Evaluating Morphological Alignment of Tokenizers in 70 Languages

Anonymous Authors[1]

Abstract

While tokenization is a key step in language modeling, with effects on model training and performance, it remains unclear how to effectively evaluate tokenizer quality. One proposed dimension of tokenizer quality is the extent to which tokenizers preserve linguistically meaningful subwords, aligning token boundaries with morphological boundaries within a word. Here, we expand on previous work and develop datasets for 86 languages, which can be used to study tokenizer quality crosslinguistically. We also develop a new evaluation framework, addressing limitations of previous evaluations and providing flexible evaluation for 71 of those languages. We then correlate out alignment scores with downstream task performance for five pre-trained languages models on seven tasks, with at least one task in each of the languages in our sample. We find that morphological alignment does not explain very much variance in model performance, suggesting that morphological alignment alone does not measure dimensions of tokenization quality relevant to model performance.

1. Introduction

Tokenization is the first step of language modeling, in which strings of text are segmented into discrete units in the tokenizer's vocabulary. Tokenization has been shown to have effects on speed and efficiency of language model training (Dagan et al., 2024; Ali et al., 2024; Asgari et al., 2025), performance (Ali et al., 2024), and inference cost and latency (Ahia et al., 2023; Petrov et al., 2023). Despite this, it is still unclear how to best evaluate tokenizers. Finding reliable intrinsic tokenizer evaluation would be enormously valuable, as it would enable tokenizer selection before model training, leading to significant computational and financial savings.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

One of the most frequently used intrinsic tokenizer evaluations is compression. Compression is often measured as the number of tokens it takes to encode a text given a particular tokenizer. It is relatively easy to measure, as it requires simply tokenizing a text and calculating token counts. One metric of compression is fertility, i.e. the number of tokens per word (Rust et al., 2021). Fertility is simple to implement but can be difficult to generalize crosslinguistically, as wordhood is often operationalized as whitespace-separated orthographic units. Not all languages use whitespaces, e.g. Mandarin Chinese, Thai, and Khmer. Corpus token count (CTC; Schmidt et al., 2024) is the total tokens it takes to represent a text for a given tokenizer. CTC can be compared, therefore across tokenizers of different types, vocabulary sizes, etc. It has also been used to compare compression crosslinguistically, by calculating CTC over parallel text in order to determine crosslinguistic differences in compression (Arnett & Bergen, 2025).

Some have argued that increased compression increases the information density for a sequence of fixed length, which could lead to improved model performance (Deletang et al., 2024). There has been empirical evidence to support the claim that more tokenizer compression is correlated with better task performance (Goldman et al., 2024; Gallé, 2019). However, more recent work has shown that there is no robust relationship between tokenizer compression and language model performance (Schmidt et al., 2024).

Other intrinsic tokenizer evaluations have been proposed, such as Rényi efficiency (Zouhar et al., 2023), which takes into account frequency distribution. More optimal Rényi efficiency is associated with having more compression for higher-frequency items and less compression for lower-frequency items. Zouhar et al. (2023) released the `tokenization-scorer` package to support calculation of Rényi efficiency for any tokenized text. However, later work argues it may not provide a holistic metric of good tokenization quality (Cognetta et al., 2024).

Another property of tokenizers that has been studied is how morphologically aligned tokenization is, or to what extent do token boundaries align with morpheme boundaries for a given word. For example, the English word 'books' is composed of the stem 'book' and the plural suffix '-s'. The morphologically aligned segmentation would be [book + s].

Non-aligned segmentations include [boo + ks] or [bo + oks].

There are several studies which show that morphologically aligned tokenization is associated with improved performance on a variety of NLP tasks (Park et al., 2020; Vasiu & Potolea, 2020; Bostrom & Durrett, 2020; Hofmann et al., 2021; Nzeyimana & Niyongabo Rubungo, 2022; Erkaya, 2022; Toraman et al., 2023; Dŕík & Forgac, 2024; Libovický & Helcl, 2024; Jabbar, 2024; Uzan et al., 2024; Bauwens & Delobelle, 2024; Asgari et al., 2025). Despite the volume of work on this topic, it is still difficult to conclude whether morphological alignment of tokenizers *generally* improves downstream performance. Prior work varies widely in language coverage, model architectures, amount of supervision (zero shot through full supervised finetuning), and evaluation metrics (e.g. perplexity versus performance on various downstream tasks).

Batsuren et al. (2024) developed an evaluation in which the tokenization of a given word was classified according to whether words were split into morphemic tokens or non-morphemic tokens, or were stored whole as a single token. The authors found that morphemic tokenization was correlated with better performance. MorphScore (Arnett & Bergen, 2025) expands on this idea and measures how often tokenizer boundaries align with morpheme boundaries for 22 languages. However, the authors found that MorphScore was not predictive of model performance (Arnett & Bergen, 2025). Arnett et al. (2024) found that morphemic tokenization had only a small effect on performance at a subject-verb agreement task in Spanish. There is also evidence from a variety of different languages that morphologically aligned tokenization did not benefit model performance (Macháček et al., 2018; Saleva & Lignos, 2021; Choo & Kim, 2023).

MorphScore is limited, however. While relatively diverse, the language coverage does not include many high-resource languages that are commonly represented in language model research, e.g. French or German. There are also design choices in the creation of MorphScore that limit its potential utility. The items in MorphScore do not have any context. While this does not impact tokenization which uses whitespace pre-tokenization, this makes it impossible to accurately evaluate morphological alignment of superword tokenizers, e.g. SuperBPE (Liu et al., 2025) and BoundlessBPE (Schmidt et al., 2025). Other information from the Universal Dependencies (UD), which were used to create MorphScore was also not included, such as part-of-speech (POS) information or morphological information.

MorphScore also does not take into consideration item frequency. As discussed in Zouhar et al. (2023), optimal tokenization may be dependent on frequency distribution. It may be more important for tokenization of more frequent items to be morphologically aligned, as they occur more often. Or, it may be more important for low-frequency items to be tokenized morphemically, as lower-frequency words are more likely to be segmented into multiple tokens using popular tokenization algorithms like Byte-Pair Encoding (BPE; Gage, 1994; Sennrich et al., 2016).

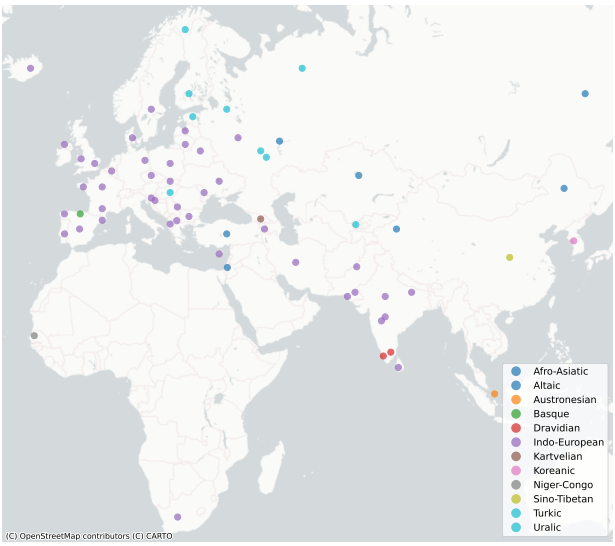

Figure 1. Geographical coverage of language sample.

An expanded and updated evaluation of morphological alignment of tokenizers is key to determining under which settings morphologically aligned tokenization contributes to better model performance. Given the mixed evidence in previous work, a more comprehensive study is necessary. In this paper, we propose a modified and expanded version of MorphScore. We create evaluations for tokenizers in 71 languages. We test the effects of various design decisions, such as including frequency information and the scoring of single-token words on our tokenizer evaluation. Our datasets also include sentential context, POS information, and the morphological information included in UD. While we do not analyze these factors here, we include them in order to enable a broad range of future work.

2. Creating Evaluation Datasets

Data. All datasets are built using the annotations from Universal Dependencies[1]. The exact treebanks we used are listed in Appendix A. For each language, we chose the largest available treebank and used all available splits (train, dev, and test). For each annotated word, we use the wordform and the lemma to determine a proposed segmentation. For example, for the wordform 'launched', the provided lemma is 'launch'. Therefore, by identifying the longest shared sequence between the wordform and lemma, we de-

[1] https://universaldependencies.org/

termine 'launch' to be the stem and '-ed' to be the affix. Any preceding and subsequent characters are treated as the prefix and suffix, respectively. Thus, the gold segmentation will have at least two morphemes (the stem and an affix) and at most three morphemes (a prefix, stem, and suffix).Following, Arnett & Bergen (2025), we only select cases where there the wordform can be recomposed by concatenating the proposed stem and the affixes, in order to remove irregular forms and examples of non-concatenative morphology, where determining a gold segmentation is less straightforward.

Following MorphScore, we used only examples where the identified stem did not undergo suppletion, umlaut, etc., and the wordform could be composed of the stem and either a prefix, as suffix, or both. We observe that without this criterion, we could get gold segementations that would not be informative about the quality of tokenization. For example, the infinitival form of the verb 'to be' in Afrikaans is *wees*. The present form for all persons and numbers is *is*. Under our segmentation approach, the stem would be identified as -s and the proposed gold segmentation would be [i + s]. However, *is* is an irregular form and it should not be thought of as having the stem -s.

In the process of creating and filtering the datasets, despite having very large treebanks, there were not sufficient remaining items from any of the Semitic languages (Amharic, Arabic, and Hebrew) or most isolating languages (e.g. Chinese, Vietnamese, and Thai), which are introflexive languages. In these languages, many morphological processes are encoded using non-concatenative morphology. In particular, these languages often use root template patterns, where a group of consonants is used for a series of related words. Changing the intervening vowels changes the meaning, e.g. from verb to noun (cf. kataba 'he wrote' and kātib 'writer'; Figure 2). Recent work has sought solutions for effective tokenization in languages with these morphological patterns (Gazit et al., 2025).

Isolating languages like Vietnamese and Chinese are not included, because there are not sufficient affixation patterns to create the kind of examples that are selected for by our dataset creation process. In these languages, most words do not have overt morphological markings for number, tense, etc. Therefore, this approach only covers fusional and agglutinative languages. Future work could focus on how to determine gold segmentations for both irregular items, such as the example from Afrikaans, and non-concatenative morphology.

Arabic root k-t-b (ك-ت-ب)

(a) كَتَبَ
kataba
'he wrote'

(b) كاتِب
kātib
'writer'

Figure 2. Example of root template pattern in Arabic.

Once our datasets were created, we filtered out languages for which there were fewer than 100 items. This leaves a set of 71 languages. Their geographical distribution is shown in Figure 1 and all languages are listed in Appendix A. We release the unfiltered datasets, including those that ultimately had too few examples to be scored, on Hugging Face.[2]

Scoring. We expand on MorphScore by incorporating both boundary-level and subword-level evaluations. Specifically, we calculate:

- macro average *boundary* precision and recall
- micro and macro average *subword* precision, recall, and F1

Boundary metrics evaluate whether the predicted tokenization correctly identifies morpheme boundaries, focusing solely on boundary placement. In contrast, subword metrics assess whether the predicted subword spans exactly match gold morphemes.

For example, if the gold segmentation is [book + s] and the predicted tokens are [boo + k + s], only the boundary between 'k' and 's' is correct. This yields a boundary precision of 1/2 and a boundary recall of 1/1. However, for subword metrics, only the token 's' matches a gold morpheme exactly, resulting in a subword precision of 1/3 and recall of 1/2. The code for running scoring is released on GitHub[3].

Oversegmentation and Accuracy. If morphological alignment is measured using accuracy, then a tokenizer can achieve a perfect alignment score by segmenting a word into characters. For example segmenting 'books' into [b + o + o + k + s] leads to an accurate segmentation. This should not be considered a morphologically aligned tokenization. The Llama tokenizers, for several of the languages with non-Latin scripts, tokenize words into tokens more granular than characters, e.g. separating characters and diacritics or decomposing into bytes. Therefore, oversegmentation leads to high accuracy. We find that tokenizing words into more tokens is strongly correlated with morphological alignment as measured with accuracy. In contrast to Arnett & Bergen (2025), we use precision and recall as evaluation

[2]Link removed to preserve anonymity.
[3]Link removed to preserve anonymity.

metrics. Precision, in particular, penalizes tokenizers for oversegmentation.

3. Effect of Design Decisions

Here, we explore the effects of two parameters of the scoring function on alignment score and how they interact with each other. Our goal is to determine the optimal default settings for evaluating morphological alignment.

Frequency Scaling. One parameter we set is whether we weight the morphological alignment score by the wordform frequency, as measured in the UD treebank we used to create the dataset for a given language. Higher-frequency items would be weighted more heavily in the final score than lower-frequency items. Taken frequency distribution into account could lead to a more informative measurement of tokenization quality.

We also test whether there is a correlation between an item's frequency and the likelihood that a tokenizer segments it in a morphologically aligned way. We compute Spearman's rank correlation coefficient across all items and find a weak but statistically significant correlation ($\rho = 0.119$, $p < 0.0001$). The relationship is positive, so more frequent items are more likely to be morphemically segmented.

One-Token Words. Next, we test whether there is a difference in scores depending on whether items that are tokenized into a single token are included in the score calculation. If they are included, the tokenization receives the score associated with a morphologically aligned tokenization. One argument for excluding these items is that these cases do not give any indication of how morphologically aligned a segmentation of a word is, given that there is a segmentation. The alignment score can be inflated for languages where it is possible for the tokenizer to store many whole words in its vocabulary. However, excluding these cases might also essentially penalize a tokenizer for segmenting less. Fewer segmentations leads to better compression, which is thought to be an ideal feature of a tokenizer, as discussed above.

We find there is a significant difference based on the inclusion of one-token items. Morphological alignment scores are generally higher with the inclusion of one-token items, which is what we predicted. We also find an interaction between word frequency and the likelihood that a tokenizer represents a word as a single token. This is a feature of most tokenization algorithms. More frequent items are more likely to be stored in the vocabulary, instead of having to be composed of multiple tokens. In an item-wise test, there is a negative correlation between word frequency and the number of tokens a word is segmented into (Spearman's ρ = -0.108, $p < 0.0001$).

Table 1. Morphological alignment of pre-trained tokenizers.

	Morph. Alignment	
Tokenizer	**Recall**	**Precision**
BLOOM	0.33 ± 0.00	0.11 ± 0.00
Gemma3	0.35 ± 0.00	0.12 ± 0.00
Llama2	**0.56 ± 0.00**	0.13 ± 0.00
Llama3	0.45 ± 0.00	0.12 ± 0.00
XGLM	0.52 ± 0.00	**0.23 ± 0.00**

Optimal Default Settings. We test whether there are differences in morphological alignment scores as we vary frequency scaling and the inclusion of one-token words. We fit a linear mixed effects model with morphological alignment precision as the dependent variable. Frequency scaling, one-token words, and training split are each fixed effects. We test for effects of each of these and their interactions. We include the tokenizer as a random intercept. We report the full statistical results in Appendix B.

There are significant differences across the different categories. We compare the relative ranks according to precision score for the different conditions for five pre-trained tokenizers (Table 1). XGLM consistently has the highest morphological alignment as measured by precision. The other tokenizers' rankings change depending on the different conditions. Measured with recall, Llama2 has the best recall. This is likely due to pervasive oversegmentations. Because of the variable rankings across different metrics and conditions, we determine the optimal default evaluation settings not by maximizing alignments scores, but by determining which is most predictive of language model performance.

4. Correlation with Language Model Performance

Following the analysis in Arnett & Bergen (2025), we take reported model performance scores on a variety of downstream tasks in a range of languages. We test whether there is a correlation between morphological alignment and downstream performance. This serves two purposes. First, we can determine which settings are most predictive of model performance. This could inform choice of settings. Second, we replicate the analysis in Arnett & Bergen (2025), but with the inclusion of many more languages and additional models and tasks.

Method. We use reported model task performance results from Arnett & Bergen (2025). This includes tasks such as XCOPA (Ponti et al., 2020), XNLI (Conneau et al., 2018), and SIB-200 (Adelani et al., 2024). Scores come from Llama2 8B (Touvron et al., 2023), BLOOM (560M, 1.1B,

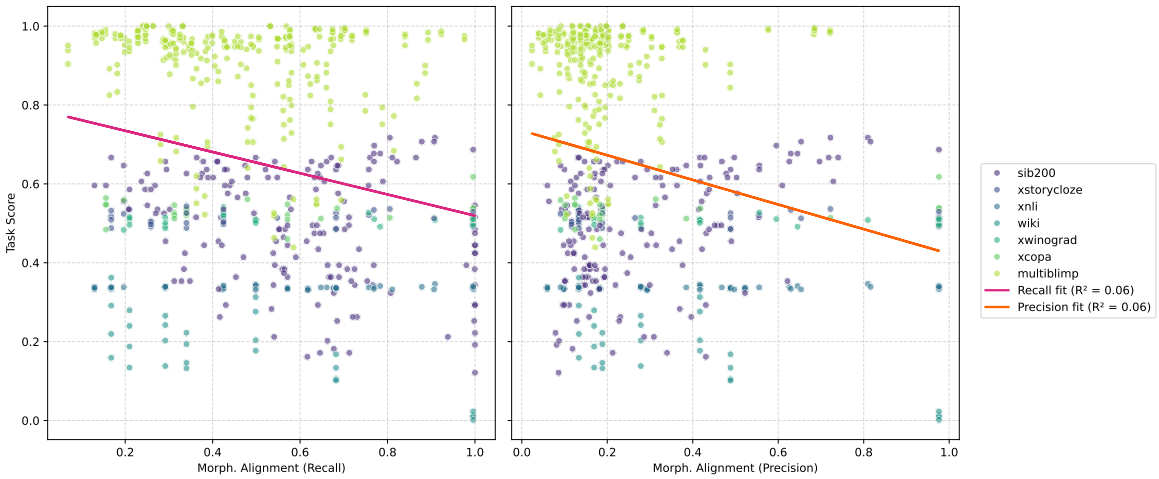

Figure 3. Truex3 condition (check)

3B, 7.1B; Le Scao et al., 2023), and XGLM 7.5B (Lin et al., 2021). We add MultiBLiMP (Jumelet et al., 2025), which tests models' subject-verb agreement performance. We use the results for Llama3 (8B and 70B; Grattafiori et al., 2024) and Gemma3 (4B, 12B, 27B; Team et al., 2025), as reported in the MultiBLiMP paper. The inclusion of MultiBLiMP means we have performance results for all languages in our sample, since MultiBLiMP is also derived from UD. Following the previous study, we use the estimated training data proportions from Hayase et al. (2024), as the model developers do not release that information about the pre-training data.

We test the correlation using linear mixed effects models. As it is known that model size, in parameters, and proportion of the training data in each language impact performance (Kaplan et al., 2020 and Bagheri Nezhad & Agrawal, 2024; Li et al., 2024, respectively), we include these factors as fixed effects. We included benchmark task as a random intercept, as the tasks have different levels of difficulty. We test whether morphological alignment explains additional variance above and beyond these factors using an ANOVA. We also use a simple linear regression to test how much variance morphological alignment explains in the model performance scores.

Results. We find that the fixed effects, number of parameters and proportion of training data in each languages, explains significantly more variance than the intercept ($\chi^2(2) = 25.67$, $p < 0.001$). Morphological alignment, as measured with recall, explains additional variance above and beyond these factors ($\chi^2(1) = 391.42$, $p < 0.001$); however, precision does not ($\chi^2(1) = -6.99$, $p = 1$).

Next, we report the amount of variance explained by morphological alignment. We find that the full linear mixed effects model only explains a small fraction of the variance (recall $R^2 = 0.024$, precision $R^2 = 0.005$). We also plot the relationship between both metrics of morphological alignment and model performance in Figure 3.

In addition to being a very small effect, the correlation between morphological alignment and model performance is *negative*. This is consistent with the findings in Arnett & Bergen (2025), and challenges claims that morphologically aligned tokenization can contribute to better model performance.

Comparing across condition, we find that the condition which frequency-scales scores and does not include one-token words has slightly more explanatory power for model performance, though we note this difference is numeric and the amount of variance is still quite small. All of the conditions still show small negative correlations with model performance. Therefore, we consider this setting to be an appropriate set of default scoring parameters. We report correlations for each condition in Appendix C.

5. Discussion

Optimal Settings. We tested a variety of parameters in our scoring function, and frequency scaling the scores and leaving out one-token words leads to slightly better prediction of model performance. This suggests that including frequency information does improve predictive power of our morphological alignment metrics. Our frequency metrics came only from the treebanks we used to create our datasets, meaning for some languages the sample was very small. Additionally, many treebanks are created with data from one source, e.g. news articles. In the future, word

frequency could be calculated using larger corpora from a wider range of domains. Another possible change would be to use lemma frequency instead of wordform frequency. Particularly for agglutinative languages, e.g. Turkish, individual wordforms tend to be lower frequency. Any given verb, for example, can have thousands of different forms (Hakkani-Tür et al., 2002). Especially if we aim for our morphological alignment metric to capture how often a tokenizer encodes a word with semantically meaningful tokens, like stems, measuring frequency by the lemma may improve predictive power of our morphological alignment score.

The Relevance of Morphological Alignment. Our results show that our version of morphological alignment score explains relatively little variance in model performance, even after taking into account model size and training data proportion. Given large amount of evidence in support of and against the claim that morphological tokenization helps model performance, these results should not be taken as conclusive. But, maybe it suggests that the relationship should be measured differently. Perhaps, taken in isolation, morphological alignment is not sufficient to classify tokenization as optimal. This seems plausible, given that we saw such a strong tradeoff between compression and morphological alignment, when we use accuracy as a metric. Combining morphological alignment with other intrinsic tokenizer evaluation metrics, like compression or Rényi efficiency, could potentially be more informative.

Future Work. While morphological alignment is not predictive of model performance as it is measured here, we hope our datasets and evaluation metric can be used to better understand multilingual tokenization. There are aspects of our evaluation we do not discuss here. Our implementation offers the ability to retrieve morphological alignment score broken down by POS, for instance. Our evaluation framework is flexible to allow many fine-grained analyses, which may be of interest to the wider research community.

6. Conclusion

In this paper, we develop and expanded and updated evaluation for tokenizer morphological alignment for over 70 languages. We test the impact of several design decisions in the scoring function, and find that the way that alignment is calculated leads to different morphological alignment scores and relative rankings between tokenizers. We also test whether morphological alignment is predictive of model performance, which is predicted by previous work. We find, however, that morphological alignment offers only a small negative correlation. This is consistent with the claim that morphologically aligned tokenization does not positively impact model performance. We release our evaluation framework and our datasets to support more work in this area to better understand what features of tokenizers are associated with better performance.

Limitations

While we significantly expand language coverage of this type of tokenizer evaluation, our language sample is far from comprehensive. Additionally, European languages are over-represented in our sample. This is a result of systemic over-representations in the field and in resources like Universal Dependencies. Other resources like UniMorph could be used to improve language coverage, but UniMorph does not provide sentential context, so additional work would be needed to fully integrate UniMorph data into the framework we developed. We hope that as language coverage continually expands and diversifies, it will be easier to represent a more diverse sample of languages.

In this paper, we use only a small number of tasks to represent model performance. We used evaluations which were available for a wide variety of languages, but such evaluations are limited and generally do not represent most of the languages in our sample. For example XCOPA (Ponti et al., 2020) represents 11 languages and XNLI (Conneau et al., 2018) represents 15 languages. Many of these are high-resource European languages like English, Italian, German, and French or widely spoken languages that have been historically underrepresented in NLP like Swahili and Urdu.

Our focus is on large, autoregressive LMs, which allows for cleaner comparisons but excludes encoder models or those trained with masked language modeling. Our sample of models was not very large, because many models do not provide critical information about their training data. BLOOM and XGLM are the only models which report their training data proportions. We were able to expand our sample of models because of the work by Hayase et al. (2024) estimating training data proportions by language for closed-data models. We also chose to exclude instruction-tuned models, because similarly information about fine-tuning data proportions by language was not available. Furthermore,it was not clear about how to calculate proportion of training data, taking into account pre-training data proportions and fine-tuning data proportions.

Impact Statement

Our work aims to understand tokenization quality, which is an issue that disproportionately affects low-resource languages (Petrov et al., 2023; Ahia et al., 2023). We hope that our work positively contributes towards understanding relevant features of tokenization in a multilingual context and helps improve equity in language technology performance across languages.

References

Adelani, D. I., Liu, H., Shen, X., Vassilyev, N., Alabi, J. O., Mao, Y., Gao, H., and Lee, E.-S. A. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In Graham, Y. and Purver, M. (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 226–245, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.14/.

Agić, Ž. and Ljubešić, N. Universal Dependencies for Croatian (that work for Serbian, too). In Piskorski, J., Pivovarova, L., Šnajder, J., Tanev, H., and Yangarber, R. (eds.), *The 5th Workshop on Balto-Slavic Natural Language Processing*, pp. 1–8, Hissar, Bulgaria, September 2015. INCOMA Ltd. Shoumen, BULGARIA. URL https://aclanthology.org/W15-5301/.

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D., Smith, N., and Tsvetkov, Y. Do all languages cost the same? tokenization in the era of commercial language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9904–9923, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614/.

Ahrenberg, L. LinES: An English-Swedish parallel treebank. In Nivre, J., Kaalep, H.-J., Muischnek, K., and Koit, M. (eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pp. 270–273, Tartu, Estonia, May 2007. University of Tartu, Estonia. URL https://aclanthology.org/W07-2441/.

Akhundjanova, A. and Talamo, L. Universal Dependencies treebank for Uzbek. In Holdt, Š. A., Ilinykh, N., Scalvini, B., Bruton, M., Debess, I. N., and Tudor, C. M. (eds.), *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pp. 1–6, Tallinn, Estonia, March 2025. University of Tartu Library, Estonia. ISBN 978-9908-53-121-2. URL https://aclanthology.org/2025.resourceful-1.1/.

Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., Klug, K., Ebert, J., Doll, N., Buschhoff, J., Jain, C., Weber, A., Jurkschat, L., Abdelwahab, H., John, C., Ortiz Suarez, P., Ostendorff, M., Weinbach, S., Sifa, R., Kesselheim, S., and Flores-Herr, N. Tokenizer choice for LLM training: Negligible or crucial? In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3907–3924, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.247. URL https://aclanthology.org/2024.findings-naacl.247/.

Aranzabe, M. J., Atutxa, A., Bengoetxea, K., Diaz, A., de Ilarraza, I. G., Gojenola, K., and Uria, L. Automatic conversion of the basque dependency treebank to universal dependencies. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, pp. 233, 2015.

Arnardóttir, ., Hafsteinsson, H., Sigursson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., and Steingrímsson, S. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pp. 16–25, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.udw-1.3.

Arnardóttir, ., Hafsteinsson, H., Jasonarson, A., Ingason, A., and Steingrímsson, S. Evaluating a Universal Dependencies conversion pipeline for Icelandic. In Alumäe, T. and Fishel, M. (eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 698–704, Tórshavn, Faroe Islands, May 2023. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.69.

Arnett, C. and Bergen, B. Why do language models perform worse for morphologically complex languages? In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6607–6623, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.441/.

Arnett, C., Rivière, P. D., Chang, T. A., and Trott, S. Different tokenization schemes lead to comparable performance in Spanish number agreement. In Nicolai, G., Chodroff, E., Mailhot, F., and Çöltekin, Ç. (eds.), *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 32–38, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigmorphon-1.4. URL https://aclanthology.org/2024.sigmorphon-1.4/.

Asgari, E., Kheir, Y. E., and Javaheri, M. A. S. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity

for efficient llm training across morphologies. *arXiv preprint arXiv:2502.00894*, 2025.

Augustinus, L., Dirix, P., van Niekerk, D., Schuurman, I., Vandeghinste, V., Van Eynde, F., and van Huyssteen, G. AfriBooms: An online treebank for Afrikaans. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 677–682, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1107/.

Badmaeva, E. and Tyers, F. M. Dependency treebank for buryat. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pp. 1–12, 2017.

Bagheri Nezhad, S. and Agrawal, A. What drives performance in multilingual language models? In Scherrer, Y., Jauhiainen, T., Ljubešić, N., Zampieri, M., Nakov, P., and Tiedemann, J. (eds.), *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pp. 16–27, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.vardial-1.2. URL https://aclanthology.org/2024.vardial-1.2/.

Batchelor, C. Universal dependencies for Scottish Gaelic: syntax. In Lynn, T., Prys, D., Batchelor, C., and Tyers, F. (eds.), *Proceedings of the Celtic Language Technology Workshop*, pp. 7–15, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6902/.

Batsuren, K., Vylomova, E., Dankers, V., Delgerbaatar, T., Uzan, O., Pinter, Y., and Bella, G. Evaluating subword tokenization: Alien subword composition and oov generalization challenge. *arXiv preprint arXiv:2404.13292*, 2024.

Bauwens, T. and Delobelle, P. BPE-knockout: Pruning pre-existing BPE tokenisers with backwards-compatible morphological semi-supervision. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5810–5832, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.324. URL https://aclanthology.org/2024.naacl-long.324/.

Bejček, E., Hajič, J., Hajičová, E., Kolářová, V., and Vidová-Hladká, B. Ud_czech-cac: Czech cac treebank. https://github.com/UniversalDependencies/UD_Czech-CAC, 2022. Universal Dependencies 2.10 release.

Bhat, R. A., Bhatt, R., Farudi, A., Klassen, P., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., Vaidya, A., Vishnu, S. R., et al. The hindi/urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press, 2017.

Bielinskienė, A., Boizou, L., Kovalevskaitė, J., and Rimkutė, E. Lithuanian dependency treebank alksnis. In *Human language technologies–the Baltic perspective*, pp. 107–114. IOS Press, 2016.

Borges Völker, E., Wendt, M., Hennig, F., and Köhn, A. HDT-UD: A very large Universal Dependencies treebank for German. In Rademaker, A. and Tyers, F. (eds.), *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pp. 46–57, Paris, France, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-8006. URL https://aclanthology.org/W19-8006.

Bostrom, K. and Durrett, G. Byte pair encoding is suboptimal for language model pretraining. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4617–4624, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.414. URL https://aclanthology.org/2020.findings-emnlp.414.

Branco, A., Silva, J. R., Gomes, L., and António Rodrigues, J. Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5617–5626, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.603/.

Choo, S. and Kim, W. A study on the evaluation of tokenizer performance in natural language processing. *Applied Artificial Intelligence*, 37(1):2175112, 2023.

Chun, J., Han, N.-R., Hwang, J. D., and Choi, J. D. Building Universal Dependency treebanks in Korean. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. Euro-

pean Language Resources Association (ELRA). URL https://aclanthology.org/L18-1347/.

Cognetta, M., Zouhar, V., Moon, S., and Okazaki, N. Two counterexamples to tokenization and the noiseless channel. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16897–16906, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1469/.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. XNLI: Evaluating cross-lingual sentence representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269/.

Dagan, G., Synnaeve, G., and Roziere, B. Getting the most out of your tokenizer for pre-training and domain adaptation. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ZFYBnLljtT.

Deletang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, M., Orseau, L., Hutter, M., and Veness, J. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jznbgiynus.

Dione, C. B. Ud wolof-wtb. https://github.com/UniversalDependencies/UD_Wolof-WTB, 2024. Version 2.15.

Dobrovoljc, K. and Ljubešić, N. Extending the SSJ Universal Dependencies treebank for Slovenian: Was it worth it? In Pradhan, S. and Kuebler, S. (eds.), *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pp. 15–22, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.law-1.3/.

Dobrovoljc, K., Erjavec, T., and Krek, S. The Universal Dependencies treebank for Slovenian. In Erjavec, T., Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J., and Yangarber, R. (eds.), *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pp. 33–38, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1406. URL https://aclanthology.org/W17-1406/.

Drík, D. and Forgac, F. Slovak morphological tokenizer using the byte-pair encoding algorithm. *PeerJ Computer Science*, 10, 2024. URL https://api.semanticscholar.org/CorpusID:274275647.

Droganova, K., Lyashevskaya, O., and Zeman, D. Data conversion and consistency of monolingual corpora: Russian ud treebanks. In *Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018)*, volume 155, pp. 53–66. Linköping University Electronic Press Linköping, Sweden, 2018.

Eli, M., Zeman, D., and Tyers, F. Ud uyghur-udt. https://github.com/UniversalDependencies/UD_Uyghur-UDT, 2024. Version 2.14.

Erkaya, E. A comprehensive analysis of subword tokenizers for morphologically rich languages. Master's thesis, Boğaziçi University, 2022.

Eslami, S. and Çağrı Çöltekin. UD_Azerbaijani-TueCL: Universal dependencies for azerbaijani (tuecl), May 2024. URL https://github.com/UniversalDependencies/UD_Azerbaijani-TueCL.

Etezadi, R., Karrabi, M., Zare, N., Sajadi, M. B., and Pilehvar, M. T. DadmaTools: Natural language processing toolkit for Persian language. In Hajishirzi, H., Ning, Q., and Sil, A. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pp. 124–130, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-demo.13. URL https://aclanthology.org/2022.naacl-demo.13/.

Faryad, J. and Zeman, D. Ud pashto-sikaram. https://github.com/UniversalDependencies/UD_Pashto-Sikaram, 2024. Version 2.14.

Gage, P. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

Gallé, M. Investigating the effectiveness of bpe: The power of shorter sequences. 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)*, pp. 1375–1381, 2019.

Gazit, B., Shmidman, S., Shmidman, A., and Pinter, Y. Splintering nonconcatenative languages for better tokenization. *arXiv preprint arXiv:2503.14433*, 2025.

Goldman, O., Caciularu, A., Eyal, M., Cao, K., Szpektor, I., and Tsarfaty, R. Unpacking tokenization: Evaluating text compression and its correlation with model performance. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2274–2286, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.134. URL https://aclanthology.org/2024.findings-acl.134/.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guillaume, B., de Marneffe, M.-C., and Perrier, G. Conversion et améliorations de corpus du français annotés en Universal Dependencies [conversion and improvement of Universal Dependencies French corpora]. *Traitement Automatique des Langues*, 60(2):71–95, 2019. URL https://aclanthology.org/2019.tal-2.4/.

Guinovart, X. G. Recursos integrados da lingua galega para a investigación lingüística. In *Gallæcia: Estudos de lingüística portuguesa e galega*, pp. 1037–1048. Universidad de Santiago de Compostela, 2017.

Hakkani-Tür, D. Z., Oflazer, K., and Tür, G. Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities*, 36:381–410, 2002.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3):493–531, 2014. ISSN 1574-020X. doi: 10.1007/s10579-013-9244-1. URL http://dx.doi.org/10.1007/s10579-013-9244-1. Open access.

Hayase, J., Liu, A., Choi, Y., Oh, S., and Smith, N. A. Data mixture inference: What do bpe tokenizers reveal about their training data? In *Proceedings of the ICML 2024 FM-Wild Workshop*. OpenReview.net, 2024. URL https://openreview.net/forum?id=0SRg6Cwx3h. Poster presentation.

Heinecke, J. and Tyers, F. M. Development of a Universal Dependencies treebank for Welsh. In *Proceedings of the Celtic Language Technology Workshop*, pp. 21–31, Dublin, 2019. European Association for Machine Translation. URL https://www.aclweb.org/anthology/W19-6904.

Hellwig, O., Scarlata, S., Ackermann, E., and Widmer, P. The treebank of Vedic Sanskrit. In *Proceedings of the LREC*, 2020.

Hellwig, O., Nehrdich, S., and Sellmer, S. Data-driven dependency parsing of Vedic Sanskrit. *Language Resources & Evaluation*, 57:1173–1206, 2023.

Hladká, B., Hajic, J., Hana, J., Hlavácová, J., Mírovskỳ, J., and Raab, J. The czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89:41, 2008.

Hofmann, V., Pierrehumbert, J., and Schütze, H. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3594–3608, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.279. URL https://aclanthology.org/2021.acl-long.279/.

Irimia, E. and Mititelu, V. B. Racai-rotb: nucleu de corpus de limbă română adnotat sintactic cu relaţii de dependenţă. *Revista Română de Interacţiune Om-Calculator*, 8(2):101–120, 2015.

Jabbar, H. MorphPiece: A linguistic tokenizer for large language models. *arXiv*, 2024. URL https://arxiv.org/pdf/2307.07262.pdf.

Johannsen, A., Alonso, H. M., and Plank, B. Universal dependencies for danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, pp. 157, 2015.

Jumelet, J., Weissweiler, L., and Bisazza, A. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*, 2025.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kote, N., Rushiti, R., Cepani, A., Haveriku, A., Trandafili, E., Meçe, E. K., Rakipllari, E. S., Xhanari, L., and Deda, A. Universal dependencies treebank for standard albanian: A new approach. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pp. 80–89, 2024.

Kotsyba, N., Moskalevskyi, B., Romanenko, M., Samoridna, H., Kosovska, I., Lytvyn, O., Orlenko, O., Dyka, L., Brovko, H., Matushko, B., Onyshchuk, N., Pareviazko, V., Rychyk, Y., Stetsenko, A., Umanets, S., and Masenko, L. Ud ukrainian-iu. https://github.com/UniversalDependencies/UD_Ukrainian-IU, 2024. Version 2.15.

Kuzgun, A., Cesur, N., Yıldız, O. T., Kuyrukçu, O., Yenice, A. B., Arıcan, B. N., and Sanıyar, E. Ud turkish-kenet. https://github.com/UniversalDependencies/UD_Turkish-Kenet, 2021. Version 2.8.

Laan, K. Ud veps-vwt. https://github.com/UniversalDependencies/UD_Veps-VWT, 2024. Version 2.14.

Larasati, S. D., Kuboň, V., and Zeman, D. Indonesian morphology tool (morphind): Towards an indonesian corpus. In *International Workshop on Systems and Frameworks for Computational Morphology*, pp. 119–129. Springer, 2011.

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.

Li, Z., Shi, Y., Liu, Z., Yang, F., Liu, N., and Du, M. Quantifying multilingual performance of large language models across languages. *arXiv e-prints*, pp. arXiv–2404, 2024.

Libovický, J. and Helcl, J. Lexically grounded subword segmentation. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://api.semanticscholar.org/CorpusID:270620835.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., et al. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*, 2021.

Liu, A., Hayase, J., Hofmann, V., Oh, S., Smith, N. A., and Choi, Y. Superbpe: Space travel for language models. *arXiv preprint arXiv:2503.13423*, 2025.

Liyanage, C., Sarveswaran, K., Nadungodage, T., and Pushpananda, R. Sinhala dependency treebank (STB). In Grobol, L. and Tyers, F. (eds.), *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pp. 17–26, Washington, D.C., March 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.udw-1.3/.

Lobzhanidze, I. *Finite-state computational morphology: An analyzer and generator for Georgian*. Springer Nature, 2022.

Lynn, T. *Irish Dependency Treebank*. PhD thesis, Dublin City University, 2016. Available at https://github.com/tlynn747/IrishDependencyTreebank.

Macháček, D., Vidra, J., and Bojar, O. Morphological and language-agnostic word segmentation for NMT. In *International Conference on Text, Speech, and Dialogue*, pp. 277–284. Springer, 2018.

Makazhanov, A., Sultangazina, A., Makhambetov, O., and Yessenbayev, Z. Syntactic annotation of kazakh: Following the universal dependencies guidelines. a report. In *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pp. 338–350, 2015.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. Universal Dependency annotation for multilingual parsing. In Schuetze, H., Fung, P., and Poesio, M. (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 92–97, Sofia, Bulgaria, August 2013a. Association for Computational Linguistics. URL https://aclanthology.org/P13-2017/.

McDonald, R. T., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K. B., Petrov, S., Zhang, H., Täckström, O., et al. Universal dependency annotation for multilingual parsing. In *Proc. of ACL*, 2013b.

Merzhevich, T. and Gerardi, F. F. Ud yakut-yktdt. https://github.com/UniversalDependencies/UD_Yakut-YKTDT, 2022. Version 2.15.

Miletic, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C., and Sibille, J. A four-dialect treebank for Occitan: Building process and parsing experiments. In Zampieri, M., Nakov, P., Ljubešić, N., Tiedemann, J., and Scherrer, Y. (eds.), *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 140–149, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL). URL https://aclanthology.org/2020.vardial-1.13/.

Muischnek, K., Müürisep, K., Puolakainen, T., Aedmaa, E., Kirt, R., and Särg, D. Estonian dependency treebank and its annotation scheme. In *Proceedings of 13th workshop on treebanks and linguistic theories (tlt13)*, pp. 285–291, 2014.

Nzeyimana, A. and Niyongabo Rubungo, A. KinyaBERT: a morphology-aware Kinyarwanda language model. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5347–5363, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.367. URL https://aclanthology.org/2022.acl-long.367/.

Ojha, A. K. and Zeman, D. Universal dependency treebanks for low-resource indian languages: The case of bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pp. 33–38, Marseille, France, May 2020. European Language Resources Association (ELRA).

Palmer, M., Bhatt, R., Narasimhan, B., Rambow, O., Sharma, D. M., and Xia, F. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pp. 14–17, 2009.

Park, K., Lee, J., Jang, S., and Jung, D. An empirical study of tokenization strategies for various Korean NLP tasks. In Wong, K.-F., Knight, K., and Wu, H. (eds.), *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 133–142, Suzhou, China, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.aacl-main.17. URL https://aclanthology.org/2020.aacl-main.17/.

Partanen, N., Blokland, R., Lim, K., Poibeau, T., and Rießler, M. First komi-zyrian universal dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 126–132, 2018. URL https://aclanthology.org/W18-6015.

Petrov, A., La Malfa, E., Torr, P., and Bibi, A. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36: 36963–36990, 2023.

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., and Korhonen, A. XCOPA: A multilingual dataset for causal commonsense reasoning. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185/.

Pretkalniņa, L., Rituma, L., and Saulīte, B. Deriving enhanced universal dependencies from a hybrid dependency-constituency treebank. In *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings 21*, pp. 95–105. Springer, 2018.

Prokopidis, P. and Papageorgiou, H. Universal Dependencies for Greek. In de Marneffe, M.-C., Nivre, J., and Schuster, S. (eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pp. 102–106, Gothenburg, Sweden, May 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-0413/.

Prokopidis, P., Desipri, E., Koutsombogera, M., Papageorgiou, H., and Piperidis, S. Theoretical and practical issues in the construction of a greek dependency corpus. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT-2005)*, 2005.

Pyysalo, S., Kanerva, J., Missilä, A., Laippala, V., and Ginter, F. Universal dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa 2015)*, pp. 163–172. NEALT, 2015. URL https://aclweb.org/anthology/W15-1821.pdf.

Rahman, M. U., Qureshi, S., Pirzada, S., Shah, S., Shaheer, M., Talpur, M. A. A., Sanjrani, Z., and Bauer, J. Ud sindhi-isra. https://github.com/UniversalDependencies/UD_Sindhi-Isra, 2024. Version 1.0.

Ramasamy, L. and Žabokrtský, Z. Prague dependency style treebank for Tamil. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 1888–1894, İstanbul, Turkey, 2012. ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/456.html.

Ravishankar, V. A Universal Dependencies treebank for Marathi. In Hajič, J. (ed.), *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pp. 190–200, Prague, Czech Republic, 2017. URL https://aclanthology.org/W17-7623/.

Rueter, J. Erme ud moksha. Version v1.0, January 2018. URL https://doi.org/10.5281/zenodo.1156112. Zenodo.

Rueter, J. and Tyers, F. Towards an open-source universal-dependency treebank for erzya. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pp. 106–118, 2018.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. How good is your tokenizer? on the monolingual performance of multilingual language models. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

Saleva, J. and Lignos, C. The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In Sorodoc, I.-T., Sushil, M., Takmaz, E., and Agirre, E. (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 164–174, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-srw.22. URL https://aclanthology.org/2021.eacl-srw.22.

Samardžić, T. and Ljubešić, N. Ud serbian-set. https://github.com/UniversalDependencies/UD_Serbian-SET, 2024. Version 2.4.

Sazdov, S. *Sovremen makedonski jazik 4*. Tabernakul, Skopje, 2 edition, 2012. English title: Contemporary Macedonian Language, page 84.

Scannell, K. Universal Dependencies for Manx Gaelic. In de Marneffe, M.-C., de Lhoneux, M., Nivre, J., and Schuster, S. (eds.), *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pp. 152–157, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.udw-1.17/.

Schmidt, C. W., Reddy, V., Zhang, H., Alameddine, A., Uzan, O., Pinter, Y., and Tanner, C. Tokenization is more than compression. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 678–702, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.40. URL https://aclanthology.org/2024.emnlp-main.40/.

Schmidt, C. W., Reddy, V., Tanner, C., and Pinter, Y. Boundless byte pair encoding: Breaking the pre-tokenization barrier. *arXiv preprint arXiv:2504.00178*, 2025.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A. (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

Sharma, T., Varma, D. A., Das, M., and Bajpai, S. Ud_malayalam-ufal: Universal dependencies treebank for malayalam. https://github.com/UniversalDependencies/UD_Malayalam-UFAL, 2021. Universal Dependencies Treebank.

Sheyanova, M. and Tyers, F. M. Annotation schemes in north sámi dependency parsing. In *Proceedings of the 3rd International Workshop for Computational Linguistics of Uralic Languages*, pp. 66–75, 2017.

Shishkina, Y. and Lyashevskaya, O. Sculpting enhanced dependencies for belarusian. In *International Conference on Analysis of Images, Social Networks and Texts*, pp. 137–147. Springer, 2021.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. D. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

Simov, K., Osenova, P., Simov, A., and Kouylekov, M. Design and implementation of the bulgarian hpsg-based treebank. *Research on Language and Computation*, 2: 495–522, 2004.

Solberg, P. E., Skjærholt, A., Øvrelid, L., Hagen, K., and Johannessen, J. B. The Norwegian dependency treebank. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 789–795, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1273/.

Taguchi, C. Ud tatar-nmctt. https://github.com/UniversalDependencies/UD_Tatar-NMCTT, 2024. Version 2.14.

Talamo, L. Introducing staf: The saarbrücken treebank of albanian fiction. *Journal of Open Humanities Data*, 11 (1), 2025.

Taulé, M., Martí, M. A., and Recasens, M. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pp. 96–101, 2008.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Toraman, C., Yilmaz, E. H., Şahinuç, F., and Ozcelik, O. Impact of tokenization on language models: An analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4): 1–21, 2023. URL https://dl.acm.org/doi/10.1145/3578707.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tsarfaty, R. A unified morpho-syntactic scheme of stanford dependencies. In *Proc. of ACL*, 2013.

Tyers, F. M. and Ravishankar, V. A prototype dependency treebank for breton. In *Actes de la 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, 2018. *to appear*.

Tyers, F. M. and Washington, J. N. Towards a free/open-source universal-dependency treebank for kazakh. In *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pp. 276–289, 2015.

Uzan, O., Schmidt, C. W., Tanner, C., and Pinter, Y. Greed is all you need: An evaluation of tokenizer inference methods. *arXiv preprint arXiv:2403.01289*, 2024. URL https://arxiv.org/abs/2403.01289.

Van der Beek, L., Bouma, G., Malouf, R., and Van Noord, G. The alpino dependency treebank. In *Computational linguistics in the Netherlands 2001*, pp. 8–22. Brill, 2002.

Vasiu, M. A. and Potolea, R. Enhancing tokenization by embedding romanian language specific morphology. *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 243–250, 2020. URL https://api.semanticscholar.org/CorpusID:227232820.

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., and Csirik, J. Hungarian dependency treebank. In *LREC*, volume 10, pp. 1855–1862. Citeseer, 2010.

Wróblewska, A. Extended and enhanced polish dependency bank in universal dependencies format. In de Marneffe, M.-C., Lynn, T., and Schuster, S. (eds.), *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 173–182. Association for Computational Linguistics, 2018.

Yavrumyan, M. M. and Anna, S. D. Universal dependencies and the armenian treebank. *Herald of the Social Sciences*, 2:231–244, 2020.

Zeman, D. Slovak dependency treebank in universal dependencies. *Jazykovedny Casopis*, 68(2):385–395, 2017.

Zeman, D. and Nedoluzhko, A. Ud upper sorbian-ufal. https://github.com/UniversalDependencies/UD_Upper_Sorbian-UFAL, 2024. Version 2.14.

Zouhar, V., Meister, C., Gastaldi, J., Du, L., Sachan, M., and Cotterell, R. Tokenization and the noiseless channel. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5184–5207, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.284. URL https://aclanthology.org/2023.acl-long.284/.

İbrahim Benli. Ud_kyrgyz-ktmu: Universal dependencies treebank for kyrgyz. https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU, 2023. Universal Dependencies v2 treebank.

A. Language Sample

Table 2 reports the number of items for each language after filtering and the UD treebank used to create the datset for that language.

Table 2. List of languages, UD sources, and number of items after filtering

Language	ISO 639-3	ISO 15924	Num. Items	Data Source
Afrikaans	afr	latn	1397	UD_Afrikaans-AfriBooms (Augustinus et al., 2016)
Albanian	sqi	latn	366	UD_Albanian-STAF (Talamo, 2025; Kote et al., 2024)
Armenian	hye	armn	5441	UD_Armenian-ArmTDP (Yavrumyan & Anna, 2020)
Azerbaijani	aze	latn	220	UD_Azerbaijani-TueCL (Eslami & Çağrı Çöltekin, 2024)
Basque	eus	latn	12089	UD_Basque-BDT (Aranzabe et al., 2015)
Belarusian	bel	cyrl	9935	UD_Belarusian-HSE (Shishkina & Lyashevskaya, 2021)
Bhojpuri	bho	deva	177	UD_Bhojpuri-BHTB (Ojha & Zeman, 2020)
Breton	bre	latn	233	UD_Breton-KEB (Tyers & Ravishankar, 2018)
Bulgarian	bul	cyrl	5443	UD_Bulgarian-BTB (Simov et al., 2004)
Buriat	bur	cyrl	1983	UD_Buryat-BDT (Badmaeva & Tyers, 2017)
Catalan	cat	latn	1230	UD_Catalan-AnCora (Taulé et al., 2008)
Croatian	hrv	latn	7749	UD_Croatian-SET (Agić & Ljubešić, 2015)
Czech	ces	latn	15059	UD_Czech-CAC (Hladká et al., 2008) (Bejček et al., 2022)
Danish	dan	latn	6680	UD_Danish-DDT (Johannsen et al., 2015)
Dutch	nld	latn	3606	UD_Dutch-Alpino (Van der Beek et al., 2002)
English	eng	latn	3688	UD_English-EWT (Silveira et al., 2014)
Erzya	myv	cyrl	2309	UD_Erzya-JR (Rueter & Tyers, 2018)
Estonian	est	latn	19261	UD_Estonian-EDT (Muischnek et al., 2014)
Finnish	fin	latn	10172	UD_Finnish-TDT (Haverinen et al., 2014) (Pyysalo et al., 2015)
French	fra	latn	6082	UD_French-GSD (Guillaume et al., 2019)
Galician	glg	latn	2879	UD_Galician-CTG (Guinovart, 2017)
Georgian	kat	geor	2535	UD_Georgian-GLC (Lobzhanidze, 2022)
German	deu	latn	31281	UD_German-HDT (Borges Völker et al., 2019)
Greek	ell	grek	691	UD_Greek-GDT (Prokopidis et al., 2005) (Prokopidis & Papageorgiou, 2017)
Hebrew	heb	hebr	4641	UD_Hebrew-HTB (Tsarfaty, 2013) (McDonald et al., 2013b)
Hindi	hin	deva	1301	UD_Hindi-HDTB (Palmer et al., 2009; Bhat et al., 2017)
Hungarian	hun	latn	6350	UD_Hungarian-Szeged (Vincze et al., 2010)
Icelandic	isl	latn	13155	UD_Icelandic-IcePaHC (Arnardóttir et al., 2020; 2023)
Indonesian	ind	latn	2785	UD_Indonesian-GSD (Larasati et al., 2011) (McDonald et al., 2013a)
Irish	gle	latn	2576	UD_Irish-IDT (Lynn, 2016)
Kazakh	kaz	cyrl	2442	UD_Kazakh-KTB (Tyers & Washington, 2015) (Makazhanov et al., 2015)
Kirghiz	kir	cyrl	4221	UD_Kyrgyz-KTMU (İbrahim Benli, 2023)
Komi-Zyrian	kpv	cyrl	1038	UD_Komi_Zyrian-Lattice (Partanen et al., 2018)
Korean	kor	hang	316	UD_Korean-Kaist (Chun et al., 2018)
Latvian	lav	latn	8332	UD_Latvian-LVTB (Pretkalniņa et al., 2018)
Lithuanian	lit	latn	667	UD_Lithuanian-ALKSNIS (Bielinskienė et al., 2016)
Macedonian	mkd	cyrl	153	UD_Macedonian-MTB (Sazdov, 2012)

Table 2. List of languages, UD sources, and number of items after filtering (continued)

Language	ISO 639-3	ISO 15924	Num. Items	Data Source	
Malayalam	mal	mlym	131	UD_Malayalam-UFAL	(Sharma et al., 2021)
Manx	glv	latn	224	UD_Manx-Cadhan	(Scannell, 2020)
Marathi	mar	deva	171	UD_Marathi-UFAL	(Ravishankar, 2017)
Moksha	mdf	cyrl	615	UD_Moksha-JR	(Rueter, 2018)
Northern Sami	sme	latn	664	UD_North_Sami-Giella	(Sheyanova & Tyers, 2017)
Norwegian	nob	latn	13017	UD_Norwegian-Bokmaal	(Solberg et al., 2014)
Occitan	oci	latn	878	UD_Occitan-TTB	(Miletic et al., 2020)
Pashto	pus	arab	155	UD_Pashto-Sikaram	(Faryad & Zeman, 2024)
Persian	fas	arab	11859	UD_Persian-PerDT	(Etezadi et al., 2022)
Polish	pol	latn	10886	UD_Polish-PDB	(Wróblewska, 2018)
Portuguese	por	latn	4559	UD_Portuguese-CINTIL	(Branco et al., 2022)
Romanian	ron	latn	10129	UD_Romanian-RRT	(Irimia & Mititelu, 2015)
Russian	rus	cyrl	21569	UD_Russian-SynTagRus	(Droganova et al., 2018)
Sanskrit	san	deva	16184	UD_Sanskrit-Vedic	(Hellwig et al., 2020; 2023)
Scottish Gaelic	gla	latn	1004	UD_Scottish_Gaelic-ARCOSG	(Batchelor, 2019)
Serbian	srp	latn	3874	UD_Serbian-SET	(Samardžić & Ljubešić, 2024)
Sindhi	snd	arab	3874	UD_Sindhi-Isra	(Rahman et al., 2024)
Sinhala	sin	sinh	196	UD_Sinhala-STB	(Liyanage et al., 2023)
Slovak	slk	latn	3590	UD_Slovak-SNK	(Zeman, 2017)
Slovenian	slv	latn	11383	UD_Slovenian-SSJ	(Dobrovoljc et al., 2017) (Dobrovoljc & Ljubešić, 2022)
Spanish	spa	latn	6658	UD_Spanish-AnCora	(Taulé et al., 2008)
Swedish	swe	latn	6223	UD_Swedish-LinES	(Ahrenberg, 2007)
Tamil	tam	taml	1179	UD_Tamil-TTB	(Ramasamy & Žabokrtský, 2012)
Tatar	tat	cyrl	627	UD_Tatar-NMCTT	(Taguchi, 2024)
Turkish	tur	latn	30076	UD_Turkish-Kenet	(Kuzgun et al., 2021)
Uighur	uig	arab	3073	UD_Uyghur-UDT	(Eli et al., 2024)
Ukrainian	ukr	cyrl	4182	UD_Ukrainian-IU	(Kotsyba et al., 2024)
Upper Sorbian	hsb	latn	867	UD_Upper_Sorbian-UFAL	(Zeman & Nedoluzhko, 2024)
Urdu	urd	arab	981	UD_Urdu-UDTB	(Bhat et al., 2017)
Uzbek	uzb	latn	1867	UD_Uzbek-UT	(Akhundjanova & Talamo, 2025)
Veps	vep	latn	159	UD_Veps-VWT	(Laan, 2024)
Welsh	cym	latn	757	UD_Welsh-CCG	(Heinecke & Tyers, 2019)
Wolof	wol	latn	1355	UD_Wolof-WTB	(Dione, 2024)
Yakut	sah	cyrl	250	UD_Yakut-YKTDT	(Merzhevich & Gerardi, 2022)

B. Full Statistical Results

Tables 3 and 4 report the results of the linear mixed effects models described in Section 4.

Table 3. Precision

Variable	**Coef.**	**Std. Err.**	**z**	**p-value**
Intercept	0.169	0.030	5.541	0.000
Frequency Scaling	**0.102**	**0.008**	**13.565**	**0.000**
Single-Token	**-0.045**	**0.008**	**-5.929**	**0.000**
Frequency Scaling × Single-Token	**-0.087**	**0.011**	**-8.073**	**0.000**
Group Var	0.005	0.025		

Table 4. Recall

Variable	Coef.	Std. Err.	z	p-value
Intercept	0.476	0.042	11.336	0.000
Frequency Scaling	**0.065**	**0.013**	**5.000**	**0.000**
Single-Token	**-0.036**	**0.013**	**-2.787**	**0.005**
Frequency Scaling × Single-Token	**-0.064**	**0.018**	**-3.459**	**0.001**
Group Var	0.008	0.027		

C. Correlation with Model Performance in All Conditions

Figures 4 and 5 show the correlation between task performance by condition. The `True_True` condition indicates that scores were scaled by word frequency and single-token words were excluded. The `True_False` condition indicates that scores were scaled by word frequency and single-token words were included. The `False_True` condition indicates that scores were not scaled by word frequency and single-token words were excluded. The `False_False` condition indicates that scores were not scaled by word frequency and single-token words were included.

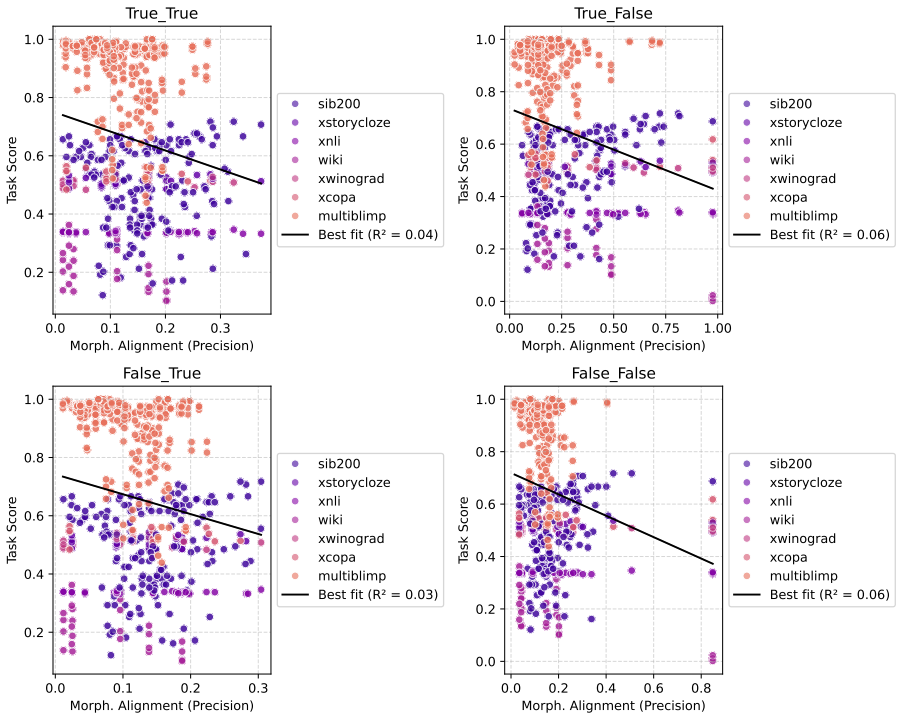

Figure 4. Correlation between morphological alignment measured with precision and task score. Model task is indicated by color.

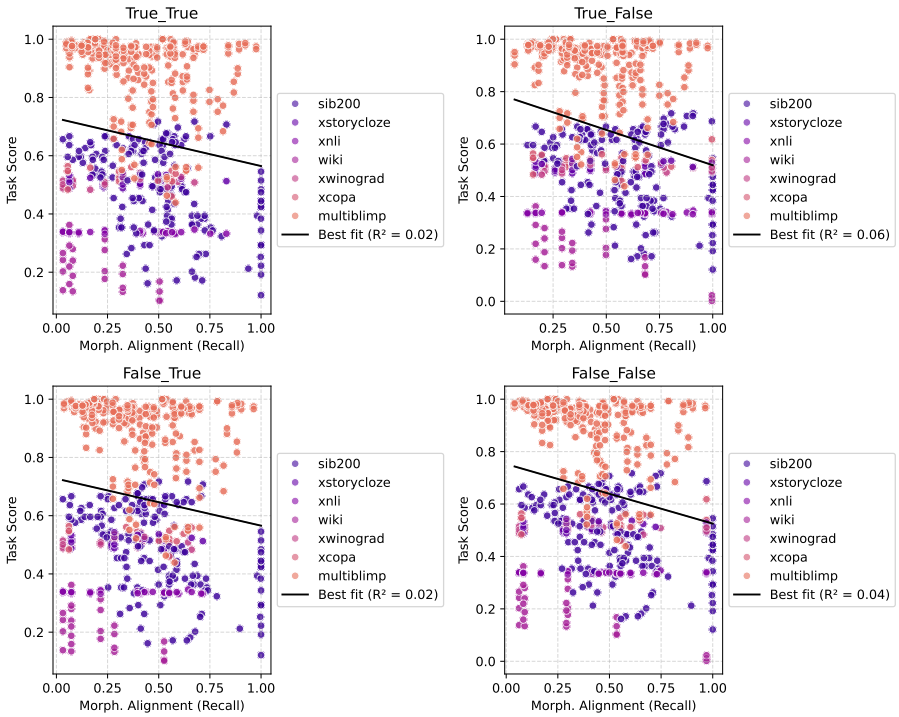

Figure 5. Correlation between morphological alignment measured with recall and task score. Model task is indicated by color.