latin Dependency Treebank 2.0. Accessed: February 08, 2024.

Avihay Chriqui and Inbal Yahav. 2022. HeBERT & HebEMO: a Hebrew BERT model and a tool for polarity analysis and emotion recognition. *INFORMS Journal on Data Science*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Oksana Dereza. 2024. ACHILLES: Ancient and Historical Language Evaluation Set.

Oksana Dereza, Theodorus Fransen, and John P. Mccrae. 2023a. Do not trust the experts: How the lack of standard complicates NLP for historical Irish. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 82–87, Dubrovnik, Croatia. Association for Computational Linguistics.

Oksana Dereza, Theodorus Fransen, and John P. Mccrae. 2023b. Temporal domain adaptation for historical Irish. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 55–66, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aleksei Dorkin and Kairit Sirts. 2024. TartuNLP @ SIGTYP 2024 Shared Task: Adapting XLM-RoBERTa for ancient and historical languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024)*, St. Julians, Malta. Association for Computational Linguistics.

Adrian Doyle. 2018. Würzburg Irish Glosses. Accessed: February 14, 2023.

Loïc Grobol, Mathilde Regnault, Pedro Ortiz Suarez, Benoît Sagot, Laurent Romary, and Benoît Crabbé. 2022. BERTrade: Using Contextual Embeddings to Parse Old French. In *13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

HAS Research Institute for Linguistics. 2018. Old Hungarian Codices.

Dag T. T. Haug and Marius L. Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

Brandon Hawk, Antonia Karaisl, and Nick White. 2018. Modelling Medieval Hands: Practical OCR for Caroline Minuscule. *Faculty Publications*, (416).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Johannes Heinecke. 2024. UDParse @ SIGTYP 2024 Shared Task: Modern language models for historical languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024)*, St. Julians, Malta. Association for Computational Linguistics.

Hai Hu, Patrícia Amaral, and Sandra Kübler. 2021. Word embeddings and semantic shifts in historical Spanish: Methodological considerations. *Digital Scholarship in the Humanities*, 37(2):441–461.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pages 4411–4421.

Kyle P Johnson, Patrick J Burns, John Stewart, Todd Cook, Clément Besnier, and William JB 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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kevin Krahn, Derrick Tate, and Andrew C. Lamicela. 2023. Sentence embedding models for Ancient Greek using multilingual knowledge distillation. In *Proceedings of the Ancient Language Processing Workshop*, pages 13–22, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Yuxian Liang, Nan Duan, Yizhe Gong, Nan Wu, Fangxiang Guo, Weizhen Qi, Ming Gong, Lin Shou, Daxin Jiang, Gang Cao, Xinyu Fan, Ruofei Zhang, Rishabh Agrawal, Emo Cui, Siqi Wei, Tanmay Bharti, Yu Qiao, Ji-Hong Chen, Wei Wu, and et al. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint:1806.08730*.

Wouter Mercelis and Alek Keersmaekers. 2022. An ELECTRA model for Latin token tagging tasks. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 189–192, Marseille, France. European Language Resources Association.

Lester James V. Miranda. 2024. Allen Institute for AI @ SIGTYP 2024 Shared Task on word embedding evaluation for ancient and historical languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024)*, St. Julians, Malta. Association for Computational Linguistics.

Piotr Nawrot. 2023. nanoT5: Fast & simple pre-training and fine-tuning of T5 models with limited resources. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 95–101, Singapore. Association for Computational Linguistics.

Leonard Neidorf, Madison S. Krieger, Michelle Yakubek, Pramit Chaudhuri, and Joseph P. Dexter. 2019. Large-scale Quantitative Profiling of the Old English Verse Tradition. *Nature Human Behaviour*, 3(6):560–567.

Donnchadh Ó Corráin, Hiram Morgan, Beatrix Färber, Benjamin Hazard, Emer Purcell, Caoimhín Ó Dónaill, Hilary Lavelle, Julianne Nyhan, and Emma McCarthy. 1997. CELT: Corpus of Electronic Texts. Retrieved: March 15, 2021.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Frederick Riemenschneider and Kevin Krahn. 2024. Heidelberg-Boston @ SIGTYP 2024 Shared Task: Enhancing low-resource language analysis with character-aware hierarchical transformers. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024)*, St. Julians, Malta. Association for Computational Linguistics.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Maheswaran Kale, Mengting Ma, Massimo Nicosia, Shyam Rijhwani, Patrick Riley, Joudy-Maysaa Abdel Sarr, Xiyang Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Daniel L. Dickinson, Brian Roark, Bitan Samanta, Chen Tao, David I. Adelani, and et al. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. *arXiv preprint: 2305.11938*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jie Fu, Pengcheng Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of*

the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10215–10245.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint:1910.01108*.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Eszter Simon. 2014. Corpus Building from Old Hungarian Codices. In Katalin É. Kiss, editor, *The Evolution of Functional Left Peripheries in Hungarian Syntax*. Oxford University Press, Oxford.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Rachele Sprugnoli, Marco Passarotti, Cecchini Flavio Massimiliano, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), Language Resources and Evaluation Conference (LREC 2022)*, pages 183–188.

David Stifter, Bernhard Bauer, Fangzhe Qiu, Elliott Lash, Nora White, Siobhán Barret, Aaron Griffith, Romanas Bulatovas, Francesco Felici, Ellen Ganly, Truc Ha Nguyen, and Lars Nooij. 2021. Corpus PalaeoHibernicum (CorPH). Accessed: 19-02-2023.

Silvia Stopponi, Saskia Peels-Matthey, and Malvina Nissim. 2024. AGREE: a new benchmark for the evaluation of distributional semantic models of Ancient Greek. *Digital Scholarship in the Humanities*.

Li Sun, Florian Luisier, Kayhan Batmanghelich, Dinei Florencio, and Cha Zhang. 2023. From characters to words: Hierarchical pre-trained language model for open-vocabulary language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3605–3620, Toronto, Canada. Association for Computational Linguistics.

Zuoyu Tian, Dylan Jarrett, Juan Escalona Torres, and Patricia Amaral. 2021. BAHP: Benchmark of assessing word embeddings in historical Portuguese. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 113–119.

Sergio Torres Aguilar. 2022. Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France. European Language Resources Association.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

Krzysztof Wróbel and Krzysztof Nowak. 2022. Transformer-based part-of-speech tagging and lemmatization for Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 193–197, Marseille, France. European Language Resources Association.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu,

Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ivan Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in plutarch's shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, ..., and Rayan Ziane. 2023. Universal dependencies 2.12. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Shared Task Results

		AVG	chu	cop	fro	got	grc	hbo	isl	lat	latm	lzh	ohu	orv	san
POS-tagging															
Baseline		92.76	93.36	94.98	91.57	93.73	90.33	94.07	94.00	92.39	97.22	90.91	93.59	90.33	89.37
Constrained	HDB-BOS	**95.25**	**96.57**	**96.92**	**93.10**	**95.41**	**96.39**	**96.68**	**96.08**	**95.54**	**98.43**	**92.92**	**95.98**	**94.46**	**89.71**
	Team 21a	82.47	94.62	42.65	85.14	93.48	93.49	27.26	93.85	92.43	94.41	81.79	94.42	91.23	87.32
Unconstrained	UDParse	**96.09**	**97.00**	**97.33**	**96.01**	**96.47**	**96.49**	**97.84**	**96.88**	**96.83**	**98.79**	**93.76**	**96.71**	**94.99**	**90.02**
	TartuNLP	85.67	66.35	60.99	94.51	92.72	95.72	94.15	96.67	95.86	**98.79**	83.28	75.14	75.67	83.83
Lemmatisation															
Baseline		91.95	89.60	95.74	91.93	91.95	91.06	95.28	93.78	92.08	97.03	98.81	89.43	84.44	84.24
Constrained	HDB-BOS	**93.67**	**94.49**	**95.07**	**92.63**	**93.31**	**94.08**	**97.29**	**96.63**	**96.00**	**98.46**	99.18	**85.92**	**90.09**	**84.59**
	Team 21a	81.98	79.59	46.32	83.32	90.79	88.30	61.75	94.58	92.35	97.22	**99.84**	69.97	78.44	83.21
Unconstrained	UDParse	86.47	59.56	74.78	92.47	92.81	**94.02**	96.85	**97.96**	96.74	**98.91**	**99.96**	63.43	68.55	88.10
	TartuNLP	**94.88**	**92.70**	**98.28**	**95.11**	**95.41**	93.39	**98.15**	97.23	**96.99**	98.69	99.91	**86.91**	**89.23**	**91.48**
Morphological feature prediction															
Baseline		33.32	85.07	47.41	28.27	18.95	25.10	42.78	35.83	18.17	30.94	43.58	23.20	25.55	08.34
Constrained	HDB-BOS	**96.18**	**96.04**	**98.60**	**97.87**	**95.32**	**97.46**	**97.46**	**95.29**	**95.17**	**98.68**	**95.52**	**96.30**	**95.00**	**91.58**
	Team 21a	90.70	94.06	80.47	94.08	93.96	96.50	71.20	94.79	93.31	97.98	85.98	94.64	92.16	90.00
Unconstrained	UDParse	**96.68**	**96.49**	**98.88**	**98.33**	**96.23**	**97.78**	**97.05**	**95.92**	**96.66**	**98.83**	**96.24**	**96.62**	**95.16**	**92.60**
	TartuNLP	88.14	67.14	74.86	98.01	92.40	97.33	95.14	95.53	95.91	**98.83**	88.75	75.62	80.00	86.33

Table 5: Results of the *SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages* for problems 1-3. The winner of each track (constrained / unconstrained) is marked in **bold**, and the overall best result is underlined. The team names are as provided by participants, except HDB-BOS, which stands for 'Heidelberg-Boston'. For language code reference, see Table 1.

	AVG	chu	cop	fro	got	grc	hbo	isl	lat	latm	lzh	ohu	orv	san	sga	mga	ghc
Mask filling: word-level																	
UDParse	3.77	**2.80**	0.00	3.28	2.67	**3.07**	**5.39**	3.42	3.51	4.73	6.10	**6.31**	5.03	3.86	2.79	**4.03**	3.29
TartuNLP	**5.95**	2.42	**1.87**	**7.22**	**3.40**	3.01	0.00	**16.90**	**11.45**	**14.39**	**10.46**	0.06	**6.05**	**4.79**	**3.21**	3.99	**6.00**
Mask filling: character-level																	
UDParse	**55.62**	**66.77**	0.00	**62.77**	**74.59**	**68.46**	**36.85**	**66.45**	**67.91**	**72.93**	0.00	**66.52**	**66.77**	**70.10**	**58.38**	**53.38**	**58.09**
TartuNLP	48.38	53.79	**45.10**	52.46	67.34	61.15	18.56	57.32	65.79	69.84	**0.25**	45.65	48.04	64.52	34.86	39.49	49.88

Table 6: Results of the *SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages* for problem 4. The winner is marked in **bold**, and the absolute best result across all languages is underlined. The team names are as provided by participants. For language code reference, see Table 1.

		AVG	chu	cop	fro	got	grc	hbo	isl	lat	latm	lzh	ohu	orv	san	sga	mga	ghc
	Baseline	72.68	89.35	79.38	70.59	68.21	68.83	77.38	74.54	67.55	75.07	77.77	68.74	66.77	60.65	–	–	–
Constrained	HDB-BOS	**95.02**	**95.70**	**96.65**	**94.54**	**94.68**	**95.98**	**97.14**	**96.00**	**95.57**	**98.53**	**95.88**	**92.73**	**93.18**	**88.62**	–	–	–
	Team 21a	85.05	89.42	56.48	87.51	92.74	92.76	53.41	94.41	92.69	96.54	89.21	86.34	87.28	86.84	–	–	–
Unconstrained	UDParse	**61.93**	**71.15**	**58.90**	**71.10**	**73.07**	**71.84**	**67.05**	71.98	72.38	74.79	**59.20**	**70.61**	**69.15**	**69.61**	**30.59**	**28.71**	**30.69**
	TartuNLP	55.74	49.85	51.52	68.93	69.74	70.25	60.94	**72.88**	**73.15**	**76.15**	56.54	51.98	55.66	65.51	19.03	21.74	27.94

Table 7: Overall results of the *SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages* averaged across all problems for a given language. The winner for each setting is marked in **bold**. The team names are as provided by participants, except HDB-BOS, which stands for 'Heidelberg-Boston'. For language code reference, see Table 1.