Where are we Still Split on Tokenization?

Rob van der Goot
IT University of Copenhagen
robv@itu.dk

Abstract

Many Natural Language Processing (NLP) tasks are labeled on the token level, for these tasks, the first step is to identify the tokens (tokenization). Because this step is often considered to be a solved problem, gold tokenization is commonly assumed. In this paper, we investigate if this task is solved with supervised tokenizers. To this end, we propose an effient multi-task model for tokenization that performs on-par with the state-of-the-art. We use this model to reflect on the status of performance on the tokenization task by evaluating on 122 languages in 20 different scripts. We show that tokenization performance is mainly dependent on the amount and consistency of annotated data as well as difficulty of the task in the writing systems. We conclude that besides inconsistencies in the data and exceptional cases the task can be considered solved for Latin languages for in-dataset settings (>99.5 F1). However, performance is 0.75 F1 point lower on average for datasets in other scripts and performance deteriorates in cross-dataset setups.[1]

1 Introduction

Because many tasks in Natural Language Processing (NLP) are annotated on the token level, identifying the tokens is a crucial first step for NLP models. However, in most work on token-level tasks in NLP, gold tokenization is used, implicitly making the assumption that tokenization is a solved problem. Notable exceptions include the CoNLL 2018 shared task (Zeman et al., 2018) and work on languages where whitespaces are not used as word separators, and tokenization is more challenging (e.g. Tian et al., 2020; Hiraoka et al., 2020).

Traditionally, tokenization was done with rule-based systems (Marcus et al., 1993b; Dridan and Oepen, 2012), with rules usually adapted towards English datasets (Figure 1: 2). With the introduction of machine learning, and later neural networks, tokenization was also framed as a character level labeling task (Figure 1: 3) (Xue, 2003; Evang et al., 2013; Shao et al., 2018). However, since most recent NLP models are based on Contextualized Language Models (CLM), which commonly use subwords, subword level labeling for tokenization has been proposed (Nguyen et al., 2021) (Figure 1: 4), leading to even higher performance. However, Nguyen et al. (2021) do not extend to multi-lingual models, and their training procedure is compute intensive. Hence, we propose to tackle tokenization simultaneously with other NLP tasks while finetuning the CLM. This setup has competitive performance, while being universally applicable; we train one multi-task, multi-lingual model that does tokenization, pos tagging and dependency parsing; which is desirable in terms of efficiency, dependencies, and simplicity. We then use this model to evaluate and analyze the performance in a variety of setups.We tackle the following question in this work: 1) Is the tokenization task solved in supervised setups? 2) How robust are supervised tokenizers across datasets?

1)	`Dr. Dron is his backup.`							
2)	`s=[·][.][]])} > ”′]∗⋆$=\1 \2\3 =g`							
3)	`biiobiiiobiobiiobiiiiib`							
4)	Dr b	. i	Dro b	##n i	is b	his b	backup b	. b

Figure 1: Example sentence (1), regular expression tokenizing punctuation (2), sequence labeling on the character level (3), sequence labeling on the subword level (4). All of these strategies lead to the same tokenization: “Dr. Dron is his backup .”

2 The Tokenization Task

Since the increased popularity of subword tokens, the word “tokenization” is commonly used to re-

[1]Code is available on bitbucket.org/robvanderg/tok, note that our implementation is also available as part of the MaChAmp toolkit: https://github.com/machamp-nlp/

Findings of the Association for Computational Linguistics: EACL 2024, pages 118–137
March 17-22, 2024

Input: If␣momma␣ain’t␣happy,␣nobody␣ain’t␣happy.
Tokenization: If␣momma␣ain’t␣happy␣,␣nobody␣ain’t␣happy␣.
Multi-word expansions: If␣momma␣is␣not␣happy,␣nobody␣is␣not␣happy.
Subword segmentation: If␣mo␣##mma␣ai␣##n␣’␣t␣happy␣,␣no␣##body␣ai␣##n␣’␣t␣happy␣.

Table 1: Examples of the scope of tasks, we use the ␣ character to indicate whitespaces. The tokenization and multi-word expansion examples are from the UD, and the subword segmentation is based on mBERT, which does tokenization and subword segmentation. In UD, tokenization and multi-word expansions are annotated separately, but we do not consider multi-word expansions as part of the tokenization task.

fer to the task of subword segmentation. However, traditionally, “tokenization” referred to the task of identifying tokens in a segment of text. We follow the traditional usage, and follow the definition of token as used in the Universal Dependencies project (Zeman et al., 2022)[2], which to the best of our knowledge, is the largest and most diverse manually annotated dataset for this task. Furthermore, it has downstream tasks and tokenization annotated on the same utterances, which allows for more elaborate evaluations. We consider the transformation to *multiword tokens* (e.g. splitting clitics, undoing contractions) not to be part of the tokenization task. [3] We remove the multiword tokens with the UD-conversion tools (Agić et al., 2016), which propagates the annotations of the sub-token closest to root to the multiword token. An overview of the different tasks and the terminology we follow is shown in Table 1.

3 Tokenization with Subword-level Labels

Because the subword level is central in most modern language models, we label subwords for the tokenization task (Figure 1: 4). This approach has a limitation; there is a theoretical upper bound, as there is a limitation on the possible boundaries (i.e. splits are not possible within subwords). To increase this upper bound, we first apply the BasicTokenizer from the transformers library (Wolf et al., 2020), which is a rule-based tokenizer that separates punctuation characters. This leads to an upper bound above 99% F1 score for 122 out of 123 treebanks of the datasets we use (Appendix D) when using the mBERT subword segmenter (Devlin et al., 2019). Only the Japanese GSD treebank has a lower score (80.4). [4] To increase this upperbound, we consider all Hiragana and Katakana characters as a single subword (note that BERT tokenizers already do this for CJK characters, including Kanji). It should be noted that character normalizations and unknown tokens make the conversion of the output of the CLM to the original text non trivial. More details on how we handled these specific cases can be found in Appendix A.

If we would train a separate CLM for tokenization and one for a downstream task, this would lead to very inefficient training as well as inference. Note that they can’t run in paralellel, as tokenization should be done first. Hence, we propose a multi-task setup, where we share an encoder and model multiple tasks in separate decoder heads (linear layers). At train time, we use gold tokenization to obtain the loss for the other tasks, as labels for incorrect tokenizations are non-trivial to obtain. At inference time we use the predicted tokenization as input for the other tasks.

Setup We implemented our model in MaChAmp (van der Goot et al., 2021) v0.4.2, and have included it in the public version. We use all default parameters in MaChAmp (see Appendix B; note that we fully fine-tune the CLM in all our settings). We implemented tokenization with cross-entropy loss and a feedforward layer which transforms the output of the CLM to a binary label (B or I, see Figure 1). In the multi-task setup, we use the default implementations for UPOS tagging, lemmatization, morphological tagging and dependency parsing. We report F1 scores from the official CoNLL 2018 evaluation script (Zeman et al., 2018). We used UD v2.10 and multilingual BERT for our main evaluations. Note that we also evaluated on XLM-R Large (Conneau et al., 2020), but found that it underperforms for tokenization while being computationally more expensive (Appendix E).

We evaluate a variety of settings: **ST**: Single Task; an CLM encoder with only a tokenization head; **MT**: Multi-Task: learn tokenization simultaneously with POS tagging, lemmatization, morphological tagging and dependency parsing, **ML+MT**:

[2] https://universaldependencies.org/u/overview/tokenization.html

[3] In other words, we do not consider annotations where the word index contains a ‘-’, and we focus on the ‘tokens’ column in the evaluation script instead of ‘words’

[4] Short Unit Word tokenization (Den et al., 2008) was used for annotation of this dataset, which mismatches with the subword segmentation in mBERT.

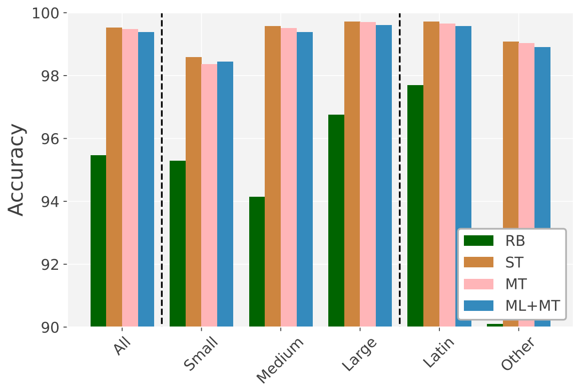

Figure 2: F1 scores for tokenization task (dev set). ST=Single Task (tokenization only), MT=Multi Task, RB=Rule-Based, ML=Multi-Lingual.

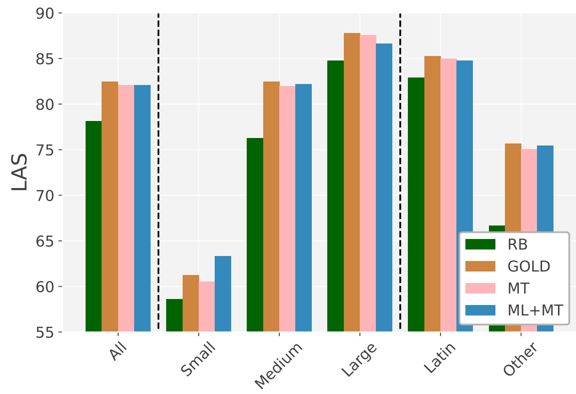

Figure 3: LAS F1 scores for dependency parsing (dev set). GOLD refers to using gold tokenization. Single Task (ST) is left out here, as it is an impractical in this setup (twice as slow, see Section 3).

Multi-Lingual, Multi-Task: train on the training splits of all treebanks for all tasks. To better interpret our results, we compare against five rule-based (**RB**) tokenizers (more information in Appendix G). We use the highest performing tokenizer (through an oracle) for each dataset.

4 Results

In this section we only consider treebanks that contain a train-split to be able to fairly compare to single-treebank models. We report averages over all dev splits (to avoid over analyzing the test data, note that we did not tune the models), but also averages over subsets of the data; we compare datasets in the Latin script (93 datasets) and all other scripts (38 datasets),[5] and we inspect the effect of dataset size by separating datasets in small (0<#tokens<20,000, 11 datasets), medium, (20,000<#tokens<100,000, 43 datasets) and large (>100,000, 51 datasets) train size. We focus here on tokenization and dependency parsing, results on other tasks can be found in Appendix F.

Starting with the results on tokenization (Figure 2), we can see that the differences in performance for the different settings are small for the tokenization task; but every error for this task has a catastrophic effect on downstream task performances, so even small differences can be important. The **single task setting (ST) outperforms all other models** in almost all setups. However, this setting is impractical due to computational costs. **Multi-task (MT) and Multi-lingual (ML) learning slightly harm performance**, but **Multi-lingual (ML) models outperform mono-lingual models on small datasets**. It should be noted that treebanks in non-Latin scripts are not consistently smaller (Appendix F), and the **lower performance on non-lating datasets can thus mainly be ascribed to under-representation in the underlying language model and the complexity of the task**. To interpret our results in a larger context, we attempt to compare to rule-based baselines; which are non-trivial to find for our varied set of languages (Appendix G), but it is clear that **rule-based approaches underperform with a large margin**; averages for all treebanks are around 91-92 F1.

Interestingly, downstream results on dependency parsing (Figure 3) show different trends compared to the tokenization results; **multi-lingual training (ML) is beneficial for this task**, except for large datasets which have slightly lower performance. Furthermore, we see that **the predicted tokenization performs very close to the gold tokenization** (GOLD) for parsing.

4.1 Test Data

We evaluate against the best rule-based tokenizers (RB) on the dev-data for each treebank; similary, we pick the best model of the CoNLL 2018 shared task (Zeman et al., 2018) for each treebank (UD v2.2); which are mostly Bi-LSTM character level BIO labelers. Finally, we compare to Trankit (Nguyen et al., 2021), who employ XLM-R with adapters (UD v2.5). [6] Results (Table 2) show that performance of our proposed model is on par

[5] Note that most other scripts contain less than 3 treebanks, we refer to Appendix F for per treebank results and % of unknown subwords

[6] Note that training Trankit for all tasks on UD_English-EWT was ~10 times slower compared to our approach with default parameters on an A100 GPU.

	Train treebanks			All		
	UD2.2	UD2.5	UD2.10	UD2.2	UD2.5	UD2.10
RB	95.98	94.99	94.40	91.67	91.67	92.71
SOTA	99.53	99.32	—	—	—	—
ST	99.42	99.41	99.39	—	—	—
ML+MT	99.33	99.31	99.09	97.59	97.18	95.64

Table 2: Average tokenization F1 scores on test data. SOTA on v2.2 is the highest score of each treebank in the CoNLL 2018 shared task, and v2.5 is Trankit. RB=RuleBased.

with the state-of-the-art both for UD v2.2 and v2.5. Furthermore, we confirm small loss in performance when training a multi-task, multi-lingual model (ML+MT) compared to the single task model (ST). Performance on all treebanks is substantially lower than the treebanks with a training split (lowest on UD v2.10, because there are more low-resource treebanks).

5 Analysis

Quantitative In general, precision is higher than recall for all the proposed models (results available in repository), showing that the model mostly misses splits instead of over-tokenizing. Performance detoriates on test-only treebanks (Table 3). As expected, performance is worst for treebanks in unseen scripts; however, F1 is still 80.11. For dependency parsing performances are much lower, this is mainly due to the amount of [UNK] tokens and the low coverage for these languages and scripts in mBERT training data.

Qualitative Latin data We picked the single task (ST) model for qualitative analysis to avoid any influence from the other adaptations. We selected the six lowest performing Latin treebanks. For Swedish_Sign_Language-SSLC (97.73), low performance is likely caused by non-standard use of capitalization and punctuation. For Estonian-EWT (97.93) inconsistency in splitting multiple periods was the main source of error, whereas in Romanian-Nonstandard (98.73), the '-' character is sometimes appended to the previous and sometimes to the following token, which is challenging for the model. The Dutch_Alpino treebank (99.17) has a mismatch between gold tokenization of numbers in the training and dev splits.[7] For Italian_PoSTWITA (99.47), we found cases where usernames, hashtags, URLs were wrongly tokenized by the model, and some cases similar to the errors found in English_EWT treebank (99.67), which are discussed in more detail in the following paragraph.

Common errors in the English EWT were due to ambiguity, for example, due to possesive markers being similar as the plural inflection; "salons ↦ salon␣s" was not tokenized by the model (but it was in gold), but "boys ↦ boy␣s" was. Other cases were difficult because of absence of any punctuation or white space clues: "so goand get dancing", "is there anyway", "andthere". In some cases, the model did not separate punctuation; "18+ ↦ 18␣+" "<>" ↦ "<␣>". Finally, there were also cases where the gold tokenization was inconsistent: "f/2 ↦ f/2", but "f/2.7↦f␣/␣2.7".

Qualitative Non-Latin data We manually inspected all treebanks with a performance <99 F1 score (11 total). For the treebanks that were included in previous work, performance of our model is highly competitive, indicating that these are generally challenging datasets. For four of the treebanks, the main issue where unknown subwords, due to special characters (Old East Slavic *2, Uyghur) or emojis (Russian); where the latter also had errors with Twitter usernames. We confirm this trend by checking the Pearson correlation between the % of unknown tokens and the performance for tokenization (F1) as well as the correlation between the % of unknown tokens and dependency parsing performance (LAS) on our full data (the % of unknown tokens for each treebank can be found in Table 15 in Appendix I). The correlations are -0.19, and -0.64, indicating that a higher percentage of unknown tokens indeed leads to worse tokenization (although dependency parsing is affected worse).

Vietnamese-VTB is a notoriously difficult treebank to tokenize in UD, due to tokens including whitespaces. For the Japanese and Chinese treebanks (five total); the problem of tokenization is harder, as there are no whitespaces and token segmentation can be a more ambiguous (i.e. subjective) task. For these languages,[8] we identified three main trends: 1) Adpositions: the model oversplits on adpositions, which are considered to be part of the word in the gold annotation. On the other hand, politeness markers for Japanese are usually attached to the word by the model (which is not con-

[7]We confirmed this with the treebank creators, this is the effect of merging datasets with different pre-processing

[8]We consulted native speakers for a qualitative inspection

setting	F1 tok.	F1 LAS	# treebanks
all	93.23	38.72	90
in-language	95.11	68.20	34
in-script	94.16	40.45	84
new-script	80.11	14.41	6

Table 3: Results on test-only treebanks, separated into treebanks with an in-language training treebank, an in-script training treebank, and neither (new-script).

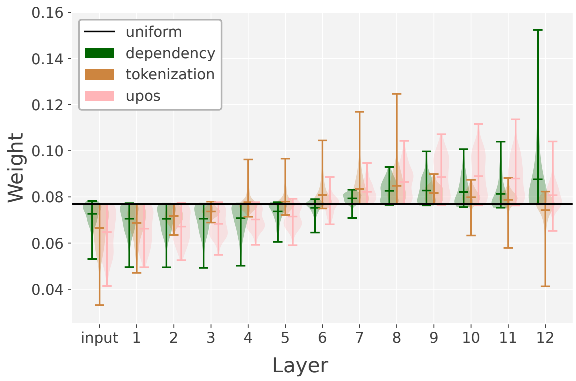

Figure 4: Violin plots of the attention at each layer for tokenization, UPOS tagging and dependency parsing for the MT+SPL models. Layer 'input' represent the (uncontextualized) word embeddings. Uniform weight (== no layer attention) would be $1/13 \approx 0.077$.

sistently the case in the treebanks) 2) Names: the model usually oversplits, For example for Japanese, the model splits "クモハ123-1" which is a train type, into: "クモ␣ハ␣123␣-␣1", because "クモ" can be read as the phoneticized "cloud" or "spider". . In general, for both Chinese and Japanese, names are often split into lexical tokens. 3) Compound words: for example 'homerun' (ホームラン) and 'copy protection' (コピープロテク) are not split by the model, but are split in the treebanks. Whereas for 'Kyoto-style' (京風) it is the other way around.

Rule-based baselines The performance of the rule-based baselines is substantially worse. Upon inspection, we found this is mainly due to 1) a different understanding of the tokenization task; rule based tokenizers consistently have different preferences (for example won't -> wo n't or ->won't) 2) scripts that were not considered while developing the tokenizers

Annotation consistency Our findings of the qualitative analyses indicate that annotation consistency is the main source of remaining errors for in-dataset settings, especially for Latin datasets. This is underlined by the the scores on test-only treebanks with in-language training data available; where F1 is only 95.11 (Table 3). It should be noted that another possible explanation is domain transfer, but our manual inspection suggested that annotation consistencies are the main source.

Attention To investigate where in the model the tokenization task is best represented, we analyze in which layer the tokenization task is best learned for the MT+SPL models. Instead of using a probing method (e.g. Tenney et al., 2019), we choose to use layer attention, (as implemented by Kondratyuk and Straka (2019), with the hope of improving performance further[9], saving computation costs, and finding the importance of each layer as assigned by the model itself. Results (Figure 4) show that tokenization is better presented in the middle layers (4-8). This suggests that context is necessary to perform this task (the input layer has a very low weight).

[9]Performance went down a little instead (Appendix F).

6 Conclusion

We have investigated which problems are still open for the task of tokenization. We conclude that tokenization in supervised setups for Latin languages can be considered solved, with some dataset inconsistencies as remaining errors. But for lower-resource languages and especially languages without whitespaces for word boundaries challenges remain. Furthermore, we showed that performance in cross-dataset setups deteriorates, even when training on the target language. This highlights the need for clear annotation guidelines, and confirms the presence of annotation inconsistenties.

Furthermore, we have implemented a new tokenization model that is faster to train than previous work. We include handling of unknown tokens and character normalizations as well as missed word boundaries. Furthermore, multi-task learning as well as multi-lingual learning slightly harm performance, but allow for a single model for multiple tasks and languages.

7 Acknowledgements

I would like to thank my colleagues at NLP North and MaiNLP and most of the anonymous reviewers for their feedback on earlier versions of this paper. I acknowledge the IT University of Copenhagen

HPC resources made available for conducting the research reported in this paper. Furthermore, special thanks go to Yiping Duan and Max Müller-Eberstein for the qualitative analysis of Chinese and Japanese.

8 Limitations

In our experiments, we have mainly focused on mBERT, we also evaluated on XLM-R Large (Appendix E), but for tokenization mBERT performs highly competitive while being computationally cheaper. We did test our implementation with other language models as well, but due to computational limitations we have not done the full evaluations. Furthermore, we were limited to evaluate on languages for which annotated data is available (including 20 of the 165 scripts defined in Unicode). It should be noted that we have limited ourselves to the definition of UD for the tokenization task.

We also only focused on syntactic downstream tasks, as annotation was readily available, although we do believe that the main gains from correct tokenization do not come from the shared parameters, but from having the correct word-boundaries. It should be noted that some of the datasets are created using automatic tokenization, and parts of the data can thus be considered silver (this is unfortunately not documented per treebank, as for other tasks in UD). Other datasets are trivial to tokenize, for example sign language (which includes transcriptions of signs) and treebanks on transcribed spoken data (without punctuation). However, even in these setups, it is important to have a tokenizer that mimics the treebank standard and that is consistent, and the original tokenizer that was used to create the data is often unknown or not available anymore. We did not perform significance testing, because to do this properly, multiple runs would have to be done (Dror et al., 2019), which is computationally expensive. Furthermore, multiple runs from previous work are not available, and due the size of the datasets used, even small differences will usually lead to significant differences.

Recently, character and byte level language models have been proposed(e.g. Xue et al., 2022; Clark et al., 2022), which do not have the theoretical upper-bound discussed in Section 3. However, their performance on syntactic word-level tasks was empirically not on par with the subword-based models (see Appendix C). Further improvements on downstream tasks might be obtained by using predicted tokenization during training. However, the current evaluation metrics do not take incorrectly tokenized tokens into account for the downstream tasks, and it is non-trivial to obtain a loss for downstream takss on a non-perfect tokenization.

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

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.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

Rebecca Dridan and Stephan Oepen. 2012. Tokenization: Returning to a long solved problem — a survey, contrastive experiment, recommendations, and toolkit —. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–382, Jeju Island, Korea. Association for Computational Linguistics.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence, Italy. Association for Computational Linguistics.

Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426, Seattle, Washington, USA. Association for Computational Linguistics.

Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. 2020. Optimizing word segmentation for downstream task. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1341–1351, Online. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993a. Building a large annotated corpus of english: The penn treebank.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993b. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Yan Shao, Christian Hardmeier, and Joakim Nivre. 2018. Universal word segmentation: Implementation and interpretation. *Transactions of the Association for Computational Linguistics*, 6:421–435.

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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020. Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online. 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.

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48.

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, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Erik Andersen, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas

Arıcan, Ĥórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, 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, Kepa Bengoetxea, Yifat Ben Moshe, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, 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, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, 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, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, 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, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, 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, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóğa, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, 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, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, 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, 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, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot,

Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, 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, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Vivian Stamou, Steinh͡ór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, 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, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, 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, Anna Zhuravleva, and Rayan Ziane. 2022. Universal dependencies 2.10. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

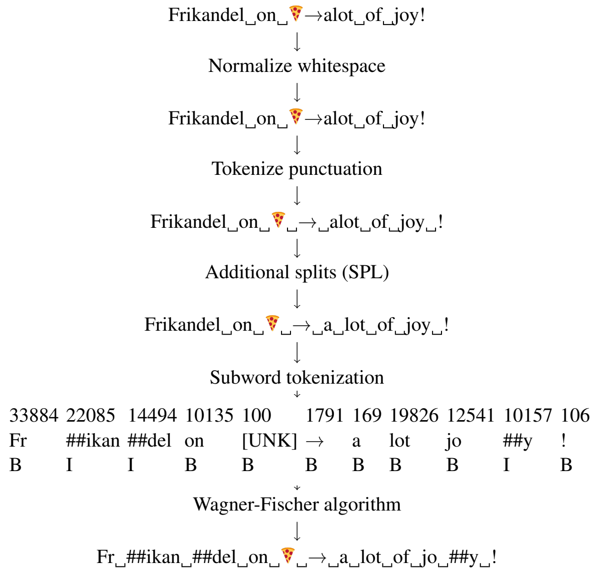

Figure 5: Detailed overview of the steps of proposed tokenization model.

A Detailed Overview of Model

The steps of our proposed tokenization procedure is shown in Figure 5. We start with whitespace normalization, converting all whitespace characters (tabs, no-break space etc.) to normal whitespaces, so that they are threated equally in the subword segmentation (There are no changes in our example, most input does not contain non-standard whitespaces). The next step is a basic tokenization based on punctuation, we use the `BasicTokenizer` from huggingface for this step (with `strip_accents=False`, `do_lower_case=False`, `tokenize_chinese_chars=True`). Next, we perform additional splits learned from the training data. This is done to overcome the upperbound because of the limitation that we can only split on subword boundaries (e.g. if 'alot' is split into 'al' and 'ot' by the subword tokenizer, there is no correct tokenization possible). We automatically extract all missed word-boundaries within words (e.g. alot $\mapsto$ a lot) from the *training* data. These additional splits lead to higher upper bounds on the development data for some datasets (Appendix D), but eventually harmed performance in more cases, so they are not included in the results reported in the paper. In the appendix we use **SPL** to indicate runs that use these additional splits. Then, we use the slow subword tokenizer from Huggingface, and set `do_basic_tokenize` to false.

We require one last step, because most language models do some (Unicode) normalization on the data and include special unknown tokens to represent (sequences of) characters that were unseen during the training of the tokenizer. These break the evaluation of tokenization, as no alignment between the gold tokenization and the prediction can be found. To solve this, we align the subwords to the original input automatically. This mapping is non-trivial, and we empirically found that character edit rules are a robust solution for this. We use the Wagner-Fischer (Wagner and Fischer, 1974) algorithm as implemented by (Straka, 2018). We calculate the character edit transformation from the segmented subwords to the original text (after removing whitespaces for both), and insert or substitute characters that differ.

Parameter	Value
Optimizer	Adam
β_1, β_2	0.9, 0.99
Dropout	0.2
Epochs	20
Batch size	32
Learning rate (LR)	1e-4
LR scheduler	slanted triangular
Weight decay	0.01
Decay factor	0.38
Cut fraction	0.3

Table 4: Hyperparameter settings (taken from MaChAmp v0.4beta).

B Hyperparameters

Hyperparameters we used for all experiments are reported in Table 4, and match the default settings of MaChAmp 0.4 (van der Goot et al., 2021). Note that no early stopping is used, because the learning rate scheduler lowers the learning rate dynamically; so even if performance does not improve in the current epoch, it might still improve in future epochs.

C Results Character-level Models

We experimented with character/byte level models in a similar setup for a selected set of treebanks. We picked treebanks that are challenging (Chinese/Japanese treebanks), even when trained in-dataset, as well as a common benchmark (English-EWT). Results are shown in Table 5 for the tokenization task, and Table 6 for downstream performance on dependency parsing. Results show that mBERT substantially outperforms both other

Treebank	mBERT	byt5-base	Canine-C
UD_Chinese-GSD	99.09	88.49	93.98
UD_Chinese-GSDSimp	99.10	88.53	94.07
UD_Classical_Chinese-Kyoto	98.16	98.71	-
UD_English-EWT	99.81	99.59	98.25
UD_Japanese-GSDLUW	99.36	93.00	98.78
UD_Japanese-GSD	99.30	91.33	97.92

Table 5: Tokenization F1 scores for character level models versus mBERT

Treebank	mBERT	byt5-base	Canine-C
UD_Chinese-GSD	84.95	80.28	59.90
UD_Chinese-GSDSimp	84.94	81.20	59.67
UD_Classical_Chinese-Kyoto	78.70	77.68	56.32
UD_English-EWT	90.04	89.30	79.10
UD_Japanese-GSDLUW	94.71	93.97	90.16
UD_Japanese-GSD	94.48	93.83	89.66

Table 6: LAS scores for character level models and mBERT

models, but Canine-C seems to be better at tokenization and byt5-base at parsing. To avoid waste of compute, we decided to not train byt5-base and Canine-C on the rest of the data.

D Upper Bound

Table 7 shows the theoretical upper bound of performance of the tokenization task for each treebank in UD 2.10. The table shows the upper bound on the training and the dev data, and also shows the performance after extracting the splits for impossible cases from the training data (for example "alot ↦ al ##ot" make it impossible to get "a lot", see also Section 3 and Appendix A).

E Comparison mBERT to XLM-R Large

In Table 8 we compare the scores for all 5 tasks for all treebanks with a training split in UD v2.10. Results show that XLM-R large (Conneau et al., 2020) is substantially better than mBERT for most tasks; however, for tokenization it only outperforms mBERT in the single task setting.

F Full Scores Tokenization

Per treebank results on UD v2.10 dev splits for all our proposed models are shown in Table 9.

G Scores Rule-based Baselines

We used the `BasicTokenizer` from the Transformers library (Wolf et al., 2020), without normalization. The other rule-based tokenizers are all taken from NLTK (Bird et al., 2009). `Destructive` is an extended version of the TreebankTokenizer, which in turn is a python version of the `tokenizer.sed` script originally used for the Penn Treebank (Marcus et al., 1993a). The TweetTokenizer is a tokenizer focused on data from Twitter, and Toktok is a fast simple tokenizer based on regular expressions. We automatically checked the output for changed characters and reverted these using the strategy described in Appendix A. Results (Table 10) show that altough for some treebanks performance around 99-100 F1 can be achieved, average performance is around 91-92%, which is substantially lower compared to the supervised results in Table 9. There are some outliers dragging the average down,[10] but also many treebanks with scores in the mid- and low 90's. Interestingly, for some treebanks 100% was achieved only by the rule-based models;[11] these are treebanks for which the gold tokenization is most likely automatically created.

H Scores on Other Tasks

We include performance on the other UD tasks included in our multi-task model. Dependency parsing in Table 11, UPOS tagging in Table 12, Morphological tags in 13, Lemmatization in 14. All reported scores are obtained with the official conll 2018 script.

I Full Scores on Test data

In Table 15 we report the performance of ST and MT-ML on the test splits of UD v2.2, v2.5 and v2.10 per treebank.

[10] Chinese, Japanese, Maltese, Old east Slavic (Birchbark) Swedish Sign Language, and Vietnamese treebanks.

[11] Ancient Greek (*2), Czech-CAC, Latin-PROIEL, Old Church Slavonic, and Tamil treebanks

Treebank	dev	+splits	#splits	Treebank	dev	+splits	#splits
UD_Afrikaans-AfriBooms	100.0000	100.0000	0	UD_Japanese-BCCWJLUW	100.0000	100.0000	0
UD_Ancient_Greek-PROIEL	100.0000	100.0000	0	UD_Japanese-GSD	99.1478	99.1478	514
UD_Ancient_Greek-Perseus	100.0000	100.0000	0	UD_Japanese-GSDLUW	99.1385	99.1385	421
UD_Ancient_Hebrew-PTNK	100.0000	100.0000	0	UD_Korean-GSD	99.8244	99.8285	36
UD_Arabic-NYUAD	100.0000	100.0000	0	UD_Korean-Kaist	100.0000	100.0000	0
UD_Arabic-PADT	100.0000	100.0000	0	UD_Latin-ITTB	100.0000	100.0000	0
UD_Armenian-ArmTDP	100.0000	100.0000	0	UD_Latin-LLCT	100.0000	100.0000	0
UD_Armenian-BSUT	100.0000	100.0000	4	UD_Latin-PROIEL	100.0000	100.0000	0
UD_Basque-BDT	100.0000	100.0000	0	UD_Latin-UDante	100.0000	100.0000	0
UD_Belarusian-HSE	99.9435	99.9435	311	UD_Latvian-LVTB	100.0000	100.0000	3
UD_Bulgarian-BTB	100.0000	100.0000	0	UD_Lithuanian-ALKSNIS	100.0000	100.0000	0
UD_Catalan-AnCora	100.0000	100.0000	0	UD_Lithuanian-HSE	100.0000	100.0000	0
UD_Chinese-GSD	100.0000	100.0000	0	UD_Maltese-MUDT	99.9804	99.9804	0
UD_Chinese-GSDSimp	100.0000	100.0000	0	UD_Marathi-UFAL	100.0000	100.0000	0
UD_Classical_Chinese-Kyoto	100.0000	100.0000	0	UD_Naija-NSC	99.9177	100.0000	3
UD_Coptic-Scriptorium	100.0000	100.0000	0	UD_Norwegian-Bokmaal	100.0000	100.0000	3
UD_Croatian-SET	100.0000	100.0000	0	UD_Norwegian-Nynorsk	100.0000	100.0000	2
UD_Czech-CAC	100.0000	100.0000	33	UD_Norwegian-NynorskLIA	100.0000	100.0000	0
UD_Czech-CLTT	99.9583	99.9583	1	UD_Old_Church_Slavonic-PROIEL	100.0000	100.0000	0
UD_Czech-FicTree	100.0000	100.0000	3	UD_Old_East_Slavic-Birchbark	99.6482	99.6482	4
UD_Czech-PDT	100.0000	100.0000	41	UD_Old_East_Slavic-TOROT	100.0000	100.0000	0
UD_Danish-DDT	100.0000	100.0000	0	UD_Old_French-SRCMF	100.0000	100.0000	0
UD_Dutch-Alpino	100.0000	100.0000	0	UD_Persian-PerDT	100.0000	100.0000	0
UD_Dutch-LassySmall	100.0000	100.0000	0	UD_Persian-Seraji	100.0000	100.0000	1
UD_English-Atis	100.0000	100.0000	0	UD_Polish-LFG	99.3590	99.7100	251
UD_English-ESL	100.0000	100.0000	0	UD_Polish-PDB	100.0000	100.0000	7
UD_English-EWT	99.9516	99.9839	17	UD_Pomak-Philotis	100.0000	100.0000	0
UD_English-GUM	100.0000	100.0000	4	UD_Portuguese-Bosque	100.0000	100.0000	1
UD_English-GUMReddit	100.0000	100.0000	0	UD_Portuguese-GSD	100.0000	100.0000	0
UD_English-LinES	99.6035	100.0000	14	UD_Romanian-Nonstandard	99.9785	99.9785	6
UD_English-ParTUT	100.0000	100.0000	7	UD_Romanian-RRT	100.0000	100.0000	0
UD_Estonian-EDT	100.0000	100.0000	0	UD_Romanian-SiMoNERo	100.0000	100.0000	0
UD_Estonian-EWT	99.9800	99.9800	8	UD_Russian-GSD	100.0000	100.0000	2
UD_Faroese-FarPaHC	99.8684	99.9371	5	UD_Russian-SynTagRus	99.9954	99.9967	14
UD_Finnish-FTB	100.0000	100.0000	0	UD_Russian-Taiga	99.9406	99.9406	101
UD_Finnish-TDT	100.0000	100.0000	2	UD_Scottish_Gaelic-ARCOSG	100.0000	100.0000	0
UD_French-FTB	100.0000	100.0000	0	UD_Serbian-SET	100.0000	100.0000	0
UD_French-GSD	99.9899	99.9899	16	UD_Slovak-SNK	100.0000	100.0000	0
UD_French-ParTUT	100.0000	100.0000	5	UD_Slovenian-SSJ	100.0000	100.0000	2
UD_French-Rhapsodie	100.0000	100.0000	0	UD_Spanish-AnCora	100.0000	100.0000	1
UD_French-Sequoia	99.9794	99.9794	0	UD_Spanish-GSD	100.0000	100.0000	3
UD_Galician-CTG	99.9926	99.9926	4	UD_Swedish-LinES	100.0000	100.0000	0
UD_German-GSD	100.0000	100.0000	2	UD_Swedish-Talbanken	100.0000	100.0000	0
UD_German-HDT	100.0000	100.0000	1	UD_Swedish_Sign_Language-SSLC	100.0000	100.0000	0
UD_Gothic-PROIEL	100.0000	100.0000	0	UD_Tamil-TTB	100.0000	100.0000	0
UD_Greek-GDT	100.0000	100.0000	0	UD_Telugu-MTG	100.0000	100.0000	0
UD_Hebrew-HTB	100.0000	100.0000	0	UD_Turkish-Atis	100.0000	100.0000	0
UD_Hebrew-IAHLTwiki	99.9783	99.9783	0	UD_Turkish-BOUN	99.9582	99.9708	13
UD_Hindi-HDTB	100.0000	100.0000	0	UD_Turkish-FrameNet	100.0000	100.0000	0
UD_Hindi_English-HIENCS	100.0000	100.0000	0	UD_Turkish-IMST	100.0000	100.0000	0
UD_Hungarian-Szeged	100.0000	100.0000	0	UD_Turkish-Kenet	100.0000	100.0000	0
UD_Icelandic-IcePaHC	99.9885	99.9957	26	UD_Turkish-Penn	100.0000	100.0000	0
UD_Icelandic-Modern	99.9444	100.0000	17	UD_Turkish-Tourism	100.0000	100.0000	0
UD_Indonesian-GSD	100.0000	100.0000	3	UD_Turkish_German-SAGT	100.0000	100.0000	0
UD_Irish-IDT	100.0000	100.0000	0	UD_Ukrainian-IU	99.9841	99.9841	2
UD_Italian-ISDT	100.0000	100.0000	0	UD_Urdu-UDTB	100.0000	100.0000	0
UD_Italian-MarkIT	100.0000	100.0000	0	UD_Uyghur-UDT	100.0000	100.0000	0
UD_Italian-ParTUT	100.0000	100.0000	6	UD_Vietnamese-VTB	100.0000	100.0000	0
UD_Italian-PoSTWITA	99.9535	99.9535	13	UD_Welsh-CCG	99.9555	99.9555	2
UD_Italian-TWITTIRO	100.0000	100.0000	2	UD_Western_Armenian-ArmTDP	100.0000	100.0000	0
UD_Italian-VIT	100.0000	100.0000	0	UD_Wolof-WTB	100.0000	100.0000	0
UD_Japanese-BCCWJ	100.0000	100.0000	0				

Table 7: Upper bounds of performance of development splits of UD 2.10 treebanks with mBERT (‘bert-base-multilingual-cased’). * For Japanese_GSD, we achieved 80.3969 and 92.1994 respectively (with 6,266 splits) without splitting each character (Section 3).

Task	CLM	ST	MT	MT+SPL	MT+SPL+LA	MT+ML	MT+ML+SPL
Tokenization	mBERT	**99.4782**	98.6299	98.5744	98.9350	99.0533	99.0319
	XLM-R L.	**99.5204**	98.6018	98.5031	98.5509	99.0472	99.0274
Dependency	mBERT		**81.5181**	81.4892	79.9496	81.2555	81.1588
	XLM-R L.		**85.0159**	84.1389	80.1694	81.3341	81.1333
UPOS	mBERT		93.7492	93.7111	**93.8782**	93.6883	93.6524
	XLM-R L.		**95.0951**	94.5530	94.6112	93.6962	93.6305
UFeats	mBERT		89.9223	89.9172	**90.6450**	85.5533	85.3939
	XLM-R L.		**92.2903**	92.1143	91.3762	85.5791	85.4916
Lemma	mBERT		89.8071	89.8243	90.9796	**90.9957**	90.9396
	XLM-R L.		91.4172	91.2470	**91.6976**	91.0358	90.9591

Table 8: Results of mBERT versus XLM-R large for all tasks considered in this paper.

Treebank	train_size	dev_size	script	ST	MT	MT+SPL	MT+SPL+LA	MT+ML	MT+ML+SPL
UD_Afrikaans-AfriBooms	33880	5317	Latin	99.6801	**99.7461**	99.6802	99.6708	99.5483	99.5483
UD_Ancient_Greek-PROIEL	187033	13652	Greek	**99.9670**	99.9414	**99.9670**	99.9561	99.9451	99.9451
UD_Ancient_Greek-Perseus	159895	22135	Greek	99.7178	**99.9729**	99.5911	99.5436	99.9593	99.9593
UD_Ancient_Hebrew-PTNK	12530	7340	Hebrew	**99.9728**	**99.9728**	**99.9728**	99.9319	**99.9728**	**99.9728**
UD_Arabic-PADT	191869	25986	Arabic	99.9211	**99.9731**	99.8980	99.9115	99.9577	99.9577
UD_Armenian-ArmTDP	41801	5348	Armenian	**99.8785**	99.7662	99.8224	99.7944	99.7571	99.7571
UD_Armenian-BSUT	21024	10267	Armenian	**99.7858**	99.6790	99.5816	99.6983	99.0757	99.0856
UD_Basque-BDT	72974	24095	Latin	**99.9647**	99.9378	99.9523	99.9357	99.9128	99.9128
UD_Belarusian-HSE	273172	15931	Cyrillic	**99.2842**	99.2118	99.1931	99.1334	99.2180	99.0667
UD_Bulgarian-BTB	124336	16089	Cyrillic	**99.8912**	99.8881	99.8726	99.8415	99.8601	99.8601
UD_Catalan-AnCora	416680	56322	Latin	99.8970	99.9014	**99.9059**	**99.9059**	99.8908	99.8908
UD_Chinese-GSD	98616	12663	Han	**98.1187**	97.8495	97.7687	97.8950	96.9087	96.9087
UD_Chinese-GSDSimp	98616	12663	Han	**98.1128**	97.7134	97.7144	97.7999	96.9089	97.0802
UD_Classical_Chinese-Kyoto	236067	28793	Han	97.2069	97.4586	97.4586	**97.6134**	97.1215	97.0453
UD_Coptic-Scriptorium	14581	5165	Coptic	99.9419	**99.9710**	99.9419	99.9419	**99.9710**	**99.9710**
UD_Croatian-SET	152857	22292	Latin	**99.8475**	99.8161	99.8318	99.7959	99.8430	99.8430
UD_Czech-CAC	471594	10888	Latin	**99.9862**	**99.9862**	**99.9862**	**99.9862**	**99.9862**	**99.9862**
UD_Czech-CLTT	27752	4800	Latin	99.7395	99.6562	99.7396	**99.7707**	99.4381	99.3346
UD_Czech-FicTree	133137	16652	Latin	**100.0000**	99.9910	**100.0000**	**100.0000**	99.9730	99.9910
UD_Czech-PDT	1171190	158958	Latin	99.9902	99.9858	**99.9918**	99.9912	99.9597	99.9572
UD_Danish-DDT	80378	10332	Latin	99.7532	**99.7725**	99.7386	99.6950	99.7532	99.7532
UD_Dutch-Alpino	186026	11541	Latin	99.1749	99.1750	**99.1751**	99.1446	99.1190	99.1190
UD_Dutch-LassySmall	75134	11397	Latin	99.7895	99.7324	**99.8464**	99.7544	99.6931	99.6931
UD_English-Atis	48655	6644	Latin	**100.0000**	**100.0000**	**100.0000**	**100.0000**	99.9774	99.9548
UD_English-EWT	201962	24788	Latin	**99.6671**	99.6247	99.5177	99.6107	99.4512	98.8592
UD_English-GUM	123243	19337	Latin	**99.8888**	99.8759	99.8759	99.8319	99.4751	99.0058
UD_English-LinES	57372	19170	Latin	**99.9452**	99.5072	99.8905	99.9166	98.6138	98.8698
UD_English-ParTUT	43477	2721	Latin	**99.7796**	99.3748	99.6694	99.7060	98.6893	98.8005
UD_Estonian-EDT	344613	44748	Latin	99.4486	**99.4872**	99.4436	99.4537	99.3719	99.3719
UD_Estonian-EWT	55073	10002	Latin	97.9300	**98.2639**	98.0380	97.8753	98.1716	97.9191
UD_Faroese-FarPaHC	23089	8739	Latin	99.6738	99.5193	99.6222	99.6909	99.6851	**99.7595**
UD_Finnish-FTB	127359	15694	Latin	99.8917	99.8917	99.8980	99.8758	**99.9267**	**99.9267**
UD_Finnish-TDT	162615	18290	Latin	99.5489	**99.6090**	99.5954	99.5927	99.5681	99.5436
UD_French-GSD	344829	34646	Latin	**99.8975**	99.8946	99.8701	99.8744	99.8874	99.8773
UD_French-ParTUT	23312	1822	Latin	99.8354	99.8354	99.7531	99.9177	**100.0000**	**100.0000**
UD_French-Rhapsodie	18891	12757	Latin	**99.9295**	99.8746	99.8707	99.8589	99.8942	99.9059
UD_French-Sequoia	49145	9717	Latin	99.6500	99.5732	99.6143	99.5474	**99.8096**	**99.8096**
UD_Galician-CTG	71928	27009	Latin	**99.8056**	99.7722	99.7315	99.7667	99.7482	99.7037
UD_German-GSD	259184	12318	Latin	**99.9594**	99.8701	99.8742	99.8660	99.2713	99.1734
UD_German-HDT	2753627	319513	Latin	99.8715	**99.9078**	99.8729	99.8775	99.8559	99.8357
UD_Gothic-PROIEL	35024	10114	Latin	**100.0000**	99.9703	99.9555	99.9703	99.9852	99.9852
UD_Greek-GDT	41212	10139	Greek	**99.8374**	99.7045	99.7143	99.7438	99.7782	99.7782
UD_Hebrew-HTB	98344	8358	Hebrew	**100.0000**	99.9641	99.9462	99.9821	99.9162	99.9162
UD_Hebrew-IAHLTwiki	88527	6916	Hebrew	**99.7327**	99.7255	**99.7327**	99.6967	99.6893	99.7110
UD_Hindi-HDTB	281057	35217	Devanagari	**100.0000**	99.9957	99.9915	99.9915	99.9957	99.9957
UD_Hungarian-Szeged	20166	11418	Latin	**99.8818**	99.8511	99.7941	99.8380	99.8337	99.8337
UD_Icelandic-IcePaHC	704716	139384	Latin	99.8231	99.8274	99.8386	99.8095	**99.8518**	99.8429
UD_Icelandic-Modern	123853	17102	Latin	**99.9912**	99.9006	99.9795	99.9708	99.8859	**99.9912**
UD_Indonesian-GSD	95868	12423	Latin	**99.6218**	99.6217	99.5492	99.5129	99.4001	99.4001
UD_Irish-IDT	95881	10000	Latin	**99.8200**	99.6899	99.6950	99.6499	99.6900	99.6900
UD_Italian-ISDT	257616	11133	Latin	**99.9326**	99.8788	99.8877	99.8473	99.8023	99.8023
UD_Italian-MarkIT	18855	9824	Latin	**99.7762**	99.5935	99.5782	99.6389	99.5984	99.6035
UD_Italian-ParTUT	45477	2786	Latin	**99.9461**	99.8744	99.7846	99.6950	99.7666	99.8205
UD_Italian-PoSTWITA	95395	11825	Latin	**99.4714**	99.3106	99.3742	99.3405	99.3105	99.0692
UD_Italian-TWITTIRO	22656	2855	Latin	99.4574	99.3691	99.4744	99.4574	99.5450	**99.5624**
UD_Italian-VIT	208506	25964	Latin	**99.8845**	99.8711	99.8422	99.8710	99.8018	99.8422
UD_Japanese-GSD	168333	12287	Hiragana	97.8668	**98.0627**	97.6166	97.8180	70.3861	70.8063
UD_Japanese-GSDLUW	130284	9531	Hiragana	**97.7005**	97.6700	97.6558	97.6818	94.6654	93.9452
UD_Korean-GSD	56687	11958	Hangul	99.3394	**99.6654**	99.2138	99.2096	99.5818	99.2053
UD_Korean-Kaist	296446	25278	Hangul	99.9209	99.9466	99.9031	99.9327	**99.9506**	**99.9506**
UD_Latin-ITTB	390785	29888	Latin	**100.0000**	**100.0000**	99.9950	**100.0000**	99.9699	99.9699
UD_Latin-LLCT	194143	24189	Latin	**99.9752**	**99.9752**	**99.9752**	99.9690	99.9628	99.9504
UD_Latin-PROIEL	172133	13939	Latin	99.9641	99.9641	99.9534	99.9641	**99.9857**	**99.9857**
UD_Latin-UDante	30335	11550	Latin	**99.9870**	99.8311	99.8311	99.8571	99.8311	99.8571
UD_Latvian-LVTB	214983	31856	Latin	**99.9168**	99.8541	99.8682	99.8619	99.8462	99.8305
UD_Lithuanian-ALKSNIS	47641	11560	Latin	99.8486	99.8530	99.8746	**99.8875**	99.8660	99.8487
UD_Lithuanian-HSE	3210	1086	Latin	99.3116	98.7586	98.4814	98.7586	**99.6324**	**99.6324**
UD_Maltese-MUDT	22880	10209	Latin	99.7503	**99.8384**	99.8041	99.7649	99.5933	99.6129
UD_Marathi-UFAL	2730	400	Devanagari	**100.0000**	**100.0000**	**100.0000**	**100.0000**	**100.0000**	**100.0000**
UD_Naija-NSC	111877	14574	Latin	**99.8867**	99.7975	99.8456	99.8250	99.7734	99.8146
UD_Norwegian-Bokmaal	243886	36369	Latin	99.9148	99.9065	99.8969	99.8859	**99.9244**	99.9051
UD_Norwegian-Nynorsk	245330	31250	Latin	99.9488	99.9520	99.9568	99.9360	**99.9664**	99.9632
UD_Norwegian-NynorskLIA	35207	10163	Latin	**99.8770**	99.8475	99.8328	99.8180	99.8327	99.8327
UD_Old_Church_Slavonic-PROIEL	37432	10100	Cyrillic	99.2983	99.9653	99.1405	99.0861	**99.9851**	**99.9851**
UD_Old_East_Slavic-Birchbark	7256	9951	Cyrillic	**85.3118**	83.9754	83.5902	82.9414	83.9948	83.3856
UD_Old_East_Slavic-TOROT	118630	15791	Cyrillic	98.7604	**99.9398**	98.4683	98.5657	99.9145	98.5188
UD_Old_French-SRCMF	158620	20553	Latin	**99.9927**	99.9854	99.9854	99.9854	99.9708	99.9708
UD_Persian-PerDT	445587	24751	Arabic	99.9212	**99.9576**	99.9333	99.9152	99.9556	99.8990
UD_Persian-Seraji	119945	15755	Arabic	99.9810	**100.0000**	99.9238	99.9238	99.9810	99.9810
UD_Polish-LFG	104750	13105	Latin	**99.7518**	99.3322	99.7366	99.7366	98.7874	99.1238
UD_Polish-PDB	279596	34429	Latin	99.9172	**99.9332**	99.9100	99.9114	99.6677	99.7329
UD_Pomak-Philotis	69223	8753	Latin	**100.0000**	99.9600	99.9600	99.9600	99.9600	99.9258
UD_Portuguese-Bosque	158985	26384	Latin	**99.8465**	99.8427	99.8446	99.8427	99.4948	99.5363
UD_Portuguese-GSD	237924	29772	Latin	99.9043	**99.9144**	99.8690	99.8405	99.6251	99.6251
UD_Romanian-Nonstandard	532881	18569	Latin	**98.7260**	98.5779	98.6722	98.5670	98.7211	98.5345
UD_Romanian-RRT	185113	17073	Latin	99.5899	**99.6017**	99.5313	99.5841	99.4463	99.4463
UD_Romanian-SiMoNERo	116857	14611	Latin	99.4727	99.4255	99.4800	99.5450	**99.5585**	99.2988
UD_Russian-GSD	74900	11709	Cyrillic	99.7181	**99.7352**	99.6285	99.6924	99.3635	99.3635
UD_Russian-SynTagRus	1204640	153325	Cyrillic	99.7871	**99.7965**	99.7857	99.7926	99.7440	99.7567
UD_Russian-Taiga	176631	10096	Cyrillic	**98.3634**	97.5559	97.5261	97.6717	96.3383	97.8517
UD_Scottish_Gaelic-ARCOSG	65721	10226	Latin	**99.5890**	99.5057	99.5499	99.3932	99.5402	99.5402
UD_Serbian-SET	74259	11993	Latin	99.8498	99.8707	99.8624	99.8332	**99.8916**	**99.8916**
UD_Slovak-SNK	80575	12733	Latin	99.9647	**100.0000**	99.9647	99.9725	99.8586	99.8586
UD_Slovenian-SSJ	215155	26500	Latin	**99.9830**	99.9755	99.9698	99.9642	99.9208	99.9208
UD_Spanish-AnCora	442591	52176	Latin	**99.9531**	99.9253	99.9262	99.9195	99.8429	99.8429
UD_Spanish-GSD	375147	36464	Latin	**99.9369**	99.8313	99.9095	99.9246	99.7736	99.8793
UD_Swedish-LinES	55451	18515	Latin	**99.9514**	99.9271	99.9217	**99.9514**	99.9217	99.9217
UD_Swedish-Talbanken	66646	9797	Latin	**99.8724**	99.8316	99.8316	99.8112	99.8366	99.8366
UD_Swedish_Sign_Language-SSLC	644	684	Latin	**97.7256**	5.4863	3.0341	44.4444	95.2864	95.2864
UD_Tamil-TTB	5734	1129	Tamil	99.4690	99.3354	99.3366	99.4695	**99.7345**	**99.7345**
UD_Telugu-MTG	5082	662	Telugu	**99.7736**	**99.7736**	**99.7736**	**99.7736**	**99.7736**	**99.7736**
UD_Turkish-Atis	36200	4862	Latin	**99.9074**	99.8766	99.8150	99.8766	99.7633	99.7324
UD_Turkish-BOUN	97257	11974	Latin	**99.2319**	99.0192	98.9816	99.0232	98.8947	98.8563
UD_Turkish-FrameNet	16333	1421	Latin	**100.0000**	99.8944	99.8944	**100.0000**	99.8592	**100.0000**
UD_Turkish-IMST	36822	9777	Latin	**99.9642**	99.9335	99.9335	99.8977	99.8209	99.6365
UD_Turkish-Kenet	143287	17554	Latin	**100.0000**	99.9915	99.9915	99.9829	99.9630	99.9658
UD_Turkish-Penn	166514	6994	Latin	98.9807	98.8227	98.8588	**98.9866**	98.6878	98.6589
UD_Turkish-Tourism	71141	10203	Latin	99.9853	99.9853	99.9706	99.9853	**100.0000**	**100.0000**
UD_Turkish_German-SAGT	10005	12959	Latin	**99.8881**	99.7298	99.7491	99.7297	99.6531	99.7570
UD_Ukrainian-IU	92355	12573	Cyrillic	**99.8211**	99.7693	99.7654	99.7773	99.6579	99.6419
UD_Urdu-UDTB	108690	14581	Arabic	**99.9486**	99.9211	99.8937	99.8868	99.9074	99.9074
UD_Uyghur-UDT	19262	10644	Arabic	**98.5743**	98.1152	98.0338	98.2202	98.0190	98.0190
UD_Vietnamese-VTB	20285	11514	Latin	**94.4922**	93.7305	93.9136	93.7197	92.8122	92.8122
UD_Welsh-CCG	18522	8991	Latin	**99.7052**	99.6440	99.5773	99.5994	99.6996	99.6439
UD_Western_Armenian-ArmTDP	94893	13261	Armenian	**99.9661**	99.9208	99.8831	99.9095	99.9208	99.9434
UD_Wolof-WTB	22817	9966	Latin	**99.8695**	99.7039	99.6888	99.6888	99.8495	99.7893
Average	**172068.2500**	21387.6379	0.0000	99.4782	98.6299	98.5744	98.9350	99.0533	99.0319

Table 9: Full results on tokenization of dev sets (F1). ST=Single Task (tokenization only), MT=Multi Task, SPL=learn additional SPLits from training data, ML=MultiLingual, LA=Layer Attention. Train and dev sizes are in number of words, and script is estimated based on the most frequent unicode category.

Treebank	scripts	BasicTokenizer	Destructive	TweetTokenizer	Toktok	TreebankTokenizer
UD_Afrikaans-AfriBooms	Latin	95.7197	99.6150	97.1971	97.4914	**99.6150**
UD_Ancient_Greek-PROIEL	Greek	99.0144	99.0144	99.0144	99.0144	**100.0000**
UD_Ancient_Greek-Perseus	Greek	99.9864	97.7400	**100.0000**	97.7400	97.7400
UD_Ancient_Hebrew-PTNK	Hebrew	99.9728	61.9607	**99.9728**	61.9607	61.9607
UD_Arabic-PADT	Arabic	97.6019	95.0274	**98.0955**	97.3448	94.9637
UD_Armenian-ArmTDP	Armenian	96.9703	91.8961	**97.0092**	90.9442	89.1156
UD_Armenian-BSUT	Armenian	97.6595	90.9219	**97.5006**	89.6702	88.2422
UD_Basque-BDT	Latin	96.8780	99.8548	99.3666	99.7160	**99.8237**
UD_Belarusian-HSE	Cyrillic	88.6854	94.2065	**96.9833**	94.2495	91.3998
UD_Bulgarian-BTB	Cyrillic	96.6032	99.7142	98.7934	**99.7142**	**99.7142**
UD_Catalan-AnCora	Latin	90.7046	93.0735	92.8945	**93.8417**	93.0685
UD_Chinese-GSD	Han	22.1135	0.2268	**23.9392**	0.3750	0.2117
UD_Chinese-GSDSimp	Han	22.1135	1.8070	**23.9254**	1.0918	0.2117
UD_Classical_Chinese-Kyoto	Han	2.2796	2.2796	**2.2796**	**2.2796**	**2.2796**
UD_Coptic-Scriptorium	Coptic	99.9710	99.9323	**99.9323**	**99.9323**	**99.9323**
UD_Croatian-SET	Latin	95.9080	99.7981	98.6165	**99.8431**	99.7847
UD_Czech-CAC	Latin	100.0000	99.9035	**100.0000**	99.9311	99.9035
UD_Czech-CLTT	Latin	90.6262	93.7449	91.8217	**93.5576**	93.3701
UD_Czech-FicTree	Latin	97.1172	99.6602	**99.7354**	99.6180	99.6572
UD_Czech-PDT	Latin	98.8252	98.0831	**99.2227**	98.1900	98.0723
UD_Danish-DDT	Latin	96.2620	99.7532	98.7377	99.6277	**99.7773**
UD_Dutch-Alpino	Latin	96.6784	98.0673	97.7014	**98.1065**	98.0542
UD_Dutch-LassySmall	Latin	93.4003	99.3911	98.7131	99.1736	**99.3779**
UD_English-Atis	Latin	98.0498	100.0000	98.4056	98.5405	**100.0000**
UD_English-EWT	Latin	93.0871	95.1030	**97.4925**	95.1881	95.2078
UD_English-GUM	Latin	95.1903	96.4848	**98.1173**	95.7891	96.8330
UD_English-LinES	Latin	96.1142	99.4019	98.1483	97.5444	**99.3704**
UD_English-ParTUT	Latin	97.0771	98.0538	96.9505	96.6611	**98.0538**
UD_Estonian-EDT	Latin	95.7130	99.5625	98.4807	**99.5807**	99.4525
UD_Estonian-EWT	Latin	95.8458	98.2525	97.4714	97.9876	**98.0447**
UD_Faroese-FarPaHC	Latin	98.0595	99.3636	**99.5014**	99.3636	99.3636
UD_Finnish-FTB	Latin	97.9686	99.6406	99.0673	99.6153	**99.6406**
UD_Finnish-TDT	Latin	95.2394	99.0678	97.4792	98.8636	**98.8732**
UD_French-GSD	Latin	90.6158	93.4095	93.0457	**93.5022**	93.3905
UD_French-ParTUT	Latin	91.9381	92.3855	92.4115	**92.4564**	92.1386
UD_French-Rhapsodie	Latin	90.0299	90.9435	91.2069	**92.0552**	90.9245
UD_French-Sequoia	Latin	88.7521	91.1366	91.2310	**91.4148**	91.1281
UD_Galician-CTG	Latin	97.0100	99.5031	99.4160	**99.4789**	**99.4789**
UD_German-GSD	Latin	98.3128	98.9768	96.4192	96.6883	**98.9524**
UD_German-HDT	Latin	90.8090	99.7248	98.2471	99.7165	**99.7278**
UD_Gothic-PROIEL	Latin	99.8617	100.0000	99.9802	**100.0000**	**100.0000**
UD_Greek-GDT	Greek	96.9599	99.5714	98.8024	99.1135	**99.1267**
UD_Hebrew-HTB	Hebrew	97.0212	97.2312	**97.2312**	**97.2312**	**97.2312**
UD_Hebrew-IAHLTwiki	Hebrew	96.8689	97.3948	**98.0288**	97.1466	97.2114
UD_Hindi-HDTB	Devanagari	99.1369	99.9233	99.5563	**100.0000**	99.7826
UD_Hungarian-Szeged	Latin	95.4270	99.9037	98.1967	99.8905	**99.9037**
UD_Icelandic-IcePaHC	Latin	98.3359	99.5196	**99.5856**	99.5002	99.5175
UD_Icelandic-Modern	Latin	97.6022	98.7501	97.9920	**98.8262**	98.7147
UD_Indonesian-GSD	Latin	96.7340	98.7599	**99.3329**	98.6475	98.6380
UD_Irish-IDT	Latin	95.9235	97.3049	98.0490	**98.3690**	97.3046
UD_Italian-ISDT	Latin	94.7139	96.0480	95.8653	**96.0800**	95.9880
UD_Italian-MarkIT	Latin	95.4557	95.8674	95.6352	95.8084	**95.8771**
UD_Italian-ParTUT	Latin	95.6182	96.0450	**96.1755**	96.1634	96.0421
UD_Italian-PoSTWITA	Latin	80.0968	79.9498	**95.8151**	92.2980	79.7246
UD_Italian-TWITTIRO	Latin	82.1405	79.4268	**96.3640**	90.0124	78.4536
UD_Italian-VIT	Latin	93.7252	95.9037	94.8151	**95.9948**	95.9015
UD_Japanese-GSD	Hiragana	18.1166	2.5073	**18.3384**	2.0790	1.7688
UD_Japanese-GSDLUW	Hiragana	21.0710	3.0602	**21.4716**	2.4402	1.9908
UD_Korean-GSD	Hangul	97.9050	98.0232	**98.4283**	97.6691	97.5360
UD_Korean-Kaist	Hangul	99.7668	99.8100	**99.8556**	99.8120	99.7981
UD_Latin-ITTB	Latin	99.1079	99.9398	99.5889	99.9398	**99.9548**
UD_Latin-LLCT	Latin	99.8161	99.7358	99.7049	**99.7358**	**99.7358**
UD_Latin-PROIEL	Latin	99.8960	100.0000	99.9247	**100.0000**	**100.0000**
UD_Latin-UDante	Latin	99.0226	99.8571	**100.0000**	98.8266	97.9727
UD_Latvian-LVTB	Latin	97.5876	99.1222	98.6913	**98.8841**	98.2688
UD_Lithuanian-ALKSNIS	Latin	97.7901	97.8846	**99.5209**	96.8244	94.7655
UD_Lithuanian-HSE	Latin	98.6188	99.4490	**99.4490**	99.3078	98.4729
UD_Maltese-MUDT	Latin	74.4567	71.4375	71.3684	**71.8197**	71.4942
UD_Marathi-UFAL	Devanagari	94.6565	97.9849	**99.4987**	97.9849	97.2222
UD_Naija-NSC	Latin	97.1491	96.4959	82.3922	84.3932	**96.4959**
UD_Norwegian-Bokmaal	Latin	97.5697	99.8157	**99.3156**	99.2367	98.6826
UD_Norwegian-Nynorsk	Latin	97.8071	99.9264	99.1574	**99.4501**	99.0638
UD_Norwegian-NynorskLIA	Latin	98.5421	98.1080	96.8166	**99.9705**	98.1080
UD_Old_Church_Slavonic-PROIEL	Cyrillic	99.9802	100.0000	**100.0000**	**100.0000**	**100.0000**
UD_Old_East_Slavic-Birchbark	Cyrillic	58.4150	58.1712	56.3522	**64.5611**	58.0344
UD_Old_East_Slavic-TOROT	Cyrillic	99.7091	99.8766	99.5670	**99.8924**	99.8766
UD_Old_French-SRCMF	Latin	94.4569	94.5155	94.3983	**94.5870**	93.7363
UD_Persian-PerDT	Arabic	99.6304	95.7376	**99.8143**	99.5817	95.3785
UD_Persian-Seraji	Arabic	99.9460	94.9495	**100.0000**	**100.0000**	94.9495
UD_Polish-LFG	Latin	96.6140	96.8738	96.8324	**96.8350**	96.7463
UD_Polish-PDB	Latin	98.6391	98.5925	**99.3056**	98.6292	98.4966
UD_Pomak-Philotis	Latin	98.9622	99.5594	**99.7999**	99.1807	98.5531
UD_Portuguese-Bosque	Latin	95.4824	99.7518	**99.1326**	98.1305	96.6265
UD_Portuguese-GSD	Latin	97.6390	99.8707	99.3028	99.8438	**99.8606**
UD_Romanian-Nonstandard	Latin	93.9927	94.0963	94.0563	**94.1047**	94.0963
UD_Romanian-RRT	Latin	95.4008	97.4519	96.7511	**97.4080**	97.1179
UD_Romanian-SiMoNERo	Latin	94.9535	97.6284	**97.7622**	97.6856	97.6284
UD_Russian-GSD	Cyrillic	92.3269	93.9545	0.0000	93.5442	**93.9545**
UD_Russian-SynTagRus	Cyrillic	97.2647	99.1475	98.9415	**99.3397**	99.1491
UD_Russian-Taiga	Cyrillic	90.4316	90.9374	94.3666	**95.9738**	90.4210
UD_Scottish_Gaelic-ARCOSG	Latin	81.9492	90.5358	88.3397	87.9921	**94.7130**
UD_Serbian-SET	Latin	96.5872	99.8999	98.5482	**99.9000**	99.8082
UD_Slovak-SNK	Latin	99.2164	98.3893	**99.9372**	98.2144	97.9275
UD_Slovenian-SSJ	Latin	98.2695	99.4478	98.9801	**99.1378**	99.0929
UD_Spanish-AnCora	Latin	97.2414	99.7038	99.6316	99.6753	**99.7173**
UD_Spanish-GSD	Latin	97.9134	99.7270	99.6384	**99.7106**	99.6486
UD_Swedish-LinES	Latin	98.4584	99.6189	**99.8596**	99.6270	99.6189
UD_Swedish-Talbanken	Latin	98.4586	99.3863	99.3485	**99.9030**	99.3709
UD_Swedish_Sign_Language-SSLC	Latin	25.7426	39.9276	30.2210	**67.4144**	40.6378
UD_Tamil-TTB	Tamil	95.9272	100.0000	96.0589	**100.0000**	**100.0000**
UD_Telugu-MTG	Telugu	99.5475	99.7736	99.5475	**99.7736**	**99.7736**
UD_Turkish-Atis	Latin	64.3649	91.3600	96.6977	64.8804	**99.9383**
UD_Turkish-BOUN	Latin	94.8207	97.7312	98.1773	94.6122	**98.1929**
UD_Turkish-FrameNet	Latin	99.4386	99.8594	**100.0000**	99.4386	**100.0000**
UD_Turkish-IMST	Latin	96.3198	99.1505	**99.5750**	96.3871	99.4002
UD_Turkish-Kenet	Latin	98.5802	99.7411	99.9715	98.6084	**99.9915**
UD_Turkish-Penn	Latin	89.0149	98.1274	95.9742	93.4662	**98.5775**
UD_Turkish-Tourism	Latin	99.7504	100.0000	99.8775	99.8237	**100.0000**
UD_Turkish_German-SAGT	Latin	97.7253	98.9693	99.1814	97.9926	**99.2797**
UD_Ukrainian-IU	Cyrillic	96.2343	97.0106	**97.3685**	94.9853	94.7347
UD_Urdu-UDTB	Arabic	96.9010	93.4296	**99.7978**	94.0515	93.4296
UD_Uyghur-UDT	Arabic	99.3277	88.1386	**99.6426**	99.0910	87.2816
UD_Vietnamese-VTB	Latin	73.1135	74.3217	74.3138	74.3038	**74.3217**
UD_Welsh-CCG	Latin	91.8942	92.7169	92.4593	**92.8141**	92.4953
UD_Western_Armenian-ArmTDP	Armenian	95.6263	89.8907	**96.1380**	89.6008	88.4475
UD_Wolof-WTB	Latin	96.5692	99.9097	99.5090	**99.8194**	99.7992
Average		91.1400	91.5459	**92.1092**	91.6538	91.3658

Table 10: Results (F1) of rule-based baselines for the tokenization task.

Treebank	MT	MT+SPL	MT+SPL+LA	MT+ML	MT+ML+SPL
UD_Afrikaans-AfriBooms	84.4164	**84.4244**	82.6860	83.7192	83.7192
UD_Ancient_Greek-PROIEL	73.1688	73.0728	71.2465	**76.1947**	**76.1947**
UD_Ancient_Greek-Perseus	61.4745	62.5805	60.6841	**65.8641**	**65.8641**
UD_Ancient_Hebrew-PTNK	36.7661	36.7116	37.5613	**37.9785**	37.9512
UD_Arabic-PADT	**82.6753**	82.4940	81.1069	82.0498	82.0498
UD_Armenian-ArmTDP	81.7391	81.5556	79.3980	**84.6786**	**84.6786**
UD_Armenian-BSUT	80.2451	80.2102	75.3990	**84.9822**	84.8858
UD_Basque-BDT	82.5372	**82.7118**	80.8201	81.2990	81.2990
UD_Belarusian-HSE	87.9314	87.9337	86.9694	**89.2944**	88.7283
UD_Bulgarian-BTB	**90.9249**	90.6723	89.9257	90.7034	90.7034
UD_Catalan-AnCora	**92.7893**	92.6428	92.3214	92.2201	92.2201
UD_Chinese-GSD	82.0897	**82.4138**	80.6919	78.7714	78.7714
UD_Chinese-GSDSimp	81.6792	**82.1853**	80.3492	79.0257	78.4564
UD_Classical_Chinese-Kyoto	**77.1275**	**77.1275**	76.8315	76.1740	76.3416
UD_Coptic-Scriptorium	14.9260	15.0407	**15.3117**	14.4420	14.4420
UD_Croatian-SET	88.8939	**89.0698**	87.6522	88.8914	88.8914
UD_Czech-CAC	92.0138	92.3352	91.7107	92.2618	**92.4822**
UD_Czech-CLTT	85.3839	85.9048	82.3260	**89.1779**	88.6879
UD_Czech-FicTree	92.5322	92.6375	91.5025	**93.8481**	93.7457
UD_Czech-PDT	**93.3442**	93.3314	93.0962	93.2325	93.1114
UD_Danish-DDT	**87.0323**	86.6770	84.6962	85.2165	85.2165
UD_Dutch-Alpino	91.8020	**92.0111**	90.7299	91.1166	91.1166
UD_Dutch-LassySmall	87.5554	87.5971	85.5539	**89.2134**	**89.2134**
UD_English-Atis	91.3606	91.4208	90.7285	**91.9395**	91.8109
UD_English-EWT	89.5767	**89.6773**	88.7656	86.8256	86.1819
UD_English-GUM	90.5405	**90.5974**	89.2256	88.7021	87.7360
UD_English-LinES	86.7729	**87.2816**	85.3969	84.0065	83.9948
UD_English-ParTUT	88.9665	**89.7502**	88.0559	84.9548	85.5877
UD_Estonian-EDT	**87.3855**	87.2088	86.4096	86.9014	86.9014
UD_Estonian-EWT	78.2609	77.8579	75.3057	**82.1119**	81.8031
UD_Faroese-FarPaHC	79.0317	79.3336	76.5884	85.0157	**85.1008**
UD_Finnish-FTB	88.2807	**88.6049**	87.1515	81.1546	81.1546
UD_Finnish-TDT	**87.9186**	87.8403	86.6344	81.4116	80.6745
UD_French-GSD	**94.7045**	94.6224	94.2538	94.0336	93.3099
UD_French-ParTUT	88.5354	**88.5597**	85.9808	88.0351	87.9254
UD_French-Rhapsodie	81.2867	81.1645	78.6425	82.0865	**82.9911**
UD_French-Sequoia	92.3741	**92.5181**	90.4434	89.9285	89.9285
UD_Galician-CTG	**81.7786**	81.6850	80.5697	80.1807	79.4993
UD_German-GSD	**87.2859**	87.1676	86.8196	85.2013	84.8394
UD_German-HDT	**96.4980**	96.4205	96.3463	96.0492	96.0361
UD_Gothic-PROIEL	75.2743	74.9048	71.2704	**80.0811**	**80.0811**
UD_Greek-GDT	90.2670	90.5536	87.5259	**91.0068**	**91.0068**
UD_Hebrew-HTB	**85.6904**	85.6613	83.8548	85.0323	85.0323
UD_Hebrew-IAHLTwiki	87.3303	**87.4521**	85.3387	86.8001	87.0087
UD_Hindi-HDTB	92.2096	**92.2168**	91.5493	91.9230	91.9230
UD_Hungarian-Szeged	84.1317	84.2626	79.4624	**84.5123**	**84.5123**
UD_Icelandic-IcePaHC	**82.2869**	82.1996	81.6687	82.2118	82.0604
UD_Icelandic-Modern	94.4324	**94.5304**	94.1826	91.0776	90.7820
UD_Indonesian-GSD	79.3448	**79.5219**	77.9777	78.5861	78.5861
UD_Irish-IDT	81.3163	**81.5941**	79.5059	81.0619	81.0619
UD_Italian-ISDT	92.2448	**92.2538**	91.8283	91.2661	91.2661
UD_Italian-MarkIT	82.3153	82.2788	79.3551	**84.7847**	84.6991
UD_Italian-ParTUT	90.4001	90.6317	88.7852	**90.7198**	90.5566
UD_Italian-PoSTWITA	79.4079	79.7463	77.9168	**79.8849**	79.2858
UD_Italian-TWITTIRO	77.6025	76.8395	73.2015	**83.0942**	82.6186
UD_Italian-VIT	**87.8005**	87.7088	87.0623	86.3861	85.6873
UD_Japanese-GSD	**91.5100**	90.5195	90.6073	45.0598	46.3854
UD_Japanese-GSDLUW	90.7221	**90.8231**	90.5641	85.0528	82.6332
UD_Korean-GSD	**82.5916**	82.2265	80.3898	70.7678	72.1850
UD_Korean-Kaist	88.0674	**88.0907**	87.5109	84.4445	84.4445
UD_Latin-ITTB	89.5811	89.4725	89.1896	**89.8602**	**89.8602**
UD_Latin-LLCT	95.6595	**95.7340**	95.2649	95.3166	95.0806
UD_Latin-PROIEL	82.2107	81.7403	80.2310	**82.5466**	**82.5466**
UD_Latin-UDante	62.2266	62.2266	58.0768	**70.6718**	70.5123
UD_Latvian-LVTB	87.0840	87.0100	86.2245	**87.2254**	86.7094
UD_Lithuanian-ALKSNIS	**83.0032**	82.8410	79.7578	82.1998	81.7313
UD_Lithuanian-HSE	62.1609	59.9172	53.4253	**69.1176**	**69.1176**
UD_Maltese-MUDT	**78.6599**	78.1391	74.7526	78.3380	78.4850
UD_Marathi-UFAL	59.5000	59.5000	54.7500	**62.5000**	**62.5000**
UD_Naija-NSC	**91.5737**	91.3615	90.9284	90.8685	91.2336
UD_Norwegian-Bokmaal	**93.1311**	92.8160	92.3563	93.1269	92.9835
UD_Norwegian-Nynorsk	91.6224	**91.6951**	91.3670	91.3370	91.3687
UD_Norwegian-NynorskLIA	74.7995	74.4541	73.2012	**76.7588**	**76.7588**
UD_Old_Church_Slavonic-PROIEL	63.9968	63.4163	61.3348	**66.8779**	**66.8779**
UD_Old_East_Slavic-Birchbark	30.7814	30.3695	27.4288	38.0637	**38.7365**
UD_Old_East_Slavic-TOROT	66.1137	64.9739	63.5979	**67.6336**	65.9382
UD_Old_French-SRCMF	**88.4299**	**88.4299**	87.4860	87.2330	87.2330
UD_Persian-PerDT	90.4797	**90.5040**	89.9725	89.1375	88.3543
UD_Persian-Seraji	**88.2450**	87.8753	86.9731	83.6169	83.6169
UD_Polish-LFG	93.8070	**94.7196**	93.8567	89.2782	90.6378
UD_Polish-PDB	**92.2020**	92.0438	91.6946	91.1990	91.1717
UD_Pomak-Philotis	80.6420	80.4135	79.1341	**80.6535**	80.4386
UD_Portuguese-Bosque	**89.5332**	89.3787	88.5545	85.5418	85.0767
UD_Portuguese-GSD	93.0233	**93.0251**	92.3245	90.5872	90.5872
UD_Romanian-Nonstandard	86.5708	86.5415	86.1653	**87.0036**	86.6810
UD_Romanian-RRT	88.5778	88.3649	87.8207	**88.7053**	**88.7053**
UD_Romanian-SiMoNERo	89.7483	89.9343	89.2690	**90.1126**	89.8649
UD_Russian-GSD	**88.4789**	88.2607	86.4246	86.6846	86.6846
UD_Russian-SynTagRus	**91.2445**	91.2358	90.9764	90.6270	90.6721
UD_Russian-Taiga	73.2837	73.5265	71.5174	73.0162	**73.8604**
UD_Scottish_Gaelic-ARCOSG	78.6648	78.8475	77.3221	**79.7084**	79.1626
UD_Serbian-SET	90.2639	**90.3307**	89.0400	89.9024	89.9024
UD_Slovak-SNK	92.0679	92.5028	89.9831	**93.2427**	**93.2427**
UD_Slovenian-SSJ	**91.8027**	91.6349	90.9197	91.5286	91.5286
UD_Spanish-AnCora	**91.8813**	91.8631	91.3336	89.7146	89.7146
UD_Spanish-GSD	89.4629	**89.7809**	89.3542	87.5403	87.8090
UD_Swedish-LinES	**85.8554**	85.7961	84.3765	85.5391	85.5391
UD_Swedish-Talbanken	86.4214	86.6167	84.8630	**86.8464**	**86.8464**
UD_Swedish_Sign_Language-SSLC	0.2494	1.0114	9.4718	**22.9152**	**22.9152**
UD_Tamil-TTB	66.1054	66.6962	59.5049	**71.9469**	**71.9469**
UD_Telugu-MTG	83.1698	83.0189	83.0189	**86.7925**	**86.7925**
UD_Turkish-Atis	**89.1447**	88.6102	88.4410	89.1405	89.1107
UD_Turkish-BOUN	70.8878	**71.2664**	69.0099	68.2795	68.5199
UD_Turkish-FrameNet	**80.6054**	80.2534	78.1140	79.6479	78.3955
UD_Turkish-IMST	66.1826	**66.2337**	62.0027	60.1934	60.5847
UD_Turkish-Kenet	74.6461	**74.7828**	72.0292	73.7631	73.0986
UD_Turkish-Penn	76.0756	76.0057	75.1927	77.0646	**77.1437**
UD_Turkish-Tourism	87.9392	87.9435	87.3805	89.2091	**89.3561**
UD_Turkish_German-SAGT	63.9620	63.3574	60.2209	**68.0413**	68.0168
UD_Ukrainian-IU	89.6039	89.4412	87.6859	**90.7637**	90.4345
UD_Urdu-UDTB	**81.7873**	81.1975	80.1619	81.6240	81.6240
UD_Uyghur-UDT	45.4158	45.2646	43.6334	**47.4692**	**47.4692**
UD_Vietnamese-VTB	**60.5940**	60.4750	57.6923	57.8233	57.8233
UD_Welsh-CCG	79.6195	79.8443	76.9308	**80.5763**	80.1491
UD_Western_Armenian-ArmTDP	81.4963	81.5792	80.0452	**83.3126**	82.7708
UD_Wolof-WTB	71.2773	71.4056	66.9276	**74.5610**	74.4331
Average	**81.5181**	81.4892	79.9496	81.2555	81.1588

Table 11: Full results on dependency parsing tagging on test sets (LAS F1). MT=Multi Task, SPL=learn additional SPLits from training data, ML=MultiLingual, LA=Layer Attention

Treebank	MT	MT+SPL	MT+SPL+LA	MT+ML	MT+ML+SPL
UD_Afrikaans-AfriBooms	**97.9968**	97.9684	97.9028	97.2897	97.2897
UD_Ancient_Greek-PROIEL	90.9830	90.9584	90.8525	**91.7134**	**91.7134**
UD_Ancient_Greek-Perseus	86.9534	87.8025	87.9626	**88.5717**	**88.5717**
UD_Ancient_Hebrew-PTNK	58.8612	58.3163	58.2834	58.8476	**60.1417**
UD_Arabic-PADT	**96.1672**	96.1147	95.9512	95.7742	95.7742
UD_Armenian-ArmTDP	96.6807	96.8496	96.7284	**96.9731**	**96.9731**
UD_Armenian-BSUT	95.7494	95.7579	95.7376	**96.4546**	96.3474
UD_Basque-BDT	**96.3481**	96.3166	96.1341	95.6796	95.6796
UD_Belarusian-HSE	**97.7232**	97.7111	97.6199	97.6730	97.3950
UD_Bulgarian-BTB	**99.0801**	99.0644	98.9773	99.0396	99.0396
UD_Catalan-AnCora	99.0366	99.0197	**99.0659**	99.0045	99.0045
UD_Chinese-GSD	94.6770	**94.7119**	94.6566	93.0870	93.0870
UD_Chinese-GSDSimp	94.5381	**94.6355**	94.6005	93.1665	93.2528
UD_Classical_Chinese-Kyoto	**90.7600**	**90.7600**	90.7461	89.9417	90.1464
UD_Coptic-Scriptorium	44.4875	44.5219	45.0832	**45.2618**	**45.2618**
UD_Croatian-SET	98.2551	98.2213	98.1675	**98.3131**	**98.3131**
UD_Czech-CAC	99.4443	**99.4811**	**99.4811**	99.2606	99.3525
UD_Czech-CLTT	99.0937	99.0937	**99.2497**	99.0219	98.9395
UD_Czech-FicTree	99.0181	98.9731	**99.0452**	98.6519	98.6939
UD_Czech-PDT	99.3712	**99.3803**	99.3703	99.2035	99.1972
UD_Danish-DDT	97.8653	**97.9280**	97.7875	97.6530	97.6530
UD_Dutch-Alpino	**97.7594**	97.7162	97.7031	97.3658	97.3658
UD_Dutch-LassySmall	97.0829	97.1439	97.1581	**97.1586**	**97.1586**
UD_English-Atis	**98.5250**	98.3444	**98.5250**	98.3668	98.2990
UD_English-EWT	96.6022	96.5752	**96.6493**	95.7269	94.8605
UD_English-GUM	97.9726	**97.9933**	97.8410	96.2789	95.7880
UD_English-LinES	97.2023	**97.6847**	97.5957	94.9502	95.0417
UD_English-ParTUT	95.4027	95.8854	**95.9941**	92.9666	92.9323
UD_Estonian-EDT	**97.1493**	97.0924	96.9640	96.8706	96.8706
UD_Estonian-EWT	92.3901	92.1021	92.3331	**93.2726**	92.9549
UD_Faroese-FarPaHC	95.5019	95.6148	95.8329	**97.4864**	97.3202
UD_Finnish-FTB	96.0872	96.1060	**96.1802**	93.8605	93.8605
UD_Finnish-TDT	**97.2578**	97.2007	97.1869	95.0467	94.7717
UD_French-GSD	**98.4571**	98.4528	98.4224	98.2161	98.1337
UD_French-ParTUT	95.7762	**96.0219**	95.9122	95.3348	95.3897
UD_French-Rhapsodie	97.5159	97.4335	97.5625	97.4174	**97.6720**
UD_French-Sequoia	98.4008	**98.4316**	98.3952	98.1936	98.1936
UD_Galician-CTG	96.9424	96.9132	**97.0147**	96.3346	96.2413
UD_German-GSD	96.2085	96.1312	**96.2777**	94.6057	94.1483
UD_German-HDT	**98.2508**	98.2150	98.2254	98.0856	98.0828
UD_Gothic-PROIEL	95.2150	95.0521	94.7998	**95.8620**	**95.8620**
UD_Greek-GDT	97.2417	**97.4882**	97.0736	97.0285	97.0285
UD_Hebrew-HTB	96.4704	**96.5006**	96.4527	95.7645	95.7645
UD_Hebrew-IAHLTwiki	95.1170	**95.1672**	94.8433	94.1550	93.8584
UD_Hindi-HDTB	**97.6389**	97.5438	97.5664	97.2045	97.2045
UD_Hungarian-Szeged	**97.0751**	96.9647	96.9397	96.9445	96.9445
UD_Icelandic-IcePaHC	96.9381	**96.9390**	96.8508	96.8977	96.9240
UD_Icelandic-Modern	98.8479	98.9213	**98.9242**	98.7396	98.7809
UD_Indonesian-GSD	**94.0192**	93.9145	93.7970	93.4418	93.4418
UD_Irish-IDT	95.4391	**95.5148**	95.3591	95.1305	95.1305
UD_Italian-ISDT	**98.3520**	98.2891	98.2310	97.9783	97.9783
UD_Italian-MarkIT	95.7516	95.7463	96.3735	**96.7719**	96.6755
UD_Italian-ParTUT	**97.4699**	97.3439	97.0393	96.6433	96.6966
UD_Italian-PoSTWITA	95.4705	**95.5518**	95.3754	95.4524	95.1261
UD_Italian-TWITTIRO	94.0414	93.9033	94.2062	**96.4648**	96.4467
UD_Italian-VIT	**97.9273**	97.8867	97.8575	97.3349	97.4323
UD_Japanese-GSD	**96.6300**	96.0440	96.2356	68.5098	68.7758
UD_Japanese-GSDLUW	96.1377	96.1001	**96.1678**	92.4487	90.8582
UD_Korean-GSD	**95.5412**	95.2074	95.1194	89.3777	89.8862
UD_Korean-Kaist	**96.4180**	96.3309	96.3328	94.3217	94.3217
UD_Latin-ITTB	98.6382	98.5864	**98.6516**	98.5949	98.5949
UD_Latin-LLCT	99.6197	**99.6238**	99.6135	99.5536	99.5659
UD_Latin-PROIEL	**97.3818**	97.2562	96.9227	97.2382	97.2382
UD_Latin-UDante	92.5735	92.5735	92.2371	**94.0096**	93.8807
UD_Latvian-LVTB	97.5850	**97.6713**	97.5173	97.2567	97.2753
UD_Lithuanian-ALKSNIS	**97.1369**	97.1063	96.9204	96.6579	96.7614
UD_Lithuanian-HSE	84.4138	83.6631	82.7586	**87.0404**	**87.0404**
UD_Maltese-MUDT	93.5201	93.4672	**93.5730**	93.1648	92.8806
UD_Marathi-UFAL	84.2500	84.2500	83.0000	**89.2500**	**89.2500**
UD_Naija-NSC	98.4314	98.3629	**98.5001**	98.2355	98.2701
UD_Norwegian-Bokmaal	**98.7681**	98.7089	98.7116	98.5990	98.5302
UD_Norwegian-Nynorsk	98.1504	**98.2385**	98.2273	97.9762	97.8738
UD_Norwegian-NynorskLIA	95.8532	95.8883	96.0606	**96.3397**	**96.3397**
UD_Old_Church_Slavonic-PROIEL	83.4018	82.6912	82.4878	**83.7681**	**83.7681**
UD_Old_East_Slavic-Birchbark	56.3633	56.5995	56.0087	**61.9864**	61.5673
UD_Old_East_Slavic-TOROT	85.0214	84.1244	83.8430	**85.0882**	84.2245
UD_Old_French-SRCMF	**97.1391**	**97.1391**	96.9250	96.5163	96.5163
UD_Persian-PerDT	**97.5053**	97.3881	97.4103	96.7279	96.3201
UD_Persian-Seraji	97.6515	**97.6893**	97.5749	94.7950	94.7950
UD_Polish-LFG	98.2562	**98.6980**	98.5988	97.2421	97.4863
UD_Polish-PDB	**98.8122**	98.7422	98.7206	98.3878	98.4759
UD_Pomak-Philotis	97.1726	97.1497	**97.2183**	96.9212	97.0015
UD_Portuguese-Bosque	**97.5648**	97.5513	97.4698	96.1570	95.9555
UD_Portuguese-GSD	**98.4166**	98.3948	98.3326	97.4665	97.4665
UD_Romanian-Nonstandard	96.3801	96.4421	96.3258	**96.4917**	96.1476
UD_Romanian-RRT	**98.1022**	98.0139	98.0492	97.6942	97.6942
UD_Romanian-SiMoNERo	97.8457	**97.9130**	97.8894	97.7857	97.6227
UD_Russian-GSD	98.0955	**98.1509**	98.1034	97.0738	97.0738
UD_Russian-SynTagRus	**98.4452**	**98.4452**	98.4373	98.0138	98.0781
UD_Russian-Taiga	92.2305	**92.4995**	92.4230	91.2850	91.7379
UD_Scottish_Gaelic-ARCOSG	94.5622	**94.6581**	94.3232	94.4232	94.5021
UD_Serbian-SET	**98.4281**	98.3780	98.3652	98.3240	98.3240
UD_Slovak-SNK	97.4868	97.3962	**97.5851**	97.3128	97.3128
UD_Slovenian-SSJ	**98.9152**	98.8831	98.8661	98.6831	98.6831
UD_Spanish-AnCora	98.9691	98.9777	**98.9787**	98.2663	98.2663
UD_Spanish-GSD	96.8846	96.9447	**96.9597**	96.1623	96.1978
UD_Swedish-LinES	97.2056	97.2110	**97.2350**	96.9134	96.9134
UD_Swedish-Talbanken	**97.9844**	97.8825	97.8828	97.7941	97.7941
UD_Swedish_Sign_Language-SSLC	4.9875	1.7699	27.6867	**59.4634**	**59.4634**
UD_Tamil-TTB	85.1573	86.5989	85.4111	**87.6991**	**87.6991**
UD_Telugu-MTG	93.2830	93.1321	93.1321	**93.5849**	**93.5849**
UD_Turkish-Atis	97.0600	**97.1217**	97.1205	97.1076	97.0976
UD_Turkish-BOUN	90.3377	**90.5000**	90.4909	86.8154	86.4930
UD_Turkish-FrameNet	93.4882	93.2066	92.9627	94.2958	**94.6517**
UD_Turkish-IMST	93.9811	93.9402	**94.2007**	89.9780	90.1039
UD_Turkish-Kenet	91.9133	**91.9987**	91.9506	90.7853	90.8449
UD_Turkish-Penn	94.5844	94.4080	**94.7331**	93.7669	93.9649
UD_Turkish-Tourism	97.6231	97.6181	97.5251	**97.6968**	97.6576
UD_Turkish_German-SAGT	89.4928	89.1651	90.2386	**91.2434**	91.2394
UD_Ukrainian-IU	97.8524	**97.9042**	97.8522	97.8282	97.6365
UD_Urdu-UDTB	94.1600	**94.2423**	94.0228	94.2153	94.2153
UD_Uyghur-UDT	74.0102	73.8618	73.4734	**75.0332**	**75.0332**
UD_Vietnamese-VTB	86.5231	**86.7846**	86.5991	84.8133	84.8133
UD_Welsh-CCG	**95.2164**	95.0945	95.1035	94.5934	94.4364
UD_Western_Armenian-ArmTDP	96.4214	96.4290	96.3650	96.4362	**96.5270**
UD_Wolof-WTB	92.3363	92.2992	91.5788	92.6944	**92.7353**
Average	93.7492	93.7111	**93.8782**	93.6883	93.6524

Table 12: Full results on UPOS tagging on dev sets (F1 ST=Single Task (tokenization only), MT=Multi Task, SPL=learn additional SPLits from training data, ML=MultiLingual, LA=Layer Attention

Treebank	MT	MT+SPL	MT+SPL+LA	MT+ML	MT+ML+SPL
UD_Afrikaans-AfriBooms	97.4513	97.2724	**97.4701**	96.2168	96.2168
UD_Ancient_Greek-PROIEL	82.4421	82.7114	82.8548	**83.4523**	**83.4523**
UD_Ancient_Greek-Perseus	82.1874	82.4029	82.7934	**84.0997**	**84.0997**
UD_Ancient_Hebrew-PTNK	49.3529	49.4211	49.4414	**49.8706**	49.8570
UD_Arabic-PADT	91.9457	91.6348	**91.9678**	90.3206	90.3206
UD_Armenian-ArmTDP	88.5086	88.3612	**89.0073**	87.6495	87.6495
UD_Armenian-BSUT	83.3674	83.5085	**85.7143**	84.4638	84.7294
UD_Basque-BDT	90.6212	90.6970	**91.0253**	88.4167	88.4167
UD_Belarusian-HSE	**94.4886**	94.4708	94.3105	94.4383	94.2903
UD_Bulgarian-BTB	97.2588	97.1933	**97.2615**	95.9189	95.9189
UD_Catalan-AnCora	98.6921	98.6646	**98.7232**	98.5978	98.5978
UD_Chinese-GSD	97.4470	97.4132	**97.4685**	96.4665	96.4665
UD_Chinese-GSDSimp	97.3342	97.3592	**97.3812**	96.4273	96.5515
UD_Classical_Chinese-Kyoto	91.8741	91.8741	**91.9679**	91.5030	91.2500
UD_Coptic-Scriptorium	46.7912	46.9996	46.7867	**47.1590**	**47.1590**
UD_Croatian-SET	95.3934	**95.4983**	95.3054	94.8270	94.8270
UD_Czech-CAC	96.3766	**96.4776**	96.4409	96.0735	96.1653
UD_Czech-CLTT	88.2592	88.0092	88.8287	92.9448	**93.4914**
UD_Czech-FicTree	95.5289	95.4480	**95.7543**	92.7853	92.7368
UD_Czech-PDT	97.6474	97.6239	**97.6786**	96.6848	96.6227
UD_Danish-DDT	96.9747	**97.0275**	97.0128	95.7561	95.7561
UD_Dutch-Alpino	96.9344	**96.9955**	96.7479	96.7148	96.7148
UD_Dutch-LassySmall	96.9338	**96.9771**	96.8336	96.6675	96.6675
UD_English-Atis	98.5099	**98.5551**	98.4046	98.4421	98.4194
UD_English-EWT	**96.7435**	96.6439	96.6008	93.6489	93.0063
UD_English-GUM	97.9260	97.9674	**98.1357**	93.1447	91.0983
UD_English-LinES	96.3836	96.7722	**96.8760**	90.6613	90.7969
UD_English-ParTUT	93.3064	**93.8281**	93.6053	82.7764	82.6352
UD_Estonian-EDT	**95.3689**	95.2653	95.2020	94.3738	94.3738
UD_Estonian-EWT	89.2280	89.2693	89.4969	**91.8399**	91.4707
UD_Faroese-FarPaHC	90.6490	90.7144	91.1162	91.5774	**91.7659**
UD_Finnish-FTB	95.3989	95.4687	**95.6641**	91.1205	91.1205
UD_Finnish-TDT	**95.5354**	95.4784	95.4810	91.2033	90.7541
UD_French-GSD	98.4109	**98.4269**	98.4108	97.8496	96.2457
UD_French-ParTUT	87.9320	88.3951	**90.3704**	86.5532	86.6630
UD_French-Rhapsodie	93.7309	93.7738	94.9056	95.4109	**95.9006**
UD_French-Sequoia	96.3439	96.5702	**97.2842**	92.1105	92.1105
UD_Galician-CTG	**99.5574**	99.5167	99.5518	39.1018	38.8054
UD_German-GSD	91.1180	**91.1785**	91.0168	74.8850	73.9097
UD_German-HDT	**87.5933**	87.5805	87.5212	86.5833	86.7260
UD_Gothic-PROIEL	82.5111	82.0918	83.0648	**85.6380**	**85.6380**
UD_Greek-GDT	92.7593	92.6517	92.8072	**92.9385**	**92.9385**
UD_Hebrew-HTB	93.3597	93.3421	**93.6292**	91.0625	91.0625
UD_Hebrew-IAHLTwiki	89.6128	**89.6771**	89.3543	86.7712	86.9220
UD_Hindi-HDTB	94.0383	**94.1023**	94.0993	93.3201	93.3201
UD_Hungarian-Szeged	87.8798	88.7916	**90.8279**	88.6797	88.6797
UD_Icelandic-IcePaHC	92.2687	**92.3210**	92.2317	91.6683	91.4378
UD_Icelandic-Modern	98.0057	98.0150	**98.2694**	96.5755	96.5473
UD_Indonesian-GSD	94.8644	94.8402	**94.8919**	94.1342	94.1342
UD_Irish-IDT	88.3677	88.4644	**88.6377**	86.3314	86.3314
UD_Italian-ISDT	**98.2352**	98.1903	98.0783	97.3583	97.3583
UD_Italian-MarkIT	90.1006	90.0849	**92.9759**	89.5633	88.1456
UD_Italian-ParTUT	96.8240	96.5901	97.1470	**97.2536**	97.0557
UD_Italian-PoSTWITA	95.5128	95.4334	**95.7136**	95.2917	94.7961
UD_Italian-TWITTIRO	89.4848	89.2081	91.8257	**95.4148**	95.2214
UD_Italian-VIT	97.7772	97.7365	**97.8460**	95.7340	95.6154
UD_Japanese-GSD	**97.6557**	97.2092	97.4020	46.9634	46.7208
UD_Japanese-GSDLUW	97.2502	97.2354	**97.2717**	59.9026	57.0674
UD_Korean-GSD	**99.0882**	98.6869	98.6659	46.9388	43.6423
UD_Korean-Kaist	**99.9466**	99.9031	99.9327	44.1289	44.1289
UD_Latin-ITTB	96.0921	96.0369	**96.1122**	94.3262	94.3262
UD_Latin-LLCT	97.2345	**97.2510**	97.2366	96.1227	96.1389
UD_Latin-PROIEL	**91.0336**	90.9150	90.8830	90.8393	90.8393
UD_Latin-UDante	66.6003	66.6003	68.7275	**70.2993**	70.2785
UD_Latvian-LVTB	94.2144	**94.2629**	94.2155	92.7997	92.9686
UD_Lithuanian-ALKSNIS	88.8331	88.8706	**89.6886**	84.5519	84.2478
UD_Lithuanian-HSE	54.6207	54.0267	57.5632	**62.5000**	**62.5000**
UD_Maltese-MUDT	**99.8384**	99.8041	99.7649	53.9468	52.7610
UD_Marathi-UFAL	52.5000	52.5000	**58.2500**	51.7500	51.7500
UD_Naija-NSC	98.8502	98.7885	**98.9326**	98.8397	98.7918
UD_Norwegian-Bokmaal	97.5610	97.5842	**97.6364**	97.1443	97.0699
UD_Norwegian-Nynorsk	97.5904	97.6498	**97.6673**	97.1091	97.1250
UD_Norwegian-NynorskLIA	93.9741	94.1373	94.2212	**95.3459**	**95.3459**
UD_Old_Church_Slavonic-PROIEL	70.0460	69.7293	68.9522	**73.0855**	**73.0855**
UD_Old_East_Slavic-Birchbark	46.5188	46.5422	47.0775	**50.8920**	50.0051
UD_Old_East_Slavic-TOROT	76.4603	75.5977	75.6350	**76.9055**	75.8930
UD_Old_French-SRCMF	**98.0149**	**98.0149**	97.8446	97.4894	97.4894
UD_Persian-PerDT	**97.2265**	97.1053	97.1315	95.6372	95.1042
UD_Persian-Seraji	97.1501	**97.2386**	97.2131	92.3004	92.3004
UD_Polish-LFG	94.0283	94.4905	**94.5974**	84.0378	82.5496
UD_Polish-PDB	94.8246	94.8353	**95.1568**	91.0859	91.6739
UD_Pomak-Philotis	89.7927	89.8384	**90.2610**	88.6845	88.2974
UD_Portuguese-Bosque	**96.5376**	96.5013	96.4653	95.6953	95.5883
UD_Portuguese-GSD	**96.5662**	96.5276	96.5157	42.1028	42.1028
UD_Romanian-Nonstandard	93.4012	**93.4903**	93.3412	93.1047	92.6778
UD_Romanian-RRT	97.3348	97.1996	**97.3872**	94.2721	94.2721
UD_Romanian-SiMoNERo	97.2370	**97.3040**	97.3010	96.5399	96.3845
UD_Russian-GSD	**93.7655**	93.5560	93.6010	90.9821	90.9821
UD_Russian-SynTagRus	**94.4689**	94.4458	94.1717	93.2841	93.2312
UD_Russian-Taiga	87.5310	**88.0741**	87.9341	85.6692	87.3426
UD_Scottish_Gaelic-ARCOSG	90.2452	**90.3532**	90.3103	90.0303	89.9041
UD_Serbian-SET	94.1417	94.1750	93.8694	**94.5802**	**94.5802**
UD_Slovak-SNK	91.3846	91.3875	**91.3967**	89.9191	89.9191
UD_Slovenian-SSJ	96.4324	96.3815	**96.4928**	95.0568	95.0568
UD_Spanish-AnCora	**98.5782**	98.5658	98.5400	97.7222	97.7222
UD_Spanish-GSD	96.9477	96.9968	**97.1133**	96.2282	96.1485
UD_Swedish-LinES	92.7671	92.7023	**92.7742**	91.7610	91.7610
UD_Swedish-Talbanken	**96.3821**	96.3723	96.3726	95.2002	95.2002
UD_Swedish_Sign_Language-SSLC	5.4863	3.0341	44.4444	**59.8985**	**59.8985**
UD_Tamil-TTB	79.0430	80.9376	**82.1397**	76.3717	76.3717
UD_Telugu-MTG	**98.2642**	**98.2642**	**98.2642**	33.5094	33.5094
UD_Turkish-Atis	95.5181	95.4564	**95.5780**	95.4606	95.5537
UD_Turkish-BOUN	90.1540	90.0242	**90.4408**	79.6997	79.4223
UD_Turkish-FrameNet	88.2084	88.2788	88.8811	**90.6338**	90.1478
UD_Turkish-IMST	87.2104	87.1491	**87.8388**	69.2485	69.2060
UD_Turkish-Kenet	89.8339	**89.8567**	89.7402	86.9285	86.6746
UD_Turkish-Penn	93.1145	93.1812	**93.1916**	91.8842	92.0816
UD_Turkish-Tourism	**96.5058**	96.4909	96.3685	96.4324	96.4814
UD_Turkish_German-SAGT	72.5006	72.3743	76.8940	78.4938	**78.9878**
UD_Ukrainian-IU	92.3719	**92.4160**	92.2600	91.2967	91.1109
UD_Urdu-UDTB	82.8710	**83.1247**	82.9670	82.8721	82.8721
UD_Uyghur-UDT	67.8916	67.7974	68.1341	**69.5545**	**69.5545**
UD_Vietnamese-VTB	90.1962	**90.3577**	90.1940	70.0560	70.0560
UD_Welsh-CCG	85.0818	85.2169	**88.1816**	87.2177	87.1592
UD_Western_Armenian-ArmTDP	89.4528	89.3246	**90.1056**	87.1290	87.3939
UD_Wolof-WTB	87.4680	87.2390	**87.7447**	85.2484	85.2699
Average	89.9223	89.9172	**90.6450**	85.5533	85.3939

Table 13: Full results on morphological tagging on dev sets (F1). ST=Single Task (tokenization only), MT=Multi Task, SPL=learn additional SPLits from training data, ML=MultiLingual, LA=Layer Attention

Treebank	MT	MT+SPL	MT+SPL+LA	MT+ML	MT+ML+SPL
UD_Afrikaans-AfriBooms	95.6268	95.8427	96.3228	**97.1391**	**97.1391**
UD_Ancient_Greek-PROIEL	78.6405	78.6831	**80.0425**	74.5249	74.5249
UD_Ancient_Greek-Perseus	71.5125	72.3268	**73.6523**	71.2034	71.2034
UD_Ancient_Hebrew-PTNK	32.0937	32.2163	31.9346	31.8894	**32.8974**
UD_Arabic-PADT	85.5922	85.3074	**86.7182**	76.4808	76.4808
UD_Armenian-ArmTDP	91.4633	91.6519	**92.8398**	92.6570	92.6570
UD_Armenian-BSUT	88.8921	89.1029	91.4169	93.5987	**93.6874**
UD_Basque-BDT	92.4098	92.4983	**93.1128**	90.9898	90.9898
UD_Belarusian-HSE	96.0902	96.1412	**96.3828**	94.8843	94.6799
UD_Bulgarian-BTB	96.1213	96.0619	**96.6709**	94.1783	94.1783
UD_Catalan-AnCora	99.1378	99.1387	**99.1725**	98.6689	98.6689
UD_Chinese-GSD	97.8495	97.7450	**97.8713**	96.9008	96.9008
UD_Chinese-GSDSimp	97.7134	97.6908	**97.7762**	96.9010	97.0723
UD_Classical_Chinese-Kyoto	97.2268	97.2268	**97.4057**	96.7718	96.6682
UD_Coptic-Scriptorium	36.1243	**36.6047**	36.5467	36.0856	36.0856
UD_Croatian-SET	95.9406	95.9424	**96.2295**	95.4596	95.4596
UD_Czech-CAC	98.5718	98.6269	**98.7187**	98.5166	98.4524
UD_Czech-CLTT	93.1764	93.2597	95.9150	**98.7929**	98.6276
UD_Czech-FicTree	97.4207	97.4177	97.8141	**98.2316**	98.2075
UD_Czech-PDT	98.9711	99.0010	**99.0129**	98.5699	98.5517
UD_Danish-DDT	94.9417	94.9651	**95.9768**	95.5432	95.5432
UD_Dutch-Alpino	93.7907	93.9302	**94.1166**	92.9306	92.9306
UD_Dutch-LassySmall	91.1436	91.1157	92.7638	**93.7297**	**93.7297**
UD_English-Atis	99.8194	99.7742	**99.8645**	99.8570	99.8194
UD_English-EWT	97.0945	97.0394	**97.2746**	96.3564	95.7393
UD_English-GUM	97.3933	97.4657	**98.0582**	96.5116	95.4472
UD_English-LinES	95.5284	95.9743	**97.0116**	95.1431	95.6406
UD_English-ParTUT	93.8580	94.6730	**96.1044**	95.1080	94.9622
UD_Estonian-EDT	92.5876	92.4619	**92.7548**	90.8560	90.8560
UD_Estonian-EWT	82.6037	82.2823	84.2053	**90.4273**	90.0867
UD_Faroese-FarPaHC	99.4621	99.5077	**99.5535**	97.6696	97.9615
UD_Finnish-FTB	91.0783	91.0777	**92.0896**	89.7442	89.7442
UD_Finnish-TDT	86.9727	86.7961	**88.4333**	84.0413	83.9105
UD_French-GSD	98.3590	98.3749	**98.4166**	97.9881	97.8970
UD_French-ParTUT	91.5524	91.6872	93.6077	**93.8529**	93.6334
UD_French-Rhapsodie	92.7357	92.8882	94.3412	97.6369	**97.7348**
UD_French-Sequoia	95.5212	95.6343	96.5539	**97.4834**	**97.4834**
UD_Galician-CTG	95.4572	95.3949	**96.1665**	96.1309	96.0413
UD_German-GSD	96.6550	96.5778	**96.8703**	91.3651	90.9232
UD_German-HDT	**96.9240**	96.8760	96.8744	95.9985	96.0133
UD_Gothic-PROIEL	83.8557	83.6538	**86.1394**	82.9980	82.9980
UD_Greek-GDT	88.1095	87.7069	**90.6986**	88.2176	88.2176
UD_Hebrew-HTB	91.0625	90.9972	**92.3730**	91.2419	91.2419
UD_Hebrew-IAHLTwiki	91.5920	91.6853	92.4310	**92.5367**	92.1098
UD_Hindi-HDTB	98.7946	**98.8443**	98.7704	98.6953	98.6953
UD_Hungarian-Szeged	86.8640	87.5914	89.7159	**92.9347**	**92.9347**
UD_Icelandic-IcePaHC	96.0570	96.0170	**96.0945**	95.2531	95.1230
UD_Icelandic-Modern	97.1811	97.3368	**97.7257**	97.4470	97.4127
UD_Indonesian-GSD	96.2569	96.2811	**96.7999**	95.9539	95.9539
UD_Irish-IDT	92.7485	92.6846	**93.2086**	91.3509	91.3509
UD_Italian-ISDT	98.2891	**98.4059**	98.3567	97.9513	97.9513
UD_Italian-MarkIT	88.6879	88.6721	90.5549	**95.6433**	95.3233
UD_Italian-ParTUT	93.1635	93.1443	93.8812	97.4331	**97.5583**
UD_Italian-PoSTWITA	92.5608	92.7526	93.0419	**93.1511**	92.7822
UD_Italian-TWITTIRO	86.3652	86.1948	88.5699	**93.5947**	93.5060
UD_Italian-VIT	97.9157	97.9059	**98.1771**	97.6736	97.6287
UD_Japanese-GSD	**96.3696**	95.9707	96.2356	67.7212	67.5701
UD_Japanese-GSDLUW	95.2771	95.2696	**95.4003**	91.2057	89.7705
UD_Korean-GSD	88.6732	88.2068	**89.2476**	88.3991	88.3888
UD_Korean-Kaist	94.0169	93.9850	**94.1688**	91.8059	91.8059
UD_Latin-ITTB	98.5780	98.5730	**98.6884**	97.9225	97.9225
UD_Latin-LLCT	97.9372	97.9579	**98.2701**	94.7090	94.7954
UD_Latin-PROIEL	93.5944	93.4902	**94.2544**	92.6040	92.6040
UD_Latin-UDante	70.8700	70.8700	72.9186	**83.4929**	83.4871
UD_Latvian-LVTB	95.6235	95.6941	**95.8193**	93.8198	93.8632
UD_Lithuanian-ALKSNIS	86.4631	86.8117	**88.4862**	87.1547	86.9114
UD_Lithuanian-HSE	58.9425	58.3525	60.5977	**83.1801**	**83.1801**
UD_Maltese-MUDT	**99.8384**	99.8041	99.7649	99.5933	99.6129
UD_Marathi-UFAL	69.0000	69.0000	**72.0000**	66.7500	66.7500
UD_Naija-NSC	99.2140	99.1935	**99.2209**	99.0457	99.0252
UD_Norwegian-Bokmaal	98.2979	98.2800	**98.3349**	98.0105	97.9362
UD_Norwegian-Nynorsk	98.1536	98.0754	**98.1761**	97.9314	97.8066
UD_Norwegian-NynorskLIA	95.3613	95.1603	96.5917	**97.5204**	**97.5204**
UD_Old_Church_Slavonic-PROIEL	67.0066	66.5382	**67.5196**	66.6007	66.6007
UD_Old_East_Slavic-Birchbark	38.1896	38.0755	39.1883	**43.4255**	43.1001
UD_Old_East_Slavic-TOROT	67.4814	67.0835	**67.9957**	65.1509	64.2895
UD_Old_French-SRCMF	**99.7470**	**99.7470**	**99.7470**	99.7324	99.7324
UD_Persian-PerDT	97.1821	97.1417	**97.5396**	95.1363	94.7084
UD_Persian-Seraji	97.2771	97.2196	**97.4416**	96.6294	96.6294
UD_Polish-LFG	94.9441	95.3915	**95.8574**	95.1612	95.3055
UD_Polish-PDB	97.0202	97.0168	**97.3322**	95.5495	95.6163
UD_Pomak-Philotis	86.9367	86.8224	**89.0501**	83.2810	83.0887
UD_Portuguese-Bosque	97.1820	97.1191	**97.3713**	91.8504	90.4858
UD_Portuguese-GSD	98.7692	98.6937	**98.7792**	98.4954	98.4954
UD_Romanian-Nonstandard	94.1123	94.1259	**94.4025**	91.9361	91.7188
UD_Romanian-RRT	96.3625	96.2622	**96.6257**	96.3406	96.3406
UD_Romanian-SiMoNERo	97.6337	97.6529	97.9715	**98.2101**	98.0400
UD_Russian-GSD	92.7577	92.5738	94.1222	**95.3565**	**95.3565**
UD_Russian-SynTagRus	97.8133	97.7890	**97.8431**	97.0192	97.0477
UD_Russian-Taiga	89.6175	89.4145	90.0355	89.3012	**91.1551**
UD_Scottish_Gaelic-ARCOSG	**94.6503**	94.5798	94.5385	94.5504	94.3553
UD_Serbian-SET	94.3668	94.4586	**95.6794**	95.4057	95.4057
UD_Slovak-SNK	94.4554	94.2230	**94.9071**	94.0363	94.0363
UD_Slovenian-SSJ	98.0700	98.0455	**98.2624**	97.3737	97.3737
UD_Spanish-AnCora	99.1492	99.1330	**99.1684**	97.7222	97.7222
UD_Spanish-GSD	98.4430	98.5436	**98.6383**	97.3615	97.2513
UD_Swedish-LinES	94.2304	94.1332	**95.3502**	95.1419	95.1419
UD_Swedish-Talbanken	94.8410	94.8722	95.8523	**95.9967**	**95.9967**
UD_Swedish_Sign_Language-SSLC	5.4863	3.0341	44.4444	**95.2864**	**95.2864**
UD_Tamil-TTB	63.2698	66.7846	72.1485	**74.1593**	**74.1593**
UD_Telugu-MTG	**99.7736**	**99.7736**	**99.7736**	**99.7736**	**99.7736**
UD_Turkish-Atis	98.0263	98.1291	98.1695	**98.5692**	98.5385
UD_Turkish-BOUN	88.3342	88.1209	89.1718	**89.7018**	89.6569
UD_Turkish-FrameNet	83.7029	84.6181	85.1513	93.3099	**94.1590**
UD_Turkish-IMST	86.8116	87.1082	88.2172	91.5435	**91.7012**
UD_Turkish-Kenet	90.8138	91.0075	91.2385	**92.0386**	91.6254
UD_Turkish-Penn	92.5580	92.6676	**93.0631**	93.0110	92.6951
UD_Turkish-Tourism	96.5744	96.5693	**97.1821**	95.4817	95.5993
UD_Turkish_German-SAGT	79.5800	79.3145	83.1956	93.7486	**93.8703**
UD_Ukrainian-IU	94.5514	94.5476	**95.5294**	95.0517	94.8194
UD_Urdu-UDTB	96.8622	96.8005	**97.0200**	96.7939	96.7939
UD_Uyghur-UDT	76.0940	75.2547	76.5218	**78.3507**	**78.3507**
UD_Vietnamese-VTB	77.3880	**77.8302**	77.5814	76.9618	76.9618
UD_Welsh-CCG	83.7134	83.8042	**85.9448**	85.7492	85.6904
UD_Western_Armenian-ArmTDP	94.2192	94.2117	**94.6078**	93.2986	93.2916
UD_Wolof-WTB	91.8545	91.8373	**92.1510**	92.1425	92.0530
Average	89.8071	89.8243	90.9796	**90.9957**	90.9396

Table 14: Full results on lemmatization on dev sets (FST=Single Task (tokenization only), MT=Multi Task, SPL=learn additional SPLits from training data, ML=MultiLingual, LA=Layer Attention

		2.2				2.5				2.10			
	% UNKs	sota	base	single	multi	sota	base	single	multi	sota	base	single	multi
UD_Afrikaans-AfriBooms	0.06	99.3003	—	99.0584	99.0881	99.3003	—	99.0877	99.3600	99.3201	—	99.0627	99.3452
UD_Akkadian-PISANDUB	1.68	—	—	—	—	91.8484	—	—	65.1432	91.8484	—	—	51.8429
UD_Akkadian-RIAO	0.10	—	—	—	—	—	—	—	—	98.0343	—	—	92.2763
UD_Akuntsu-TuDeT	0.19	—	—	—	—	—	—	—	—	100.0000	—	—	99.1924
UD_Albanian-TSA	0.00	—	—	—	—	—	—	—	—	99.5127	—	—	99.6743
UD_Amharic-ATT	97.11	100.0000	—	—	99.6763	100.0000	—	—	99.9142	100.0000	—	—	99.8570
UD_Ancient_Greek-PROIEL	5.18	100.0000	100.0000	99.9437	99.9437	100.0000	99.9100	99.9549	99.9887	100.0000	—	99.9437	99.9775
UD_Ancient_Greek-Perseus	5.61	99.9928	99.9800	99.3046	99.2680	99.9928	99.7100	99.3113	99.3295	99.9928	—	99.3808	99.4254
UD_Ancient_Hebrew-PTNK	56.00	—	—	—	—	—	—	—	—	100.0000	—	100.0000	100.0000
UD_Apurina-UFPA	0.48	—	—	—	—	—	—	—	—	100.0000	—	—	99.6119
UD_Arabic-PADT	0.00	99.3019	99.9800	99.8575	99.8430	99.3019	99.9500	99.8534	99.8120	99.3019	—	99.8781	99.8471
UD_Arabic-PUD	0.00	80.6835	—	—	80.3791	80.6835	—	—	80.4161	80.6835	—	—	80.3678
UD_Armenian-ArmTDP	0.42	97.2634	98.0900	98.2731	98.6626	94.6951	98.5200	99.8524	99.8721	94.6858	—	99.8817	99.8522
UD_Armenian-BSUT	0.17	—	—	—	—	—	—	—	—	98.0015	—	99.9265	99.4300
UD_Assyrian-AS	84.97	—	—	—	—	95.2915	—	—	77.0642	95.2915	—	—	77.0642
UD_Bambara-CRB	0.11	—	—	—	—	99.6202	—	—	99.8118	99.6202	—	—	99.8190
UD_Basque-BDT	0.00	99.8811	100.0000	99.8728	99.6920	99.8811	99.8900	99.9261	99.7763	99.8811	—	99.9241	99.6714
UD_Beja-NSC	0.82	—	—	—	—	—	—	—	—	99.4752	—	—	40.5479
UD_Belarusian-HSE	0.66	99.7101	—	99.6745	99.7831	99.9264	99.8100	96.5955	94.3874	97.2965	—	98.2588	98.1385
UD_Bengali-BRU	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	100.0000
UD_Bhojpuri-BHTB	0.45	—	—	—	—	100.0000	—	—	99.8259	99.9550	—	—	99.7975
UD_Breton-KEB	0.37	95.4954	94.4900	—	93.3171	95.4954	—	—	93.0999	95.4954	—	—	93.3740
UD_Bulgarian-BTB	0.00	99.7711	99.9300	99.8505	99.8950	99.7711	99.7800	99.8950	99.8982	99.7711	—	99.8187	99.8568
UD_Buryat-BDT	0.15	99.5905	99.2400	98.4671	99.3105	99.5905	—	98.4001	99.4857	99.5905	—	98.5036	99.3614
UD_Cantonese-HK	8.25	35.0432	—	—	77.5235	32.9637	—	—	79.9715	32.9637	—	—	79.1951
UD_Catalan-AnCora	0.00	93.6988	99.9800	99.9143	99.9195	93.7013	99.9400	99.9602	99.9161	93.7019	—	99.9265	99.9394
UD_Cebuano-GJA	0.00	—	—	—	—	—	—	—	—	99.8335	—	—	99.1674
UD_Chinese-CFL	0.37	21.0607	—	—	85.6986	21.0607	—	—	85.4503	21.0607	—	—	85.2050
UD_Chinese-GSD	0.06	24.6390	96.7100	98.2231	97.0162	24.6390	97.7500	97.8877	97.4263	24.6390	—	98.0247	96.9596
UD_Chinese-GSDSimp	0.57	—	—	—	—	24.6390	—	97.8934	97.4472	24.6390	—	98.0311	96.9540
UD_Chinese-HK	0.92	28.4281	—	—	85.8374	28.2845	—	—	86.0181	28.2845	—	—	85.0730
UD_Chinese-PUD	0.62	24.1758	—	—	92.9968	24.1758	—	—	93.0383	24.1758	—	—	92.9004
UD_Chukchi-HSE	23.15	—	—	—	—	—	—	—	—	100.0000	—	—	81.6290
UD_Classical_Chinese-Kyoto	1.82	—	—	—	—	1.2188	99.7000	99.5880	99.5311	1.2501	—	97.4758	97.8323
UD_Coptic-Scriptorium	88.21	100.0000	—	100.0000	99.8205	99.6838	—	99.5923	99.6226	99.6842	—	99.6740	99.4598
UD_Croatian-SET	0.00	99.9446	99.9300	99.8187	99.8891	99.9382	99.9300	99.8949	99.9031	99.9382	—	99.8825	99.8846
UD_Czech-CAC	0.00	99.9723	100.0000	99.9861	100.0000	99.9723	99.9900	100.0000	99.9861	99.9723	—	100.0000	100.0000
UD_Czech-CLTT	0.06	92.8049	—	99.9512	99.5615	92.8049	99.8900	99.9146	99.5859	92.8252	—	99.9636	99.4306
UD_Czech-FicTree	0.00	99.7473	100.0000	99.9730	99.9700	99.7473	99.9800	99.9730	99.9700	99.7473	—	99.9820	99.9700
UD_Czech-PDT	0.01	99.2391	99.9900	99.9856	99.9559	99.2391	99.9500	99.9891	99.9553	99.2391	—	99.9865	99.9343
UD_Czech-PUD	0.41	99.6469	99.6200	—	99.7632	99.6469	—	—	99.7955	99.6469	—	—	99.7713
UD_Danish-DDT	0.00	99.7005	99.9000	99.7905	99.8504	99.7005	99.8100	99.8354	99.8753	99.7005	—	99.8204	99.8105
UD_Dutch-Alpino	0.00	98.8547	99.9500	99.1085	99.3791	98.8547	99.4300	99.3427	99.3108	98.8547	—	99.0886	99.1285
UD_Dutch-LassySmall	0.00	99.4608	99.8800	99.4638	99.4430	99.5852	99.3600	99.4975	99.4851	99.5859	—	99.4941	99.2783
UD_English-Atis	0.00	—	—	—	—	—	—	—	—	100.0000	—	100.0000	100.0000
UD_English-EWT	0.01	96.4145	99.2600	99.3470	99.0513	96.4145	98.6700	99.3271	98.9137	96.7989	—	99.3576	98.6866
UD_English-GUM	0.90	99.2617	99.8100	99.7497	98.9651	99.1317	99.5200	99.7801	99.0362	97.8824	—	99.6745	99.0040
UD_English-LinES	0.31	99.5129	99.9600	99.9232	99.5973	99.4673	99.4600	99.9321	99.6667	99.4673	—	99.9604	98.8745
UD_English-PUD	0.48	98.5249	99.7400	—	99.3325	98.5249	—	—	99.2588	98.5249	—	—	98.8676
UD_English-ParTUT	0.13	98.8428	—	99.7944	99.3975	98.8428	99.7100	99.8972	99.2943	98.8428	—	99.8384	99.3973
UD_English-Pronouns	0.00	—	—	—	—	99.1124	—	—	98.9368	99.1176	—	—	95.0820
UD_Erzya-JR	1.37	—	—	—	—	99.5671	—	—	98.5158	99.6020	—	—	98.5678
UD_Estonian-EDT	0.34	99.7251	99.9600	99.8110	99.7856	99.6802	99.7500	99.7207	99.8030	99.6801	—	99.7062	99.8258
UD_Estonian-EWT	0.41	—	—	—	—	99.3366	97.7600	97.8406	98.0123	99.0116	—	98.2721	98.2706
UD_Faroese-FarPaHC	0.00	—	—	—	—	—	—	—	—	99.4088	—	99.7047	99.7047
UD_Faroese-OFT	0.04	99.7048	99.5100	—	99.6049	99.7048	—	—	99.5648	99.7048	—	—	99.4406
UD_Finnish-FTB	0.00	99.6133	100.0000	99.9323	99.9139	99.6133	99.8400	99.9231	99.9108	99.6133	—	99.9139	99.9201
UD_Finnish-OOD	0.14	—	—	—	—	—	—	—	—	97.4815	—	—	98.5963
UD_Finnish-PUD	0.58	98.6392	99.6900	—	99.5282	98.6486	—	—	99.5948	98.6486	—	—	99.5916
UD_Finnish-TDT	0.20	99.1225	99.7800	99.7266	99.6886	99.1083	99.7100	99.6933	99.6862	99.1083	—	99.6885	99.6720
UD_French-FQB	0.00	—	—	—	—	88.8344	—	—	99.7539	89.2963	—	—	99.7600
UD_French-GSD	0.00	92.2892	99.7300	99.8101	99.6972	92.2884	99.7700	99.8563	99.7279	92.2907	—	99.8407	99.7071
UD_French-PUD	1.17	92.8378	—	—	99.8115	92.8499	—	—	99.8798	92.8671	—	—	99.8694
UD_French-ParTUT	0.00	92.4419	—	99.8012	99.6222	92.4985	99.7600	99.6817	99.8209	92.4985	—	99.8608	99.8010
UD_French-ParisStories	0.08	—	—	—	—	—	—	—	—	92.1962	—	99.7522	99.7977
UD_French-Rhapsodie	0.35	—	—	—	—	—	—	—	—	90.4823	—	99.8797	99.9170
UD_French-Sequoia	0.00	92.1742	99.8600	99.8614	99.7486	92.1742	99.8100	99.7537	99.7998	92.1726	—	99.8150	99.7125
UD_French-Spoken	0.00	89.6971	100.0000	99.7303	99.1339	90.0200	99.3600	99.7927	99.6611	—	—	—	—
UD_Frisian_Dutch-Fame	0.00	—	—	—	—	—	—	—	—	99.9598	—	—	99.6383
UD_Galician-CTG	0.00	99.5481	99.9100	99.8171	99.7636	99.5481	99.7600	99.7857	99.7506	99.5481	—	99.7949	99.7395
UD_Galician-TreeGal	0.00	99.4475	99.6900	99.5498	99.6192	99.4475	99.4700	99.5767	99.7104	99.4475	—	99.4696	99.6461
UD_German-GSD	1.25	98.0479	99.7000	99.7688	99.7719	98.0599	99.7100	99.7719	98.5664	98.0567	—	99.8674	98.4163
UD_German-HDT	0.00	—	—	—	—	99.7942	99.9200	99.8776	99.8491	99.7942	—	99.8858	99.8426
UD_German-LIT	0.03	—	—	—	—	99.8042	—	—	99.7460	99.8042	—	—	99.7658
UD_German-PUD	0.43	98.3197	—	—	99.6547	98.3065	—	—	98.9723	98.2993	—	—	99.0058
UD_Gothic-PROIEL	1.08	100.0000	100.0000	99.9853	100.0000	100.0000	—	99.9853	99.9853	100.0000	—	99.9706	100.0000
UD_Greek-GDT	0.01	99.5019	99.8800	99.7171	99.5351	99.5019	99.8500	99.8273	99.6021	99.5019	—	99.7889	99.7076
UD_Guajajara-TuDeT	0.32	—	—	—	—	—	—	—	—	100.0000	—	—	100.0000
UD_Guarani-OldTuDeT	0.16	—	—	—	—	—	—	—	—	99.2941	—	—	95.1276
UD_Hebrew-HTB	0.00	97.5349	99.9800	99.9434	99.9037	97.5349	99.8100	99.9434	99.9207	97.5121	—	99.9263	99.8470
UD_Hebrew-IAHLTwiki	0.04	—	—	—	—	—	—	—	—	95.7169	—	99.5349	99.4655
UD_Hindi-HDTB	0.00	100.0000	100.0000	99.9831	99.9915	100.0000	99.8800	99.9944	99.9915	100.0000	—	99.9817	99.9958
UD_Hindi-PUD	0.11	99.3121	—	—	99.7902	99.3121	—	—	99.7776	99.3121	—	—	99.8154
UD_Hittite-HitTB	0.26	—	—	—	—	—	—	—	—	91.7368	—	—	45.4441
UD_Hungarian-Szeged	0.54	99.8948	99.8700	99.8421	99.8852	99.8948	99.5900	99.7560	99.8948	99.8948	—	99.7752	99.9043
UD_Icelandic-IcePaHC	0.02	—	—	—	—	—	—	—	—	99.8143	—	99.8825	99.8793
UD_Icelandic-Modern	0.02	—	—	—	—	—	—	—	—	98.9044	—	99.9563	99.8981
UD_Icelandic-PUD	0.31	—	—	—	—	—	—	—	—	99.8087	—	—	99.8195
UD_Indonesian-CSUI	0.00	—	—	—	—	—	—	—	—	97.8906	—	99.7568	99.5895
UD_Indonesian-GSD	0.14	99.5842	100.0000	99.9066	99.7707	99.5842	99.8900	99.9066	99.7410	99.4285	—	99.6231	99.3454
UD_Indonesian-PUD	0.44	82.0945	—	—	85.5327	82.0945	—	—	84.7605	99.7133	—	—	99.3682
UD_Irish-IDT	0.00	98.2073	99.6000	99.4075	99.5657	98.2073	99.4700	99.4670	99.5509	98.1729	—	99.6536	99.5991
UD_Irish-TwittIrish	0.41	—	—	—	—	—	—	—	—	97.2542	—	—	92.3558
UD_Italian-ISDT	0.05	95.7570	99.9200	99.8967	99.9277	95.7570	99.8800	99.8967	99.8088	95.7570	—	99.9277	99.7777
UD_Italian-MarkIT	0.02	—	—	—	—	—	—	—	—	95.8138	—	99.9122	99.8967
UD_Italian-PUD	0.01	95.9857	—	—	99.7543	95.9857	—	—	99.6798	95.9857	—	—	99.6641
UD_Italian-ParTUT	0.00	95.9752	—	99.7017	99.7464	95.9752	99.8100	99.7464	99.7464	95.9752	—	99.7464	99.7464
UD_Italian-PoSTWITA	0.16	95.6074	99.7600	99.1908	99.0624	95.6074	99.3400	99.2943	99.1247	95.5100	—	99.3199	99.1205
UD_Italian-TWITTIRO	0.55	—	—	—	—	97.4609	99.1500	99.3041	99.5130	97.4609	—	99.4091	99.3564
UD_Italian-VIT	0.00	—	—	—	—	96.6797	99.9700	99.9556	99.9192	96.6357	—	99.9535	99.9252
UD_Italian-Valico	0.00	—	—	—	—	—	—	—	—	96.5335	—	—	99.9001
UD_Japanese-GSD	0.37	19.4389	94.5300	94.9095	93.6476	19.5784	95.2500	94.8422	93.8462	18.5410	—	97.6517	73.8834
UD_Japanese-GSDLUW	0.37	—	—	—	—	—	—	—	—	21.2650	—	97.1609	93.3258
UD_Japanese-Modern	0.80	2.9723	75.6900	—	78.5191	3.2163	—	—	77.5036	3.2163	—	—	73.0378
UD_Japanese-PUD	0.02	21.4267	—	—	95.2093	21.5764	—	—	95.1189	20.8792	—	—	73.6933
UD_Japanese-PUDLUW	0.02	—	—	—	—	—	—	—	—	22.6221	—	—	94.1199
UD_Javanese-CSUI	0.03	—	—	—	—	—	—	—	—	99.2820	—	—	99.0875
UD_Kaapor-TuDeT	1.25	—	—	—	—	—	—	—	—	100.0000	—	—	96.7742
UD_Kangri-KDTB	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	100.0000
UD_Karelian-KKPP	1.39	—	—	—	—	98.5180	—	—	98.4461	98.5180	—	—	98.4930
UD_Karo-TuDeT	0.00	—	—	—	—	—	—	—	—	96.9283	—	—	99.2034
UD_Kazakh-KTB	0.66	97.0689	96.2700	96.8520	97.1330	97.0689	95.9800	96.5587	97.0459	97.0689	—	96.4544	96.5486
UD_Khunsari-AHA	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	66.1654
UD_Kiche-IU	0.02	—	—	—	—	—	—	—	—	99.8616	—	—	70.1648
UD_Komi_Permyak-UH	14.98	—	—	—	—	100.0000	—	—	94.1327	100.0000	—	—	84.4163
UD_Komi_Zyrian-IKDP	5.44	99.3909	—	—	95.7804	99.5327	—	—	95.9907	99.6740	—	—	95.8988
UD_Komi_Zyrian-Lattice	18.67	99.8920	—	—	87.8505	99.7702	—	—	93.9860	99.9004	—	—	84.1459
UD_Korean-GSD	0.32	98.7874	99.8800	99.4993	99.4565	98.7874	98.5700	99.6061	99.4693	98.7874	—	99.5463	99.5334
UD_Korean-Kaist	0.21	100.0000	100.0000	99.8854	99.8536	100.0000	98.7000	99.8695	99.8536	99.8995	—	99.8696	99.8624
UD_Korean-PUD	0.28	77.9232	—	—	78.8040	77.9232	—	—	78.6747	77.9232	—	—	78.9625
UD_Kurmanji-MG	0.04	97.8183	97.3000	97.3278	97.8542	97.8183	94.9500	97.4656	97.8145	97.8183	—	97.3317	97.9111

		2.2				2.5				2.10			
	% UNKs	sota	base	single	multi	sota	base	single	multi	sota	base	single	multi
UD_Latin-ITTB	0.00	99.9716	99.9900	99.9574	100.0000	99.9950	100.0000	99.9950	99.9950	99.9950	—	100.0000	99.9950
UD_Latin-LLCT	0.00	—	—	—	—	—	—	—	—	99.8402	—	99.9564	99.9066
UD_Latin-PROIEL	0.02	100.0000	100.0000	99.9361	99.9539	100.0000	99.8500	99.9539	99.9432	100.0000	—	99.9291	99.9432
UD_Latin-Perseus	0.00	100.0000	100.0000	100.0000	99.9863	100.0000	99.6000	100.0000	99.9726	100.0000	—	100.0000	99.9543
UD_Latin-UDante	0.33	—	—	—	—	—	—	—	—	99.5930	—	99.7204	99.8432
UD_Latvian-LVTB	0.35	99.0634	99.7500	99.6649	99.5921	99.1387	99.7300	99.7727	99.7746	99.1542	—	99.8234	99.7836
UD_Ligurian-GLT	0.07	—	—	—	—	—	—	—	—	89.7461	—	99.0328	98.0887
UD_Lithuanian-ALKSNIS	0.79	—	—	—	—	99.5811	99.8400	99.8755	99.8940	99.5811	—	99.8662	99.8018
UD_Lithuanian-HSE	1.53	99.8115	—	98.0778	99.4340	99.8115	97.7100	97.8048	99.2941	99.8115	—	98.4529	99.2941
UD_Livvi-KKPP	1.69	—	—	—	—	98.0366	—	95.1692	98.4574	98.0366	—	95.3603	98.2233
UD_Low_Saxon-LSDC	0.52	—	—	—	—	—	—	—	—	99.6645	—	—	99.3673
UD_Madi-Jarawara	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	100.0000
UD_Makurap-TuDeT	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	100.0000
UD_Maltese-MUDT	0.95	—	—	—	—	76.6540	—	99.5761	99.4538	76.6540	—	99.5625	99.4900
UD_Manx-Cadhan	0.27	—	—	—	—	—	—	—	—	99.7839	—	—	93.4128
UD_Marathi-UFAL	0.00	100.0000	—	100.0000	100.0000	100.0000	99.2000	100.0000	100.0000	100.0000	—	100.0000	100.0000
UD_Mbya_Guarani-Thomas	0.00	—	—	—	—	99.3187	—	—	87.5227	99.3187	—	—	94.6269
UD_Moksha-JR	0.24	—	—	—	—	100.0000	—	—	98.4889	99.9527	—	—	98.7933
UD_Munduruku-TuDeT	0.19	—	—	—	—	—	—	—	—	98.7763	—	—	81.1565
UD_Naija-NSC	0.00	98.2011	99.7100	—	86.1849	98.2013	—	—	83.2494	96.6137	—	99.9268	99.9268
UD_Nayini-AHA	0.00	—	—	—	—	—	—	—	—	98.0645	—	—	69.9301
UD_Neapolitan-RB	0.00	—	—	—	—	—	—	—	—	82.3529	—	—	84.2105
UD_North_Sami-Giella	0.06	99.3523	99.8500	99.9201	99.8901	99.3523	—	99.9351	99.9350	99.3523	—	99.9500	99.7603
UD_Norwegian-Bokmaal	0.00	99.7695	99.8700	99.8949	99.8698	99.9833	99.8800	99.8782	99.8681	99.7695	—	99.8548	99.8414
UD_Norwegian-Nynorsk	0.01	99.8647	99.9600	99.8102	99.8627	99.8647	—	99.8203	99.8627	99.8647	—	99.8042	99.8365
UD_Norwegian-NynorskLIA	0.17	99.9850	99.9900	99.7106	99.1718	99.9353	—	99.8456	99.7710	99.9353	—	99.8705	99.7859
UD_Old_Church_Slavonic-PROIEL	15.88	99.9850	100.0000	98.9109	98.7231	99.9850	—	98.8666	98.6732	99.9850	—	98.8766	98.5994
UD_Old_East_Slavic-Birchbark	12.67	—	—	—	—	—	—	—	—	80.4157	—	89.9138	89.7301
UD_Old_East_Slavic-RNC	1.08	—	—	—	—	—	—	—	—	97.6460	—	98.6809	99.0113
UD_Old_East_Slavic-TOROT	11.72	—	—	—	—	—	—	—	—	99.9252	—	99.2696	99.2078
UD_Old_French-SRCMF	0.02	93.4987	100.0000	99.9395	99.9222	93.4987	99.9100	99.9654	99.9482	93.8995	—	99.9854	99.9172
UD_Old_Russian-RNC	1.14	—	—	—	—	97.5593	—	—	98.8055	—	—	—	—
UD_Old_Russian-TOROT	11.72	—	—	—	—	99.9252	98.8700	99.2599	99.1916	—	—	—	—
UD_Old_Turkish-Tonqq	50.53	—	—	—	—	—	—	—	—	45.0593	—	—	37.4468
UD_Persian-PerDT	0.00	—	—	—	—	—	—	—	—	99.8594	—	99.9077	99.9328
UD_Persian-Seraji	0.06	100.0000	100.0000	99.8870	99.9027	100.0000	99.2600	99.8870	99.9152	100.0000	—	99.9058	99.8336
UD_Polish-LFG	0.31	96.7620	99.9400	99.7024	99.6527	96.7620	98.3400	99.7024	99.3160	96.7620	—	99.7367	99.0131
UD_Polish-PDB	0.11	—	—	—	—	99.3228	99.9300	99.9071	99.5657	99.3228	—	99.8921	99.6345
UD_Polish-PUD	0.70	—	—	—	—	99.2299	—	—	99.5970	99.2299	—	—	99.6569
UD_Polish-SZ	0.11	99.6963	100.0000	99.9159	98.7946	—	—	—	—	—	—	—	—
UD_Pomak-Philotis	2.84	—	—	—	—	—	—	—	—	99.8864	—	100.0000	99.9830
UD_Portuguese-Bosque	0.00	99.6249	99.7500	99.7568	99.3987	99.6248	99.7500	99.7991	99.2357	99.7265	—	99.8437	99.5648
UD_Portuguese-GSD	0.00	99.9115	—	99.8433	99.5701	99.9115	99.8100	99.8433	99.5308	99.9030	—	99.8161	99.5377
UD_Portuguese-PUD	0.03	99.4028	—	—	99.1354	99.4028	—	—	99.1265	99.4308	—	—	99.1936
UD_Romanian-ArT	1.12	—	—	—	—	—	—	—	—	81.9672	—	—	96.1404
UD_Romanian-Nonstandard	0.03	95.8494	—	98.7201	98.7702	95.8492	98.7400	98.8199	98.7613	95.8492	—	98.8946	98.8020
UD_Romanian-RRT	0.08	97.5932	99.7700	99.5864	99.6538	97.5932	99.6000	99.5864	99.6936	97.5932	—	99.6477	99.5894
UD_Romanian-SiMoNERo	0.01	—	—	—	—	99.4513	—	—	99.0068	98.3115	—	99.5704	99.3406
UD_Russian-GSD	1.11	95.7989	—	99.8311	99.3023	94.6997	99.7900	99.6490	99.3508	94.6997	—	99.7367	99.3418
UD_Russian-PUD	0.27	99.5213	—	—	99.2772	99.6689	—	—	99.6259	99.6689	—	—	99.6696
UD_Russian-SynTagRus	0.05	99.0720	99.7100	99.7319	99.4961	99.0720	99.7100	99.6958	99.6676	99.1204	—	99.7388	99.6995
UD_Russian-Taiga	0.56	96.6688	98.1400	98.9078	98.3041	96.6688	98.9000	98.8299	98.6760	96.6392	—	99.0891	98.7556
UD_Sanskrit-UFAL	0.04	100.0000	—	—	98.9865	100.0000	—	—	99.3234	100.0000	—	—	99.2893
UD_Sanskrit-Vedic	0.19	—	—	—	—	—	—	—	—	100.0000	—	99.8914	99.9121
UD_Scottish_Gaelic-ARCOSG	0.94	—	—	—	—	93.7824	99.4300	99.4721	99.2511	93.9589	—	99.6400	99.5661
UD_Serbian-SET	0.05	99.8715	99.9700	99.9403	99.9311	99.9168	99.9100	99.9562	99.9212	99.9168	—	99.9518	99.9081
UD_Skolt_Sami-Giellagas	16.43	—	—	—	—	99.0625	—	—	64.0867	99.4161	—	—	64.3423
UD_Slovak-SNK	0.17	99.9232	100.0000	99.9655	99.9386	99.9232	99.9400	99.9655	99.9079	99.9568	—	99.9411	99.8822
UD_Slovenian-SSJ	0.04	99.7479	99.9500	99.9218	99.9893	99.8329	99.9700	99.9218	99.9787	99.4544	—	99.9155	99.9627
UD_Slovenian-SST	0.46	87.6068	100.0000	100.0000	99.9850	87.6068	99.8400	100.0000	99.9850	87.6068	—	100.0000	99.9850
UD_Soi-AHA	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	64.6465
UD_South_Levantine_Arabic-MADAR	0.00	—	—	—	—	—	—	—	—	82.5824	—	—	82.5824
UD_Spanish-AnCora	0.02	99.7701	99.9800	99.9151	99.8417	99.7711	99.9100	99.9247	99.7969	99.7711	—	99.9113	99.8217
UD_Spanish-GSD	0.00	99.7912	—	99.9403	99.7357	99.7912	99.9300	99.9403	99.7100	99.7912	—	99.9276	99.7954
UD_Spanish-PUD	0.01	99.7611	—	—	99.6229	99.7611	—	—	99.6624	99.7611	—	—	99.6536
UD_Swedish-LinES	0.20	99.7170	99.9900	99.9501	99.9235	99.7144	99.8900	99.9647	99.9000	99.7144	—	99.9912	99.9088
UD_Swedish-PUD	0.65	99.6046	99.6900	—	99.6988	99.6203	—	—	99.6673	99.6203	—	—	99.7040
UD_Swedish-Talbanken	0.00	99.4832	99.9600	99.8650	99.9632	99.4832	99.9100	99.9019	99.9656	99.4832	—	99.8774	99.9043
UD_Swedish_Sign_Language-SSLC	0.00	65.8228	—	98.9324	98.7611	65.8228	—	98.9324	99.2933	65.8228	—	100.0000	98.7654
UD_Swiss_German-UZH	0.08	—	—	—	—	99.8962	—	—	97.2954	99.8962	—	—	96.8348
UD_Tagalog-TRG	0.00	100.0000	—	—	98.6207	100.0000	—	—	98.6207	100.0000	—	—	99.4536
UD_Tagalog-Ugnayan	0.00	—	—	—	—	—	—	—	—	97.4078	—	—	96.8907
UD_Tamil-MWTT	0.00	—	—	—	—	—	—	—	—	99.9408	—	—	99.9408
UD_Tamil-TTB	0.00	99.9154	—	97.5541	99.2668	99.9154	98.3300	97.5541	98.9876	99.9154	—	98.5352	98.9023
UD_Tatar-NMCTT	0.27	—	—	—	—	—	—	—	—	99.5876	—	—	98.6612
UD_Teko-TuDeT	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	98.7124
UD_Telugu-MTG	0.00	99.7921	—	99.3763	99.3763	99.7921	98.8900	99.3763	99.3065	99.7921	—	99.3763	99.3763
UD_Thai-PUD	0.34	8.6410	69.9300	—	69.6234	8.6410	—	—	68.5583	8.6410	—	—	67.5251
UD_Tupinamba-TuDeT	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	83.9024
UD_Turkish-Atis	0.00	—	—	—	—	—	—	—	—	100.0000	—	99.8649	99.9169
UD_Turkish-BOUN	0.00	—	—	—	—	—	—	—	—	98.5015	—	99.2613	99.0022
UD_Turkish-FrameNet	0.00	—	—	—	—	—	—	—	—	100.0000	—	99.8978	99.8636
UD_Turkish-GB	1.30	—	—	—	—	99.6969	—	—	96.8079	99.6932	—	—	98.8776
UD_Turkish-IMST	0.00	99.4665	99.8900	99.8153	99.8204	99.5968	99.8400	99.9030	99.8928	99.5968	—	99.9439	99.7651
UD_Turkish-Kenet	0.00	—	—	—	—	—	—	—	—	100.0000	—	99.9832	99.9747
UD_Turkish-PUD	0.00	99.1664	—	—	99.6825	99.1664	—	—	99.6915	99.1574	—	—	99.3003
UD_Turkish-Penn	0.10	—	—	—	—	—	—	—	—	98.9144	—	98.7246	98.5946
UD_Turkish-Tourism	0.00	—	—	—	—	—	—	—	—	99.9852	—	99.9852	99.9852
UD_Turkish_German-SAGT	0.00	—	—	—	—	—	—	—	—	99.4604	—	99.8747	99.6528
UD_Ukrainian-IU	0.62	97.6095	99.8300	99.7491	99.6988	97.5094	99.7700	99.7750	99.8101	97.5094	—	99.8627	99.6755
UD_Umbrian-IKUVINA	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	99.7285
UD_Upper_Sorbian-UFAL	0.00	98.2047	98.6400	98.0545	98.8819	98.2047	—	98.0545	98.5666	98.2047	—	98.9905	98.7464
UD_Urdu-UDTB	0.00	99.9021	100.0000	99.9223	99.9291	99.9021	99.7500	99.8615	99.8953	99.9021	—	99.9223	99.9054
UD_Uyghur-UDT	15.55	99.5300	99.9100	96.6610	96.9735	99.5300	97.9500	96.6610	96.6944	99.5300	—	96.8423	96.8643
UD_Vietnamese-VTB	0.01	73.5961	93.4600	93.7682	92.5005	73.5961	94.8800	93.7682	92.3109	73.5961	—	93.8380	92.2227
UD_Warlpiri-UFAL	0.00	100.0000	—	—	100.0000	100.0000	—	—	100.0000	100.0000	—	—	100.0000
UD_Welsh-CCG	0.07	—	—	—	—	92.4474	—	—	98.3572	92.5151	—	99.7611	99.7312
UD_Western_Armenian-ArmTDP	0.16	—	—	—	—	—	—	—	—	96.4403	—	99.8809	99.7989
UD_Wolof-WTB	0.01	—	—	—	—	99.9851	—	99.5771	99.7215	99.9851	—	99.7114	99.8408
UD_Xibe-XDT	91.98	—	—	—	—	—	—	—	—	99.0093	—	—	87.3736
UD_Yakut-YKTDT	4.01	—	—	—	—	—	—	—	—	100.0000	—	—	97.9633
UD_Yoruba-YTB	0.93	98.5030	—	—	82.5743	99.3976	—	—	83.5857	98.6458	—	—	81.0237
UD_Yupik-SLI	0.00	—	—	—	—	—	—	—	—	100.0000	—	—	99.7545

Table 15: Results of the tokenization task on the test set. % UNKs= the percentage of unknown subwords with mBERT. 'base' is the highest performing rule based baseline for each dataset.