UDParse @ SIGTYP 2024 Shared Task: Modern Language Models for Historical Languages

Johannes Heinecke
Orange Innovation
2 avenue Pierre Marzin
22300 Lannion, France
johannes.heinecke@orange.com

Abstract

SIGTYP's Shared Task on Word Embedding Evaluation for Ancient and Historical Languages was proposed in two variants, constrained or unconstrained. Whereas the constrained variant disallowed any other data to train embeddings or models than the data provided, the unconstrained variant did not have these limits. We participated in the five tasks of the unconstrained variant and came out first. The tasks were the prediction of part-of-speech, lemmas and morphological features and filling masked words and masked characters on 16 historical languages. We decided to use a dependency parser and train the data using an underlying pretrained transformer model to predict part-of-speech tags, lemmas, and morphological features. For predicting masked words, we used multilingual distilBERT (with rather bad results). In order to predict masked characters, our language model is extremely small: it is a model of 5-gram frequencies, obtained by reading the available training data.

1 Introduction

Since word embeddings and the transformer architecture (Vaswani et al., 2017) found their way into natural language processing (NLP), results for all NLP tasks improved to unseen levels. Multilingual pretrained language models like multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) include word embeddings for up to 100 languages. However, historical languages are most unlikely to be covered by these models. Since corpora of historical languages are limited in size and will most likely not grow anymore (unless an archaeological miracle unearths corpora yet unheard of) it will be difficult to include these languages to existing or new language models. In the SIGTYP Shared Task on Word Embedding Evaluation for Ancient and Historical Languages 2024 (ST 2024)[1], it is proposed to present word embeddings/models for 16 historical languages (cf. Table 1) for which part of speech (POS) (task 1), lemmas (task 2), morphological features (task 3) must be predicted. A fourth task asks to unmask masked words (task 4a) or characters (including spaces and punctuation, task 4b). Both masked words and masked characters can appear in an adjacent position. 10% of words and 5% of characters are masked. The shared task comes in two variants, constrained and unconstrained. In the first variant, only the data provided by the organizers can be used to train models, the unconstrained task allows any additional data to be used for training and inference.

The data used for the Shared Task (Dereza et al., 2024) has been compiled from various sources. Old, Middle, and Early Modern Irish is taken from Bauer et al. (2017), Doyle (2018), Ó Corráin et al. (1997), Acadamh Ríoga na hÉireann (2017); the Old Hungarian corpus origins from Simon (2014) and HAS Research Institute for Linguistics (2018), all other corpora have been published in version 2.12 of the Universal Dependencies project (UD) (Zeman et al., 2023)[2].

Both the training and the test data for the tasks 1, 2, and 3 is in CoNLL-U[3] format, i.e. the documents are segmented into tokenised sentences. The values for POS and the morphological features in tasks 1 and 3 are the UPOS and UFeats sets of the UD project. However not all languages use all possible features, e.g., the Old French data does not use the features `Number` or `Person`.

The Evaluation of the shared task is carried out by the CodaLab platform (Pavao et al., 2023) and uses the metrics shown in Table 2. In case of multiple metrics per task an unweighted average of the metrics was used.

We participated in all five tasks of the uncon-

[1] https://sigtyp.github.io/st2024.html

[2] https://universaldependencies.org, (Nivre et al., 2020)

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

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 142–150
March 22, 2024

Language	Code	Script	Dating	corpus size in tokens			corpus size in sentences		
				Train	Valid	Test	Train	Valid	Test
Ancient Greek	grc	Greek	VIII c. BCE – 110 CE	334,043	41,905	41,046	24,800	3,100	3,101
Ancient Hebrew†	hbo	Hebrew	X c. 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	I – II c. CE	57,493	7,282	7,558	1,730	216	217
Gothic	got	Latin	V – VIII c. 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 and Late Latin	lat	Latin	I c. BCE – IV c. CE	188,149	23,279	23,344	16,769	2,096	2,097
Medieval Latin	latm	Latin	774 – early XIV c. CE	599,255	75,079	74,351	30,176	3,772	3,773
Old Church Slavonic	chu	Cyrillic	X – XI c. 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: Data (†Afro-Asiatic language family, ‡Sino-Tibetan, *Finno-Ugric; all other languages are from the Indo-European language family)

Task		Metrics
1	POS-tagging	Accuracy @1, F1
2	Morph/ annotation	Macro-average of Acc. @1 per tag
3	Lemmatisation	Acc. @1, Acc. @3
4a	Filling masked words	Acc. @1, Acc. @3
4b	Filling masked chars.	Acc. @1, Acc. @3

Table 2: Evaluation metrics

strained variant of the shared task, even though our approach for filling mask characters does not use any other data than the data provided by the organizers. Apart from task 4 (filling masked words) we got the best results of all participants.

2 Related Work

Even though this shared task is not about dependency parsing, POS tagging and lemmatisation are often present in dependency parsing. The shared task in dependency parsing 2018 (Zeman et al., 2018) processed three historical languages, Ancient Greek, Latin, and Old Church Slavonic, for which annotated data was present in the Universal Dependencies project at the time. In many of the approaches word embeddings were used (calculated on corpora of these languages using word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), or fastText (Grave et al., 2018), the latter already provides word embeddings for Latin. The best results of the participants of the 2018 shared tasks for historical languages are above 98% for Latin, above 97% for Ancient Greek, and above 96% for Old church Slavonic for POS tagging. Lemmatisation for these languages also performs similarly well as modern languages. Sprugnoli et al. (2021) also studied the creation and evaluation word embeddings on Latin for the analysis of language change. More recently, several large language models for Classical Greek and Latin have been provided by (Riemenschneider and Frank, 2023) who evaluated this models on POS-tagging and lemmatisation (as this shared task), and dependency parsing. Brigada Villa and Giarda (2023) exploited models trained on Modern English to parse Old English, similar to our approach.

Evidently, word embeddings can be used for other tasks as well. E.g., Hamilton et al. (2016) use word embeddings of earlier version of English (but not going beyond the 1800s) to detect semantic shifts in English.

3 Approaches

3.1 Tasks 1 – 3: Inference of POS, Lemmas, and morphological features

Tasks 1, 2 and three consists of predicting the POS, the lemma, and morphological features of 13 historical languages (Table 1, exluding Old, Midlle and Early Modern Irish).

In order to infer POS, lemmas, and morphological features we used our syntactic dependency parser UDParse[4]. This parser is an evolution of UDpipe (Straka, 2018), which won the CoNLL

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

2018 Shared Task on dependency parsing (Zeman et al., 2018). UDpipe is a graph parser using pretrained word embeddings and embeddings of POS and characters to take the context into account. Word embeddings are loaded before the training, POS and characters embeddings are calculated from the training data. In contrast to UDPipe, UDParse uses word embeddings created by a pretrained transformer instead of contextless word embeddings produced by fastText[5]. This configuration proved to be very successful (Heinecke, 2020; Akermi et al., 2020), so we tried training models for the 13 languages for which the dependency syntax training data was available using different pretrained transformer models: bert-base-multilingual-uncased (Devlin et al., 2019), XLM-RoBERTa, GPT2 (Radford et al., 2019) and language specific models like slavicBERT (Arkhipov et al., 2019) for Old Church Slavonic and Old East Slavic or heBERT[6] for Ancient Hebrew. For the training we used 60 epochs with an initial learning rate of 10^{-3}, which decreased to 10^{-4} after 40 epochs[7], batch size was 32. We then chose for each language and each of task (1: POS, 2: lemmas, 3: morphological features) the best underlying pretrained transformer model (cf. Table 3). In nearly all cases XLM-RoBERTa produced the best results on the validation dataset (even though the difference, notably to multilingual BERT, was very small). For some languages we used different transformer models for tasks 1 to 3 to obtain the best results (on validation data).

Note that the test-data provided by the organizers was already tokenized. This simplifies enormously the tasks of assigning a POS, a lemma, or morphological features, especially for (historical) languages, which do not always come with standardized orthographies.

Even though the challenging fact was that most of these languages are not covered by any of the underlying pretrained language models, the results (Table 5, columns 2, 3, and 4) for POS, lemmas, and features, are well above 90% (except the lemmas for Old Hungarian and Old East Slavic). Partly this can be explained by the fact that the modern descendants of these languages are covered by XLM-RoBERTa etc., and at least some of the words of the historical languages still exist in the contemporary languages. Thus, the modern languages might have helped their ancestors. For comparison, UDParse on modern languages, covered by XLM-RoBERTa or mBERT has results[8] only slightly above the results obtained on historical language (Table 4).

Language code	POS	Lemma	Morphological features
chu	XLMR	XLMR	XLMR
cop	XLMR	GPT2	XLMR
fro	XLMR	mBERT	XLMR
got	XLMR	mBERT	mBERT
grc	XLMR	XLMR	XLMR
hbo	heBERT	XLMR	heBERT
isl	XLMR	XLMR	XLMR
lat	XLMR	XLMR	XLMR
latm	XLMR	XLMR	XLMR
lzh	mBERT	mBERT	mBERT
ohu	XLMR	XLMR	mBERT
orv	XLMR	XLMR	XLMR
san	mBERT	mBERT	XLMR

Table 3: Best underlying pretrained transformer models per language and task 1, 2, and 3. For language codes please refer to Table 1.

Code	UPOS	Lemma	Code	UPOS	Lemma
fr	97.93	98.41	fro	96.01	95.11
he	97.81	97.60	hbo	97.84	98.15
hu	97.07	95.51	ohu	96.71	86.91
ru	99.35	98.90	orv	94.99	89.23

Table 4: UDParse results for some modern languages (left) compared to historical languages (right, results copied from Table 5)

However, this does not explain the worse than average results for Old Hungarian and Old East Slavic whose descendants are also covered by XLM-RoBERTa. Old East Slavic contains some characters absent in its modern successors (Russian, Ukrainian and Belorussian). Similarly, the Old Hungarian corpus contains diacritics and characters not used in Modern Hungarian. This could have played a role. For the above average results for Coptic (not covered by XLM-RoBERTa and written in an alphabet totally absent in the vocabulary of XLM-RoBERTa), UDParse seems to exploit the word and character vectors produced during training to perform well in the lemmatisation.

[5] http://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[6] https://huggingface.co/avichr/heBERT

[7] Decreasing the learning rate after 40 epochs is a result of experimenting with UDParse at an earlier stage.

[8] For the results for other languages cf. https://github.com/Orange-OpenSource/UDParse/blob/master/doc/results.md

Code	UPOS	Lemma	Morph. feat.	Word fill	Char fill	Avg.
chu	97.00	92.70	96.49	2.80	66.77	71.15
cop	97.33	98.28	98.88	0.00	0.00	58.90
fro	96.01	95.11	98.33	3.28	62.77	71.10
got	96.47	95.41	96.23	2.67	74.59	73.07
grc	96.49	93.39	97.78	3.07	68.46	71.84
hbo	97.84	98.15	97.05	5.39	36.85	67.05
isl	96.88	97.23	95.92	3.42	66.45	71.98
lat	96.83	96.99	96.66	3.51	67.91	72.38
latm	98.79	98.69	98.83	4.73	72.93	74.79
lzh	93.76	99.91	96.24	6.10	0.00	59.20
ohu	96.71	86.91	96.62	6.31	66.52	70.61
orv	94.99	89.23	95.16	5.03	61.34	69.15
san	90.02	91.48	92.60	3.86	70.10	69.61
ghc	—	—	—	3.29	58.09	30.69
mga	—	—	—	4.03	53.38	28.71
sga	—	—	—	2.79	58.38	30.59

Table 5: Results (Word filling failed for Coptic (cop) and character filling missing for Coptic and Classical Chinese (lzh). For Old Irish (sga) Middle Irish (mga), and Early Modern Irish (hgc), only data for the word and character filling tasks was available)

All results for tasks 1, 2, and 3 are well above the baseline provide by the shared task's organisers (Table 6) with the exception of the lemmatisation of Old Hungarian (ohu).

Code	UPOS	Lemma	Morph. feat
chu	3.64	3.09	11.42
cop	2.35	2.54	51.47
fro	4.44	3.18	70.06
got	2.74	3.46	77.28
grc	6.16	2.33	72.68
hbo	3.77	2.86	54.26
isl	2.88	3.45	60.09
lat	4.44	4.90	78.49
latm	1.56	1.65	67.89
lzh	2.85	1.10	52.66
ohu	3.12	–2.53	73.42
orv	4.66	4.79	69.60
san	0.65	7.24	84.26

Table 6: Difference with respect to the baseline

3.2 Task 4a: Filling masked words

The task of filling one or several single words in a sentence was the most challenging task for historical languages. Consequently, our results are extremely low (Table 5, column 5). This is probably more due to the chosen approaches than to the fact that the pretrained transformers have been trained little or not at all on these languages. We tried two classical approaches, an encoder (distilbert-base-multilingual-cased, Sanh et al. (2019)) and an encoder/decoder (facebook/mbart-large-50, Tang et al. (2020)). In the first case we used Huggingface's `AutoModelForMaskedLM`, the `AdamW` (Loshchilov and Hutter, 2019) optimiser with a learning rate of $5 * 10^{-5}$, a batch size of 8 and early stopping, which stopped the training after 4 to 6 epochs depending on the language. For the training process, we did not use the masks provided in the training corpus, but masked words randomly with a probability of 15%. In the second case (with mBART) we used Huggingface's `MBartForConditionalGeneration` (other hyperparameters were identical).

The difference between distilBERT and mBART was marginal, possibly linked to a problem not identified before the shared task's deadline. We submitted the results of the first approach. However since this approach only predicts a single token (in the sense of distilBERT's vocabulary) for each masked word instead of a word (in most cases two or more tokens) our prediction was wrong for all masked words which are represented by more than one distilBERT token. In other words, masked words which are not in distilBERT's vocabulary, could not be predicted with this approach. The second approach, based on `MBartForConditionalGeneration`, did indeed return most times a word (or more) for a masked word, but we had cases where only a space was obtained.

3.3 Task 4b: Filling masked characters

For this subtask we chose a very old idea: a simple n-gram count and applying the most frequent n-gram which matches the masked character and its context. We trained our model by counting all n-grams in the unmasked part of the training corpus. We then looked for every masked character in test sentences and tried to find the frequency of all n-grams which include the masked character (we experimented with 3-grams and 5-grams, the latter proved to work much better): for instance, forthe following string "Ne␣voloi?␣aler␣nule part."[9] which includes a masked character, we take the frequencies of all the 5 character windows around the masked characters, including spaces ("␣") from the training

[9]Taken from the Old French training corpus. The shared task data used "[_]" as placeholder for masked characters. We replaced it with a single character not occurring anywhere in the data. For a better readability we use "?" here.

corpus. "?" is the masked character. The letter in inverted colors is the candidate letter:

1. "oloi?" →
   - "oloie" which has frequency of 6 in the training corpus,
   - "oloil" (frequency of 3),
   - "oloir" (8),
   - "oloit" (15, at this stage, this 5-gram is the best match. It is therefore kept while the other 5-grams are discarded)
2. "loi?␣" →
   - "loie␣" (2),
   - "loil␣" (2),
   - "loir␣" (6),
   - "lois␣" (1),
   - "loit␣" (22, new best match so far)
3. "oi?␣a" →
   - "oia␣a" (1),
   - "oie␣a" (17),
   - "oif␣a" (1),
   - "oil␣a" (3),
   - "oir␣a" (3),
   - "ois␣a" (14),
   - "oit␣a" (57, retained),
   - "oiz␣a" (8)
4. "i?␣al" →
   - "it␣al" (3); discarded since with "oit␣a" above we have found a more frequent match already
5. "?␣ale" →
   - "␣␣ale" (3),
   - "a␣ale" (1),
   - "e␣ale" (3),
   - "i␣ale" (2),
   - "l␣ale" (1),
   - "n␣ale" (2),
   - "r␣ale" (6),
   - "s␣ale" (2),
   - "t␣ale" (17),
   - "z␣ale" (1),
   - "«␣ale" (1); all discarded

In this example "oit␣a" is the most frequent replacement for one the 5-grams which contain the masked character ("oi?␣a"). So, we can replace the masked character by "t" to obtain "Ne␣voloit␣aler␣nule part.". Note that at least in this example for each of the five 5-ngrams the best match is the one where the masked character is the same ("t"), but this was not always the case.

The results of this approach can be found in Table 5, column 6. For time reason we could not implement the needed post processing for Classical Chinese (lzh) to rebuild the Hanzi characters from the decomposed characters in the train/validation and test data. Apparently we did not submit the Coptic data to the evaluation server, but a run after the deadline resulted in an accuracy of 62.26%. Interestingly, the score for Ancient Hebrew (hbo) is only half as good as for the other languages. Since the number of different characters of Ancient Hebrew is rather low (cf. Table 7), the reason of this bad result must be found elsewhere. Surprisingly the evaluation of the validation corpus, resulted in around 60% accuracy.

lang. code	characters	lang. code	characters
san	37	ghc	92
cop	41	lat	116
got	50	isl	118
fro	64	ohu	120
hbo	67	chu	124
latm	77	orv	156
mga	77	grc	176
sga	78	lzh	318

Table 7: Number of different characters in the fill masked characters test data. Many languages contain accentuated characters, digits, Classical Latin (lat) contains citations in Greek which account for the unexpected high number of different characters.

We think a more word-context-aware approach could have improved the results, even a simple word based bi- or trigram. For instance in the Old French validation corpus is the following masked character "se je ai dite [_]ne response". Our approach finds for the 5-gram "␣?ne␣" the 5-gram "␣␣ne␣" (the most frequent) instead of the correct "␣une␣". Due to the approaching deadline, we did not have the time to implement and test this.

4 Conclusion

We successfully used rather old and well-established techniques to provide a solution to the five tasks of this year's SIGTYP shared task. Putting aside the failed results for filling-masked-words task, we got very good results for POS tagging, lemmatization, and morphological feature assignment, which are as good as for modern languages and well above the baseline. We are not aware of any state-of-the-art values for filling masked characters, however, even though our results are first placed in the shared task, they are probably perfectible. For modern languages, word embedding or transformer-based methods, e.g. such as CharacterBERT, (El Boukkouri et al., 2020) will probably yield much better results.

References

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

Imen Akermi, Johannes Heinecke, and Frédéric Herledan. 2020. Transformer based natural language generation for question-answering. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 349–359, Dublin, Ireland. Association for Computational Linguistics.

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.

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

Luca Brigada Villa and Martina Giarda. 2023. Using modern languages to parse ancient ones: a test on Old English. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 30–41, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Oksana Dereza, Adrian Doyle, Priya Rani, Atul Kr. Ojha, Pádraic Moran, and John P. McCrae. 2024. Findings of the 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.

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.

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

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

HAS Research Institute for Linguistics. 2018. Old Hungarian Codices. Accessed: October 22, 2023.

Johannes Heinecke. 2020. Hybrid enhanced Universal Dependencies parsing. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 174–180, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Tyers Francis, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *LREC 2020*, Marseille. ELRA.

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.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. https://openai.com/blog/better-language-models/.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter.

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

Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2021. Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas. *Italian Journal of Computational Linguistics*, 6(1).

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. 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, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Þórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Francisco Costa, Marine Courtin, Mihaela Cristescu, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Farah

Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Tommi A. Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Ure-

šová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, 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.