UppsalaNLP at EvaLatin 2026: Multilingual parsing for Latin

Sara Stymne
Department of Linguistics and Philology
Uppsala University, Sweden
sara.stymne@lingfil.uu.se

Abstract

We describe the UppsalaNLP submission to the EvaLatin dependency parsing shared task. We explore using an out-of-the-box parser in combination with multi-treebank training on Latin and multilingual training on other ancient languages. Adding additional languages yields only small gains, but the results vary across treebanks and genres, with the largest positive effect for poetry. Our systems perform best in the shared task for prose but are less competitive for poetry, indicating the need for genre adaptation.

Keywords: Dependency parsing, Latin, Multilingual parsing

1. Introduction

We present the UppsalaNLP entry in the EvaLatin 2026 shared task on dependency parsing for Latin (Iurescia et al., 2026). Our main focus is to explore how an out-of-the-box system, MaChamp (van der Goot et al., 2021) in our case, can be improved through multilingual and multi-treebank training. Our main goal was to investigate how far it is possible to get with an existing strong parser, without any technical additions, only by combining existing treebanks during training. This is a feasible setup for many digital humanities projects.

Our base system is a multi-treebank system trained on eight Latin treebanks, compatible with the Universal Dependencies (UD, Nivre et al., 2020) annotation scheme, see Table 1 for an overview. We also follow Straka et al. (2024), who noticed that the Latin PROIEL treebank diverged from the other Latin treebanks, and used a harmonized version of it (Gamba and Zeman, 2023a,b).

For multilingual training, we also include other ancient languages, which have been shown to be useful for Latin parsing (Smith et al., 2018), including Ancient Greek, which has been shown to be useful in a Latin low-resource scenario (Karamolegkou and Stymne, 2021). In addition, we also include historical Romance languages, which are now available in UD. Many of the treebanks from the other languages stem from the PROIEL project, for which the Latin treebank has inconsistent annotations (Gamba and Zeman, 2023a,b). Thus, we also experiment with the impact of excluding them from the training. In addition, we explore the impact of multitask training on a mix of UD tasks.

We submit two systems to the EvaLatin shared task. Our primary system is trained on Latin in combination with other historical languages, but excluding PROIEL treebanks, except for Latin, where a harmonized version is used. Our secondary submission is further trained only on Latin. Our primary system ranks first on most metrics for the prose data in the EvaLatin evaluation, but it ranks fourth for the poetry data. Our secondary system is placed below the primary system for prose, second to third, but beats it for poetry, being placed third. This calls for further investigation into genre adaptation for parsing to better account for the peculiarities of poetry, an issue that has been previously investigated for other genres, domains, and languages (Stymne, 2020; Müller-Eberstein et al., 2021; Danilova and Stymne, 2023). Our multilingual approach yields only small gains compared to training only on Latin when evaluated on the test portions of the training treebank, but has a larger effect on the shared task test data, especially for poetry.

2. Related Work

EvaLatin 2026 (Iurescia et al., 2026) is a follow-up from EvaLatin 2024 (Sprugnoli et al., 2024). For both editions, the focus is on dependency parsing for Latin prose and poetry. In 2024, the winning team presented LatinPipe (Straka et al., 2024), an extension of UDPipe (Straka and Straková, 2017). LatinPipe contains both Latin-specific extensions and general architecture improvements. Specific to Latin, they train on multiple Latin treebanks, including a harmonized version of PROIEL (Gamba and Zeman, 2023a,b), use the gold UPOS (Universal part-of-speech tags de Marneffe et al., 2021) provided by the EvaLatin organizers, and combine multilingual and Latin-specific language models (LMs). They also add two BiLSTM layers on top of the LMs and perform ensembling. All of these techniques are shown to improve results on both UD and EvaLatin test sets. Other teams at EvaLatin 2024 explored using historical Latin sentence embeddings (Behr, 2024) and span-span prediction on top of a Latin LM (Mercelis, 2024), but with considerably worse performance than LatinPipe.

One approach for Latin dependency parsing is treebank combination, either for Latin alone or in

Proceedings of LT4HALA 2026 @ LREC 2026, pages 429–436.

a cross-lingual setting. An obstacle to the success of combining treebanks is the inconsistencies of the annotation, even within a standardized framework like UD. For Latin, Gamba and Zeman (2023b) show that there are considerable discrepancies across the UD Latin treebanks, especially for PROIEL. They design an automatic harmonization scheme that leads to improved results across treebanks. Kupari et al. (2024) also use these treebanks, further improving by using stronger parsers. Smith et al. (2018) explore treebank combination and find that small groups of treebanks for related languages tend to improve parsing results for most languages, especially for those with small treebanks. For Latin, they use a set of ancient-language treebanks, which improves performance on most Latin UD test sets, possibly due to shared annotation projects between these languages. In their study, they used UUParser (de Lhoneux et al., 2017), a BiLSTM-based parser. This study is revisited by van der Goot and de Lhoneux (2021), using the MaChamp parser (van der Goot et al., 2021), based on the transformer model mBERT (Devlin et al., 2019). Both overall and for Latin, they find that cross-treebank training is useful, but that it is often as effective to use all available treebanks as the designed language sets by Smith et al. (2018). In addition, they see a positive effect of using dataset embeddings. Karamolegkou and Stymne (2021) explore cross-lingual parsing in a low-resource scenario for Latin, and find that Ancient Greek is a better transfer language than modern romance languages like Italian. Kupari et al. (2024) explored the impact of using data from specific time periods on dependency parsing of Latin; however, the impact was minor.

Several works have investigated which LMs are most useful for parsing Latin. Nehrdich and Hellwig (2022) show that the Latin BERT model (Bamman and Burns, 2020) performs better than both static embeddings and mBERT. Straka et al. (2024) compare the use of two Latin-specific BERT models (Riemenschneider and Frank, 2023) to the multilingual XLM-RoBERTa (XLM-R) model (Conneau et al., 2020), and find that XLM-R is better than using the Latin models, but that it is possible to get additional gains by combining the Latin models with XLM-R. Behr (2024) investigated the usefulness of historical Latin sentence embeddings (Reimers and Gurevych, 2019) fine-tuned on top of Latin BERT (Bamman and Burns, 2020), which, however, did not improve the parsing scores.

Another line of work, for dependency parsing in general, focuses on improving parsing for specific genres. Stymne (2020) matches genres across languages, Müller-Eberstein et al. (2021) train parsers based on cross-lingual genre clusters, and Danilova and Stymne (2023) mine genre annotations from UD treebanks, and compare human genre annotations with genre clusters. None of these works specifically targets the goal of parsing poetry and prose in historical languages. This has been addressed for POS-tagging of historical Germanic languages, where it was shown that training on prose for tagging poetry, or vice versa, was not successful, and combining the genres had only a minor effect (Miani et al., 2026).

3. Data

Table 1 summarizes the Latin treebanks used. The majority of treebanks are from UD, release 2.17 (Zeman et al., 2025). Due to annotation discrepancies across UD treebanks, we follow LatinPipe (Straka et al., 2024) and explore the use of a set of harmonized treebanks (referred to as "H", Gamba and Zeman (2023a,b)) for all UD treebanks except CIRCSE, for which no harmonization is available. We try to use the harmonized treebanks for all treebanks for which they are available, or only for PROIEL. In addition, we use the same two small treebanks (and abbreviations) as LatinPipe, both annotated according to the UD scheme. For testing, we use the harmonized version of PROIEL, which we will refer to as H-PROIEL, and the UD2.17 version of all other UD treebanks. Arch has a small test set of 30 sentences, which we consider insufficient for reliable evaluation. The EvaLatin shared task provides two official testsets, covering prose, 203 sentences, and poetry, 308 sentences.

We also include UD treebanks from other ancient languages, grouped into three groups: Ancient Greek, Romance, covering Old French and Old Occitan, and Other, covering Gothic and Old Church Slavonic. Table 2 contains an overview, including the language groupings. In addition, we train a model using all these treebanks. In this set, several treebanks come from the PROIEL project (Haug and Jøhndal, 2008), which was shown to diverge from the UD framework in some respects for Latin. We thus also experiment with excluding those treebanks and train on only treebanks from other projects.

4. Dependency Parser

We use the MaChamp toolkit (van der Goot et al., 2021) for dependency parsing. MaChamp is a multipurpose toolkit targeted at a diverse set of NLP tasks, including dependency parsing, and is especially useful for multitask and multi-dataset training. It fine-tunes a language model for specific tasks. We use XLM-RoBERTa Large (Conneau et al., 2020), which performed better than any Latin-specific model in isolation and only slightly worse than in combination with two Latin-based models

Short	Treebank	Size in sentences Training	Test
Circse	UD_Circse (https://github.com/UniversalDependencies/UD_Latin-CIRCSE)	762	832
ITTB	UD_ITTB (Passarotti, 2019)	22,775	2,101
LLCT	UD_LLCT (Cecchini et al., 2020a)	7,289	894
Perseus	UD_Perseus (Bamman and Crane, 2011)	1,334	939
PROIEL	UD_PROIEL (Haug and Jøhndal, 2008)	3,387	1260
UDante	UD_UDante (Cecchini et al., 2020b)	926	421
Arch	Archimedes Latinus (Fantoli and de Lhoneux, 2022)	47	(30)
Sab	De Latinae Linguae Reparatione (Sabellicus) (Gamba and Cecchini, 2024)	246	–

Table 1: Latin treebanks used with sizes of training and test sets. Treebanks in the top of the table is part of UD, treebanks in the bottom are compatible with UD.

Language	Treebank	Training size	A. Greek	Romance	Other	No PROIEL	All
	UD_Perseus (Bamman and Crane, 2011)	11,476	X			X	X
Ancient Greek	UD_PROIEL (Haug and Jøhndal, 2008)	15,016	X				X
	UD_PTNK (Swanson et al., 2024)	727	X			X	X
Old French	UD_PROFTIEROLE (Stein and Prévost, 2013)	15,962		X		X	X
Old Occitan	UD_CorAG (Romanova et al., 2025)	912		X		X	X
Gothic	UD_PROIEL (Haug and Jøhndal, 2008)	3,387			X		X
Old Church Slavonic	UD_PROIEL (Haug and Jøhndal, 2008)	18,327			X		X

Table 2: Treebanks used for additional languages. The final columns show which groups of training languages each treebank is included in.

for LatinPipe (Straka and Straková, 2017). The MaChamp dependency parser is a non-projective graph-based parser based on the CLE algorithm (Chu and Liu, 1965; Edmonds, 1967), using deep biaffine attention (Dozat and Manning, 2018). In our multitasking experiments, we also predict UPOS, morphological tags, and lemmas, the latter by predicting conversions from word forms to lemmas. These tasks use a standard greedy sequence labelling approach. We use default parameters in MaChamp, except for the sample smoothing parameter a, used for balancing dataset size during training, which we set to 0.5 to prevent larger datasets from overwhelming smaller ones.

5. Experiments

Our main focus is on combining treebanks for Latin and other ancient languages. We also explore the impact of multitask training, comparing training MaChamp only for dependency parsing, training it also for UPOS, and training it also for UPOS, morphology prediction, and lemmatization.

For our treebank combination experiments, we first train models for each treebank, as a point of comparison. We then combine all available Latin treebanks in three settings: use only UD2.17 versions, use the harmonized versions of UD treebanks (where available), or use UD2.17 treebanks for all treebanks except PROIEL, for which the harmonized version was used.

We then combine the training on Latin with groups of other languages, as shown in Table 2, and with all treebanks, or all treebanks excluding PROIEL. For our final set of experiments, we take the multilingual model, trained on all data except non-Latin PROIEL treebanks, and further train it only on Latin, either on all Latin treebanks (with harmonized PROIEL) or only on CIRCSE, which we believe could be a better fit for poetry data.

Note that we do not use any gold UPOS-tags or morphology as input to our parser. While this information was available for the EvaLatin shared task and could have boosted the result, we prefer the more realistic setting of not using gold tags, which are not generally available for Latin texts.

For results on the testsets of the Latin treebanks used for training, we report labeled attachment score (LAS) without any subtypes. For the EvaLatin testdata, we report all official metrics of the EvaLatin shared task (Iurescia et al., 2026), which include LAS and CLAS, content-word LAS, which excludes relations to function words (Nivre and Fang, 2017), for both metrics with and without subtypes. All reported results are averages across three runs with different seeds, except for the official shared task results, which were calculated for the submitted systems.

6. Results

Table 3 shows LAS scores on the testsets from the training treebanks. Table 3.A shows results for multitasking experiments trained on all Latin treebanks using H-PROIEL. The effect of including other tasks than dependency parsing is minor and varies across treebanks. In the following, we will use systems trained on dependency parsing and UPOS that had the best average scores, by a small margin.

Experiment	H-PROIEL	Circse	ITTB	LLCT	Perseus	UDante	Average
A) Multitask training variants							
1 Dep only	86.83	68.43	**91.37**	**95.88**	78.52	**77.10**	83.02
2 Dep+UPOS+morph+lemma	86.54	**70.12**	91.06	95.62	**79.12**	76.70	83.19
3 Dep+UPOS	**86.93**	69.31	91.23	95.86	79.05	77.00	**83.23**
B) Latin treebank variants and combinations(Harmonized/UD)							
4 Single treebank	**88.84**	60.60	**92.54**	95.83	75.23	75.09	81.36
5 All H+ Arch+Sab	86.94	**70.30**	88.82	88.36	**79.09**	75.98	81.58
6 All UD + Arch+sab	81.98	69.49	91.19	95.83	78.93	**77.34**	82.46
3 H PROIEL+UD+Arch+Sab	86.93	69.31	91.23	**95.86**	79.05	77.00	**83.23**
C) Multilingual training, Latin (as in previous row) +							
7 Ancient Greek	87.31	70.54	91.31	95.95	79.01	77.22	83.56
8 Other	**87.36**	70.30	**91.43**	95.90	79.01	**77.40**	83.57
9 Romance	87.06	70.62	91.27	95.91	79.01	76.98	83.48
10 All	**87.36**	70.94	91.40	95.97	**79.33**	77.09	**83.68**
11 No PROIEL	86.66	**71.10**	91.32	**95.99**	**79.33**	77.22	83.60
D) Continued Latin only training (w.r.t in previous row)							
12 Circse only	84.00	70.43	75.20	77.65	63.66	62.79	72.29
13 H PROIEL+UD+Arch+Sab	**86.86**	**71.05**	**91.36**	**95.94**	**79.18**	**76.86**	**83.54**
E) Previous work							
MaChAmp (UD2.8)	–	–	92.45	95.41	74.67	74.01	–
LatinPipe (ensemble – UD2.13+H-PROIEL+Sab+Arch)	–	–	92.45	95.78	84.22	81.47	–

Table 3: LAS scores without subtypes on the testsets corresponding to the Latin training data. For PROIEL, we report scores in the harmonized version. Previous work are from Straka et al. (2024) (LatinPipe) and van der Goot et al. (2021) (MaChamp)

Table 3.B shows results with different combinations of Latin treebanks for training. For two of the larger treebanks, PROIEL and ITTB, we achieved the best results by training only on data from that treebank. For all other treebanks, training on the combination of all Latin treebanks was better, with large gains especially for the small CIRCSE treebank. Using original UD treebanks or harmonized treebanks had varying effects across treebanks, but only using the harmonized version of PROIEL led to the best average results and is used in all further experiments.

Table 3.C shows the effect of combining Latin data with other languages. Overall, this had little effect compared to training only on Latin, but led to small improvements for all treebanks, compared to the system with its corresponding Latin treebanks (system 3 in Table 3). The best language group varies across treebanks, but on average, models that used all languages performed slightly better than those that used a subset. While the model without non-Latin PROIEL treebanks performed slightly worse on average than the model trained on all treebanks, we still chose to use it in the following, since the model trained on all treebanks mainly improved H-PROIEL, indicating that the non-Latin treebanks may have the same discrepancies with UD annotation as Latin PROIEL.

The results in Table 3.D show the effect of further training the multilingual model only on Latin, either on all treebanks, or only on the small CIRCSE. Training on the small CIRCSE deterioriated the results for all treebanks, possibly due to overfitting, and even led to worse results on the CIRCSE test set. Using all Latin treebanks did not improve overall over the multilingual model, but kept the results relatively stable.

In Table 3.E, we compare our results with previous work, in the form of MaChamp and LatinPipe. Note, though, that those scores are for other UD versions than ours. For 3 out of 4 treebanks, our multilingual model is better than the original MaChamp model (van der Goot et al., 2021). For ITTB, our multilingual model falls behind, but the model trained only on ITTB is slightly ahead of MaChamp. We still fall behind the winner of the previous EvaLatin shared task, LatinPipe (Straka et al., 2024), except for the LLCT treebank. We believe that adding some of the LatinPipe features not used in this work, such as gold UPOS, Latin LMs, BiLSTM layers, and ensembling, could potentially bridge the gap between these systems.

Based on the results in Table 3, we decided to submit system 11, a multilingual system excluding non-Latin PROIEL and using harmonized Latin PROIEL as our primary system, UppsalaNLP 1, and system 13, further trained on Latin, as our secondary system, UppsalaNLP 2. Table 4 shows the official results for these systems, and also the results for a subset of contrastive systems, including the two single-source treebanks chosen for training the EvaLatin baselines, ITTB and CIRCSE (Iurescia et al., 2026). We also include ranks in the shared task for our submitted systems.

For prose, our primary system performed best on 3 out of 4 official metrics, whereas our secondary system performed better on poetry, where it ranked 3rd on all metrics. For prose, UppsalaNLP 1 performed on par with training only on ITTB, but better than all other systems. For poetry, UppsalaNLP 2 performed stronger than all our other systems. The difference between the two submitted systems was quite small on the test sets from the training treebanks, in Table 3, but considerably larger on the shared-task test sets. The multilingual training, in the submitted systems and in system 10 improves

		Prose				Poetry			
		LAS	LAS+s	CLAS	CLAS+s	LAS	LAS+s	CLAS	CLAS+s
11	UppsalaNLP 1 (Latin+no PROIEL)	**87.55**	**84.41**	**86.88**	**83.14**	69.59	66.63	68.20	67.34
	RANK	2	1	1	1	4	4	4	4
13	UppsalaNLP 2 (11 + further Latin training)	87.26	83.74	86.45	82.17	**70.55**	**67.56**	**68.47**	**68.47**
	RANK	3	2	3	2	3	3	3	3
3	Latin only	87.29	83.81	86.60	82.31	68.48	65.45	66.92	65.97
10	Latin + all	**87.37**	**84.03**	**86.65**	**82.79**	**69.88**	66.53	**68.41**	**67.23**
12	Fine-tuning on CIRCSE	73.36	70.36	78.01	76.34	69.23	**66.64**	67.74	67.18
4	CIRCSE	57.23	53.31	59.48	56.55	**59.69**	**57.35**	**58.69**	**59.14**
4	ITTB	**87.89**	**84.43**	**86.92**	**82.77**	45.43	42.73	43.33	42.02

Table 4: Official results for the submitted runs on all shared task metrics (top) and some contrastive results (bottom). Metrics marked with '+s' include the evaluation of subtypes. Numbers refer to systems presented in Table 3.

over Latin-only training, system 3, on all metrics for both genres, but especially for poetry, even though adding non-Latin languages only had a small effect on the devsets in Table 3.C.

For poetry, none of the single-treebank systems performed well, and ITTB, which was very competitive for prose, even fell behind the small CIRCSE treebank. Further training the multilingual system on only CIRCSE was competitive for poetry, but not for prose. For poetry, our multilingual system performed considerably better than the Latin-only systems. The multilingual system, including the non-Latin PROIEL treebanks, performed similarly to not using PROIEL.

In most cases, the four official metrics show the same tendencies. In the case of training on CIRCSE, system 12, though, we can see that this system performs comparatively better on the metrics with subtypes, indicating that CIRCSE is in better agreement with the test data for subtypes than the other treebanks.

We note that the accuracy for UPOS-tagging, which was performed during our multitask training, was just over 94% on poetry but over 97.5% on prose for our two submitted systems, indicating that not only is parsing poetry challenging, but also other tasks. The winners from 2024, (Straka et al., 2024) report a gain of over 1 LAS point when adding gold POS to their parser, indicating that improved tagging could also help parsing.

7. Conclusion

We have presented the UppsalaNLP contribution to the EvaLatin shared task, which placed first for parsing prose and 3–4 on parsing poetry. Our system uses multilingual training for Latin in combination with other ancient languages. The multilingual training mostly led only to minor improvements, except for poetry, for which it led to a substantial improvement. We believe that some of the most pressing future work is to improve parsing for poetry, possibly by including poetry training data for other languages.

8. Acknowledgement

Computations were enabled by resources provided by the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

9. Bibliographical References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rufus Behr. 2024. Behr at EvaLatin 2024: Latin dependency parsing using historical sentence embeddings. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 198–202, Torino, Italia. ELRA and ICCL.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020a. A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942, Marseille, France. European Language Resources Association.

Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020b. UDante: First steps towards the universal dependencies treebank of Dante's Latin works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, pages 1–7. Italian Association for Computational Linguistics (AILC).

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

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.

Vera Danilova and Sara Stymne. 2023. UD-MULTIGENRE – a UD-based dataset enriched with instance-level genre annotations. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 253–267, Singapore. Association for Computational Linguistics.

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2017. Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 99–104, Pisa, Italy. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the national Bureau of Standards B*, 71(4):233–240.

Margherita Fantoli and Miryam de Lhoneux. 2022. Linguistic annotation of neo-Latin mathematical texts: A pilot-study to improve the automatic parsing of the archimedes latinus. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 129–134, Marseille, France. European Language Resources Association.

Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Dag Trygve Truslew Haug and Marius 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.

Federica Iurescia, Marco Passarotti, and Rachele Sprugnoli. 2026. Overview of the Dependency Parsing Task at EvaLatin 2026. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Antonia Karamolegkou and Sara Stymne. 2021. Investigation of transfer languages for parsing Latin: Italic branch vs. Hellenic branch. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 315–320, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala, and Jenna Kanerva. 2024. Improving Latin dependency parsing by combining treebanks and predictions. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 216–228, Miami, USA. Association for Computational Linguistics.

Wouter Mercelis. 2024. KU Leuven / Brepols-CTLO at EvaLatin 2024: Span extraction approaches for Latin dependency parsing. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 203–206, Torino, Italia. ELRA and ICCL.

Irene Miani, Sara Stymne, and Gregory R. Darwin. 2026. Cross-lingual and cross-domain transfer learning for POS tagging in historical Germanic low-resource languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 542–558, Rabat, Morocco. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Nehrdich and Oliver Hellwig. 2022. Accurate dependency parsing and tagging of Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 20–25, Marseille, France. European Language Resources Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

Marco Passarotti. 2019. The project of the Index Thomisticus treebank. *Digital Classical Philology*, 10:299–320.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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.

Natasha Romanova, Rayan Ziane, and Barbara Francioni. 2025. Adaptation of models for parsing of Old Gascon. In *Proceedings of Journées scientifiques du réseau thématique LIFT2 linguistique informatique, formelle et de terrain*, Paris, France.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia. ELRA and ICCL.

Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts: The Syntactic Reference Corpus of Medieval French (SRCMF). In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpora*, Corpus Linguistics and International Perspectives on Language, pages 275–282. Gunter Narr Verlag.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.

Sara Stymne. 2020. Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany. Association for Computational Linguistics.

Daniel G. Swanson, Bryce D. Bussert, and Francis Tyers. 2024. Producing a parallel Universal Dependencies treebank of Ancient Hebrew and Ancient Greek via cross-lingual projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13074–13078, Torino, Italia. ELRA and ICCL.

Rob van der Goot and Miryam de Lhoneux. 2021. Parsing with pretrained language models, multi-

ple datasets, and dataset embeddings. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 96–104, Sofia, Bulgaria. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

10. Language Resource References

Federica Gamba and Flavio Massimiliano Cecchini. 2024. De latinae linguae reparatione treebank. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, et al. 2025. Universal dependencies 2.17. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).