Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages

Verena Blaschke **Hinrich Schütze** **Barbara Plank**
Center for Information and Language Processing (CIS), LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
blaschke@cis.lmu.de inquiries@cislmu.org bplank@cis.lmu.de

Abstract

One of the challenges with finetuning pretrained language models (PLMs) is that their tokenizer is optimized for the language(s) it was pretrained on, but brittle when it comes to previously unseen variations in the data. This can for instance be observed when finetuning PLMs on one language and evaluating them on data in a closely related language variety with no standardized orthography. Despite the high linguistic similarity, tokenization no longer corresponds to meaningful representations of the target data, leading to low performance in, e.g., part-of-speech tagging.

In this work, we finetune PLMs on seven languages from three different families and analyze their zero-shot performance on closely related, non-standardized varieties. We consider different measures for the divergence in the tokenization of the source and target data, and the way they can be adjusted by manipulating the tokenization during the finetuning step. Overall, we find that the similarity between the percentage of words that get split into subwords in the source and target data (the *split word ratio difference*) is the strongest predictor for model performance on target data.

1 Introduction

Transformer-based pre-trained language models (PLMs) enable successful cross-lingual transfer for many natural language processing tasks. However, the impact of tokenization and its interplay with transferability across languages, especially under-resourced variants with no orthography, has obtained limited focus so far. Tokenization splits words into subwords, but not necessarily in a meaningful way. An example with a current PLM is illustrated for Alsatian German in Figure 1a. This problem is especially pronounced for vernacular languages and dialects, where words tend to be split at a much higher rate than the standard. This has been observed on, e.g., informally written Algerian Arabic (Touileb and Barnes, 2021). As poor subword tokenization can lead to suboptimal language representations and impoverished transfer, it becomes important to understand if the effect holds at a larger scale. We are particularly interested in challenging setups in which, despite *high language similarity,* comparatively low transfer performance is obtained.

a.	M’r	redd	alemànnischi	Mundàrte	.
	M, ’, r	red, ##d	al, ##em, ##à, ##nn, ##isch, ##i	Mund, ##à, ##rte	.
b.	Wir	sprechen	alemannische	Mundarten	.
	Wir	sprechen	al, #emann, ##ische	Mund, ##arten	.
c.	W(r	sprechen	alemaInische	Mundarten	.
	W, (, r	sprechen	al, ##ema, ##In, ##ische	Mund, ##arten	.

Figure 1: “We speak Alemannic dialects”, tokenized by GBERT. Compared to Standard German (*b.*), the quality of the Alsatian German (*a.*) tokenization is poor, making cross-lingual transfer hard. Noise injection (*c.*) often improves transfer from standard to poorly tokenized non-standardized varieties.

A recent study proposes an elegant and lean solution to address this ‘tokenization gap,’ without requiring expensive PLM re-training: to *manipulate tokenization* of PLMs post-hoc (Aepli and Sennrich, 2022), i.e., during finetuning by injecting character-level noise (Figure 1c). Noise injection has been shown to successfully aid cross-lingual transfer and is an appealing solution, as it is cheap and widely applicable. In this work, we first provide a reproduction study and then broaden it by a systematic investigation of the extent to which noise injection helps. We also show how it influences the subword tokenization of the source data vis-à-vis the target data. We hypothesize that, while not emulating dialect text, injecting noise into standard language data can raise the tokenization rate to a similar level, which aids transfer.

The importance of token overlap between source

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 40–54
May 5, 2023 ©2023 Association for Computational Linguistics

and target is an on-going debate (to which we contribute): Prior research has found that subword token overlap between the finetuning and target language improves transfer (Wu and Dredze, 2019; Pires et al., 2019), although it might neither be the most important factor (K et al., 2020; Muller et al., 2022) nor a necessary condition for cross-lingual transfer to work (Pires et al., 2019; Conneau et al., 2020b).

To enable research in this direction, we contribute a novel benchmark. We collected under-resourced language variants covering seven part-of-speech (POS) tagging transfer scenarios within three language families. This collection enables also future work to study cross-lingual and cross-dialect transfer.

Our contributions are:

- We investigate the noise injection method by Aepli and Sennrich (2022) with respect to the ideal noise injection rate for different languages and PLMs.
- To the best of our knowledge, this is the broadest study that focuses specifically on transfer to closely related, non-standardized language varieties with languages from multiple linguistic families. We convert several dialect datasets into a shared tagset (UPOS) and share the conversion scripts.
- We compare the effect of noise injection on the subword tokenization differences between the source and target data, and the effect of these differences on the model performance, and find that the proportions of (un)split words are a better predictor than the ratio of seen subword tokens.

2 Method

We make our code, including scripts for reproducing the benchmark, available at `github.com/mainlp/noisydialect`.

2.1 Injecting Character-Level Noise

We follow the approach by Aepli and Sennrich (2022) to add noise to the finetuning datasets. Given a noise level $0 \leq n \leq 1$ and a finetuning dataset F with a grapheme inventory $\mathcal{I}$,[1] we inject noise into each sentence $S \in F$ as follows: we randomly select $n|S|$ words,[2] and for each of these words, we randomly perform one of the three following actions:

- delete one randomly chosen character
- replace one randomly chosen character with a random character $\in \mathcal{I}$
- insert one random character $\in \mathcal{I}$ into a random slot within the word.

Aepli and Sennrich (2022) investigate transferring POS tagging models to five target languages (Swiss German, Faroese, Old French, Livvi and Karelian) and compare set-ups with no noise ($n = 0$) to adding noise with $n = 0.15$. They find that, when the source and target languages are closely related, the configuration with noise consistently performs better. We additionally experiment with adding noise at higher levels: to 35 %, 55 %, 75 % and 95 % of each sentence's tokens.

2.2 Comparing Datasets via Subword Tokenization

We consider several simple measures of comparing the subword tokenization of the source data with that of the target data:

- *Split word ratio difference:* The (absolute) difference between the ratios of words that were split into subword tokens in the source and target data. (We additionally considered the average number of subword tokens per word, but found that that measure yielded very similar results to the split word ratio difference.)
- *Seen subwords* and *seen words:* The ratios of the target subword tokens and target words,[3] respectively, that are also in the source data. (We also included type-based versions of these measures, but found that they behaved similarly to their token-based counterparts.)
- *Type–token ratio (TTR) ratio:* The subword-level type-token ratio of the target data divided by that of the source data. This is similar to the TTR-based measures used by Lin et al. (2019) and Muller et al. (2022).

[1] Unlike Aepli and Sennrich (2022), we also include non-alphabetic characters in $\mathcal{I}$, as some of the orthographic differences are punctuation-based (see Figure 1).

[2] Excluding words that only contain numerals or punctuation marks.

[3] We consider words here as the annotated units provided by the datasets.

3 Experimental Set-up

3.1 Data

We analyze transfer between eight source and 18 target datasets in the following language varieties (see Appendix A for details):

- Modern Standard Arabic (MSA) (Hajič et al., 2009) → Egyptian, Levantine, Gulf and Maghrebi Arabic (Darwish et al., 2018)
- German (Borges Völker et al., 2019) → Swiss German (Hollenstein and Aepli, 2014), Alsatian German (Bernhard et al., 2019)
- German (Borges Völker et al., 2019), Dutch (Bouma and van Noord, 2017) → Low Saxon (Siewert et al., 2021)
- Norwegian (Nynorsk) (Velldal et al., 2017), Norwegian (Bokmål) (Øvrelid and Hohle, 2016) → West, East and North Norwegian (Øvrelid et al., 2018)
- French (Guillaume et al., 2019) → Picard (Martin et al., 2018)
- French (Guillaume et al., 2019), Spanish (Taulé et al., 2008) → Occitan (Bras et al., 2018)
- Finnish (Pyysalo et al., 2015) → six Finnish dialect groups (University of Turku and Institute for the Languages of Finland)

This list includes varieties from three language families (Afro-Asiatic, Finno-Ugric and Indo-European), written in two types of writing systems (alphabetical and abjad). It also covers a range of different degrees of linguistic relatedness (e.g., the Norwegian dialects are much more closely related to each other and to the standardized varieties than can be said of the Arabic group) and text genres (including tweets, Wikipedia articles, and professionally transcribed interviews). While orthographies for some of our target languages (e.g., Low Saxon) have been proposed, none of these languages have a sole orthography that is used by virtually all speakers.

Many of these corpora are from the Universal Dependencies (UD) project (Zeman et al., 2022), or annotated according to UD's POS tagging scheme (UPOS). For some language varieties, we first make the data compatible with UPOS: We convert the tagsets used for the Arabic dialects and the Finnish dialects to UPOS (Appendix B). To process the Occitan data, we separate contractions (ADP+DET), similarly to the way these cases are handled in other Romance UD treebanks.[4] For the Norwegian dialects, we merge parallel data from the original corpus (dialect vs. orthographic transcriptions) with the orthography-only treebank to get a treebank with dialect transcriptions.[5]

3.2 PLMs

We use two multilingual PLMs: mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a). Additionally, we include one monolingual model per source language. Both multilingual PLMs included all of our source languages in their pretraining data, and mBERT also contains two of our target languages (Low Saxon and Occitan). Details on the PLMs we used can be found in Appendix C.

We use *base*-size, cased versions of all models, and finetune the PLMs on the default training data subsets. We perform a simple grid search to choose one set of hyperparameters to be used for all experiments. This grid search was performed on the German (and Swiss German), Arabic (and Egyptian), and Finnish (and Savonian Finnish) data, using XLM-R and the respective monolingual models. Table 5 in Appendix C contains details on the hyperparameters.

4 Results and Discussion

All results we report are averaged over five different random initializations. Table 1 shows the accuracy scores of the inferred POS tags. We observe similar trends for the macro-averaged F1 score as well.

Zero-shot transfer. Performance on the unseen test languages/dialects is much lower than on the test partitions of the corpora on which the models were finetuned. This is expected, as there are not only orthographic and stylistic differences between the corpora, but also some grammatical differences between the language varieties.

The extent to which performance drops is language-dependent: For instance, the best results for the Finnish dialects are 12–17 percentage points below the best results for the Finnish standard language (XLM-R), whereas the best results for the

[4] E.g., universaldependencies.org/fr/tokenization.html

[5] The resulting scripts are available at github.com/mainlp/{convert-qcri-4dialects, convert-la-murre, convert-restaure-occitan, UD_Norwegian-NynorskLIA_dialect}.

Source	Target	Monolingual PLM						mBERT						XLM-R					
	Noise:	0	15	35	55	75	95	0	15	35	55	75	95	0	15	35	55	75	95
German	Alsatian G.	44	71	76	77	78	77	58	76	78	78	77	76	46	71	76	78	77	77
German	Swiss German	55	78	80	80	79	78	62	78	78	79	78	77	56	77	79	79	79	78
German	*German*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*
German	Low Saxon*	18	35	48	51	58	60	36	61	66	68	67	67	26	44	58	71	71	71
Dutch	Low Saxon*	52	62	63	64	64	63	73	75	75	75	73	72	63	71	73	73	73	72
Dutch	*Dutch*	*98*	*97*	*97*	*95*	*93*	*83*	*97*	*97*	*97*	*96*	*95*	*92*	*98*	*98*	*97*	*96*	*96*	*94*
Bokmål	East N.	35	60	67	65	62	60	57	60	58	57	56	54	66	63	63	62	61	59
Bokmål	North N.	36	63	69	67	65	62	61	61	61	60	60	58	70	66	66	65	64	62
Bokmål	West N.	33	59	66	63	61	59	58	57	56	55	54	53	67	62	61	60	59	57
Nynorsk	East N.	64	69	67	65	62	59	59	59	56	56	55	53	67	66	64	62	60	57
Nynorsk	North N.	67	72	69	68	65	63	62	61	59	60	59	57	71	68	67	66	64	62
Nynorsk	West N.	65	69	66	64	63	60	58	58	56	56	56	54	68	64	63	61	60	58
Bokmål	*Bokmål*	*99*	*98*	*98*	*97*	*96*	*91*	*98*	*98*	*97*	*97*	*96*	*92*	*99*	*98*	*98*	*98*	*97*	*93*
Nynorsk	*Nynorsk*	*98*	*98*	*97*	*97*	*95*	*90*	*97*	*97*	*96*	*96*	*94*	*90*	*98*	*97*	*97*	*96*	*95*	*92*
French	Picard	48	52	52	52	51	48	68	73	74	73	73	72	67	74	76	76	75	75
French	*French*	*89*	*88*	*86*	*83*	*78*	*66*	*98*	*98*	*97*	*97*	*96*	*93*	*98*	*98*	*98*	*98*	*97*	*94*
French	Occitan*	41	44	45	45	45	44	86	87	86	85	85	83	77	81	83	83	82	82
Spanish	Occitan*	62	69	70	69	69	69	83	84	83	82	81	79	72	79	78	79	78	77
Spanish	*Spanish*	*99*	*99*	*97*	*97*	*96*	*89*	*99*	*99*	*98*	*96*	*96*	*91*	*99*	*99*	*98*	*98*	*97*	*93*
MSA	Egyptian A.	67	70	66	62	57	50	59	61	60	58	54	47	64	66	65	62	57	50
MSA	Gulf Arabic	66	69	65	61	56	49	65	65	62	60	55	49	66	66	65	61	57	49
MSA	Levantine A.	64	65	62	58	53	47	56	57	55	53	50	45	59	61	60	57	53	46
MSA	Maghrebi A.	51	54	53	50	46	42	50	51	49	48	46	42	51	53	52	50	47	42
MSA	*MSA*	*94*	*93*	*89*	*83*	*78*	*67*	*96*	*95*	*91*	*85*	*79*	*69*	*96*	*95*	*91*	*86*	*80*	*70*
Finnish	Ostroboth. F.	81	80	79	77	78	75	78	78	76	74	73	70	81	85	86	86	86	84
Finnish	SE Finnish	81	79	77	75	76	73	75	75	73	70	69	66	81	84	84	84	84	82
Finnish	SW Finnish	75	73	72	71	71	70	68	68	67	64	63	61	76	80	80	81	81	79
Finnish	SW trans. area	79	78	77	76	76	74	72	72	70	68	67	65	79	84	84	85	84	83
Finnish	Savonian F.	82	80	78	76	76	73	77	79	76	73	72	69	81	84	85	85	85	83
Finnish	Tavastian F.	81	80	79	78	78	75	76	77	76	73	72	69	81	85	86	86	86	84
Finnish	*Finnish*	*98*	*98*	*98*	*97*	*96*	*94*	*96*	*96*	*96*	*95*	*94*	*93*	*98*	*97*	*97*	*97*	*96*	*94*

Table 1: **Accuracy scores (in %) by language combination, language model and noise level.** Scores are averaged over five initializations. Target languages marked with an asterisk* appear in the training data for mBERT. Rows *in italics* contain scores on the test splits of the datasets used for finetuning. The best accuracy for each language pair and PLM combination is underlined.

Norwegian dialects are 26–29 percentage points below the standard language accuracy (Nynorsk with NorBERT). When we have multiple target dialects for one source language, the target scores tend to be similar to one another across noise levels and PLM choices.

PLM choice matters for low-resource languages. While the models are for the most part indistinguishable in their performance on the source languages, the performance on the target languages can vary substantially. For instance, XLM-R outperforms mBERT and FinBERT on the Finnish dialect data. Similarly, both multilingual models perform much better than the monolingual models on the Low Saxon, Picard and Occitan data, and the reverse is true for the Arabic dialects. Neither the performance on the source languages nor the transfer performance with $n = 0$ reveal which model performs best on the target data when the ideal amount of noise is added.

Effect of noise level on accuracy. The optimal noise level depends on the language pair and on the PLM – there is no universal best noise level choice. In many (but not all) cases, the accuracy rises drastically when increasing the noise level from 0 % to 15 %, and the (positive or negative) differences between subsequent noise levels are less pronounced. The noise level of 15 % used by Aepli and Sennrich (2022) is thus a reasonable choice, although not always optimal. In some cases, the accuracy might be much greater at a different noise level (e.g., in the German→Low German XLM-R set-up the maximum gain compared to using no noise is +42 percentage points; +27 compared to 15 % noise). In other cases, adding any noise at all decreases the performance – most drastically in the case of Bokmål→West Norwegian with XLM-R, where the accuracy drops by 5 percentage points when using 15 % noise instead of no noise at all. However, the general trend is that accuracy as a function of noise has a single global maximum and no local maxima – there is a clear optimum level of noise in almost all cases.[6]

Performance on the standard language test splits from the corpora used for finetuning always decreases when noise is introduced. Whether this is detrimental depends on the language: the accuracy on the German test set only drops very slightly (less than one percentage point) whereas the quality of the tag predictions for MSA deteriorates considerably, independently of the model used.

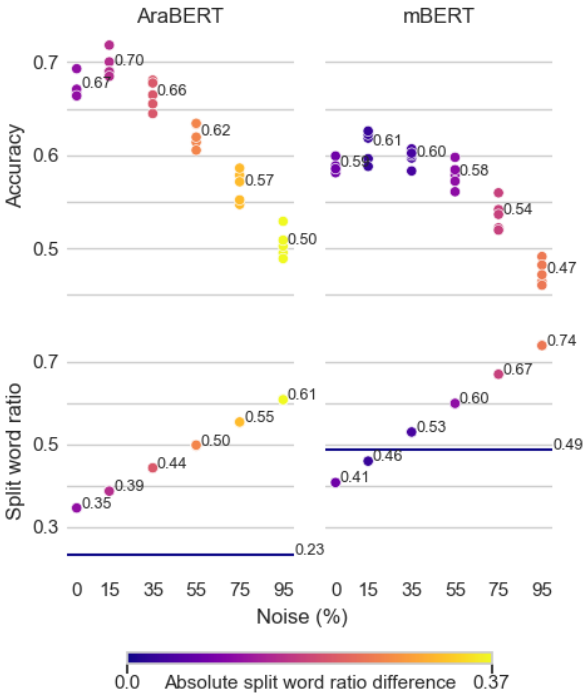

Figure 2: **Transfer from MSA to Egyptian Arabic with AraBERT (left) and mBERT (right).** **Top:** Accuracy scores per language model and noise level (five initializations per set-up; the numbers in the scatterplot indicate the mean accuracy per set-up). **Bottom:** Split word ratios per language model and noise level for the source data (dots) and the target data (dark blue lines) (five initializations per set-up). The colours indicate the (absolute) difference between the split word ratio of the training and target data (darker = smaller difference).

Effect of noise level on split word ratio difference. The words in the target data tend to be split into subword tokens more often than is the case for the source data.[7] Increasing the noise level during finetuning results in the source data being split into more subword tokens (see the rising sequences of dots in the lower part of Figure 2). In all set-ups, the split word ratio of the source data is higher than that of the target data when $n \geq 0.75$.

[6]The minor exceptions to this are FinBERT's performance on the Ostrobothnian and South-East Finnish data and XLM-R's predictions for the Spanish→Occitan transfer (see Table 1). In all of these cases, a second increase occurs after the maximum accuracy has already been reached and stays below this maximum.

[7]The exceptions to this are the tokenization of the Finnish dialects by the multilingual models and the tokenization of the Arabic dialects with AraBERT. The latter is likely due to AraBERT including a pre-tokenization step that splits words into stems and affixes (Antoun et al., 2020), but MSA and non-standard varieties of Arabic having morphological differences.

Src	Target	Monoling.		mBERT		XLM-R	
		ρ	p	ρ	p	ρ	p
Ger.	Als. G.	−0.84	0.00	−0.58	0.00	−0.68	0.00
Ger.	Swiss G.	−0.82	0.00	−0.74	0.00	−0.31	0.10
Ger.	*German*	−0.69	0.00	−0.74	0.00	−0.70	0.00
Ger.	L. Saxon	−0.75	0.00	0.37	0.05	0.44	0.02
Dutch	L. Saxon	−0.84	0.00	−0.71	0.00	−0.50	0.01
Dutch	*Dutch*	−0.89	0.00	−0.88	0.00	−0.91	0.00
Bokm.	East N.	−0.72	0.00	−0.75	0.00	−0.62	0.00
Bokm.	North N.	−0.69	0.00	−0.70	0.00	−0.68	0.00
Bokm.	West N.	−0.75	0.00	−0.85	0.00	−0.72	0.00
Nynor.	East N.	−0.70	0.00	−0.88	0.00	−0.94	0.00
Nynor.	North N.	−0.68	0.00	−0.79	0.00	−0.94	0.00
Nynor.	West N.	−0.64	0.00	−0.85	0.00	−0.95	0.00
Bokm.	*Bokm.*	−0.95	0.00	−0.96	0.00	−0.96	0.00
Nynor.	*Nynor.*	−0.97	0.00	−0.98	0.00	−0.98	0.00
French	Picard	−0.45	0.01	−0.82	0.00	−0.86	0.00
French	*French*	−0.99	0.00	−0.90	0.00	−0.97	0.00
French	Occitan	−0.76	0.00	−0.45	0.01	−0.40	0.03
Spa.	Occitan	−0.64	0.00	−0.38	0.04	−0.71	0.00
Spa.	*Spanish*	−0.95	0.00	−0.95	0.00	−0.97	0.00
MSA	Egy. A.	−0.88	0.00	−0.91	0.00	−0.90	0.00
MSA	Gulf A.	−0.89	0.00	−0.95	0.00	−0.89	0.00
MSA	Lev. A.	−0.91	0.00	−0.87	0.00	−0.83	0.00
MSA	Mag. A.	−0.72	0.00	−0.70	0.00	−0.82	0.00
MSA	*MSA*	−0.96	0.00	−0.96	0.00	−0.96	0.00
Fin.	Ost. F.	−0.46	0.01	−0.90	0.00	0.30	0.11
Fin.	SE F.	−0.71	0.00	−0.93	0.00	0.21	0.27
Fin.	SW F.	−0.09	0.63	−0.94	0.00	0.29	0.12
Fin.	SW tr.	−0.40	0.03	−0.94	0.00	0.27	0.15
Fin.	Sav. F.	−0.68	0.00	−0.89	0.00	0.24	0.20
Fin.	Tav. F.	−0.71	0.00	−0.89	0.00	0.34	0.07
Fin.	*Finnish*	−0.95	0.00	−0.96	0.00	−0.96	0.00

Table 2: **Correlation between split word ratio difference and accuracy.** Spearman's ρ with p-values for all noise levels and random initializations per language pair and PLM. Negative correlations are highlighted in blue, positive ones in yellow. P-values of 0.05 and above have a grey background.

Effect of split word ratio difference on accuracy. Out of the subword tokenization measures introduced in Section 2.2, the *split word ratio difference* correlates most consistently with the performance: the smaller the difference is (i.e., the more similar the ratios are), the higher the accuracy tends to be (Table 2). Figure 2 shows an example; note that the correlation is stronger for the model on the right-hand side (mBERT) than for the model on the left (AraBERT).

The correlation is strong enough that, if one really wants to avoid including the noise level in a hyperparameter search, only carrying out the cheap calculations needed for the *split word ratio difference* and choosing the noise level with the lowest difference can be a proxy. Nevertheless, the correlation is not perfect and this method does not necessarily pick the best noise level.

4.1 Additional Findings

The role of seen (sub)words. Adding noise to the source data initially increases the word and subword token overlap with the target data for all cross-lingual/cross-dialectal set-ups, regardless of model choice. As the noise level increases, this trend ultimately reverses, although the source and target data still have a greater (sub)word overlap at $n = 0.95$ than at $n = 0$.

The seen word ratio and seen subword ratio are much poorer predictors for the model performance than the split word ratio difference is. They are much less consistent and correlate positively with accuracy for many set-ups but negatively for many others, and the correlations tend to have larger p-values (see Tables 6 and 7 in Appendix D for details). While prior works have come to conflicting conclusions regarding the importance of subword token overlap for transfer between more distantly related (or unrelated) languages (Wu and Dredze, 2019; Pires et al., 2019; K et al., 2020; Conneau et al., 2020b; Muller et al., 2022), we find that it is a very poor predictor for the transfer between very closely related languages when injecting character-level noise. One possibility for this is that the seen target subwords contained in the noisy source data might not necessarily belong to the same POS classes.

The role of TTR ratio. For most set-ups, the TTR ratio initially decreases before ultimately increasing, with no local minima. In all of our experiments, the TTR ratio either always stays above one (the target data's TTR remains higher than that of the source data) or always below one (the source data's TTR stays higher than that of the target data; this is only the case for the cross-dialected Finnish set-ups) – adding noise does not result in bringing the TTRs to a similar level. The TTR ratio correlates positively with accuracy for some set-ups and negatively with others (see Table 8 in Appendix D). This overall very weak predictive capacity of the TTR ratio is similar to what Muller et al. (2022) find for named entity recognition and in line with Lin et al.'s (2019) results for POS tagging – their TTR-based measure is only a useful performance predictor when used in conjunction with other measures.

5 Conclusion

We have confirmed the usefulness of the noise injection method by Aepli and Sennrich (2022) for model transfer between closely related languages. To that end, we have converted additional dialectal datasets to the UPOS standard and make the conversion code available to other researchers. Furthermore, we have shown that the ideal amount of noise that should be injected at finetuning time depends on the languages and PLMs used. We have also investigated the role that subword tokenization plays in this and found that the *split word ratio difference* – the (absolute) difference between the proportion of words split into subword tokens in the source and target data – is a reliable, albeit imperfect, predictor of the performance of the transfer model.

Limitations

We include data from three linguistic families, as we were not able to find additional accessible high-quality dialect datasets manually annotated with POS tags for more linguistic families. This general lack of annotated resources is also why we were only able to focus on one NLP task. The tagsets for the Arabic and Finnish varieties were converted to UPOS by a linguist who is not a specialist of Arabic or Finnish.

We only consider one way of modifying the tokenization. In future research, it would be interesting to also consider BPE dropout (Provilkov et al., 2020), which Aepli and Sennrich (2022) show to have an effect on transfer between related languages that is somewhat similar to that of noise injection. It would also be of interest to investigate token-free models like ByT5 (Xue et al., 2022) or CharacterBERT (El Boukkouri et al., 2020), the latter of which has proven useful for processing data in a non-standard variety of Arabic (Riabi et al., 2021).

Acknowledgements

We thank the members of the MaiNLP research group as well as the anonymous reviewers for their useful feedback. This research is supported by European Research Council (ERC) Consolidator Grant DIALECT 101043235. This work was partially funded by the ERC under the European Union's Horizon 2020 research and innovation program (grant 740516).

References

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Delphine Bernhard, Pascale Erhart, Dominique Huck, and Lucie Steiblé. 2019. Annotated corpus for the Alsatian dialects. Version 2.0.

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, and Thomas Lavergne. 2018. Corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.

Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, Gothenburg, Sweden. Association for Computational Linguistics.

Myriam Bras, Louise Esher, Jean Sibille, and Marianne Vergez-Couret. 2018. Annotated corpus for Occitan. Version 1.0.

Kristen E. Brustad. 2000. *The syntax of spoken Arabic*. Georgetown University Press, Washington, D.C., USA.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *Proceedings of the Practical Machine Learning for Developing Countries Workshop*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational*

Linguistics, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. 2014. Using stem-templates to improve Arabic POS and gender/number tagging. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2926–2931, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect Arabic POS tagging: A CRF approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *Computing Research Repository*, arXiv:1912.09582.

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.

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

William Falcon and The PyTorch Lightning team. 2022. PyTorch Lightning.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because size does matter: The Hamburg Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2326–2333, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL*, 60(2):71–95.

Jan Hajič, Otakar Smrž, Petr Zemánek, Petr Pajas, Jan Šnaidauf, Emanuel Beška, Jakub Kracmar, and Kamila Hassanová. 2009. Prague Arabic dependency treebank 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531. Open access.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Ernst Håkon Jahr. 1996. Dialektane i indre Troms: Bardu og Målselv. In Ernst Håkon Jahr and Olav Skare, editors, *Nordnorske dialektar*, pages 180–184. Novus forlag, Oslo.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations (ICLR 2020)*.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Fanny Martin, Christophe Rey, and Philippe Reynés. 2018. Annotated corpus for Picard. Version 4.0.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Benjamin Muller, Deepanshu Gupta, Siddharth Patwardhan, Jean-Philippe Fauconnier, David Vandyke, and Sachin Agarwal. 2022. Languages you know influence those you learn: Impact of language characteristics on multi-lingual text-to-text transfer. *Computing Research Repository*, arXiv:2212.01757.

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1579–1585, Portorož, Slovenia. European Language Resources Association (ELRA).

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA treebank of spoken Norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Curran Associates, Inc.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of NoDaLiDa 2015*, pages 163–172. NEALT.

Arij Riabi, Benoît Sagot, and Djamé Seddah. 2021. Can character-based language models improve downstream task performances in low-resource and noisy language scenarios? In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 423–436, Online. Association for Computational Linguistics.

Karin C. Ryding. 2005. *A Reference Grammar for Modern Standard Arabic*. Cambridge University Press, Cambridge, UK.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Iris Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2022. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *Language Resources and Evaluation*.

Janine Siewert, Yves Scherrer, and Jörg Tiedemann. 2021. Towards a balanced annotated Low Saxon dataset for diachronic investigation of dialectal variation. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 242–246, Düsseldorf, Germany. KONVENS 2021 Organizers.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 789–795, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3700–3712, Online. Association for Computational Linguistics.

University of Turku and Institute for the Languages of Finland. The Finnish dialect corpus of the Syntax Archive, downloadable VRT version.

Leonoor van der Beek, Gosse Bouma, Robert Malouf, and Gertjan van Noord. 2002. The Alpino dependency treebank. In *Computational Linguistics in der Netherlands 2001*, Language and Computers: Studies in Practical Linguistics, pages 8–22. Rodopi.

Erik Velldal, Lilja Øvrelid, and Petter Hohle. 2017. Joint UD parsing of Norwegian Bokmål and nynorsk. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *Computing Research Repository*, arXiv:1912.07076.

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.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. 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.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Chiara Alzetta, Erik Andersen, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Þórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, 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, Juan Belieni, 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, Maria Clara Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Christian Ebert, 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, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Boris Katz, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yixuan 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, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning,

Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, 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, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, 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, Ivan Roksandic, 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, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Ricardo Silva, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Barbara Sonnenhauser, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, 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, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, 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, Giulia Venturi, 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, Vanessa Berwanger Wille, 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 2.11. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Dataset Details

These are the datasets we use in this study:

- Modern Standard Arabic: UD Arabic PADT (Hajič et al., 2009) – CC BY-NC-SA 3.0 – github.com/UniversalDependencies/UD_Arabic-PADT
- Egyptian, Levantine, Gulf and Maghrebi Arabic: QCRI Dialectal Arabic Resources (Darwish et al., 2018) – Apache License 2.0 – alt.qcri.org/resources/da_resources
- German: UD German HDT (Borges Völker et al., 2019; Foth et al., 2014) – CC BY-SA 4.0 – github.com/UniversalDependencies/UD_German-HDT
- Swiss German: NOAH v 3.0 (UPOS-tagged subset) (Hollenstein and Aepli, 2014; Aepli and Sennrich, 2022) – CC BY 4.0 – github.com/noe-eva/NOAH-Corpus

- Alsatian German: Annotated Corpus for the Alsatian Dialects (Bernhard et al., 2019, 2018) – CC BY-SA 4.0 – zenodo.org/record/2536041. Like Swiss German, Alsatian German is a variety of Alemannic German. Note that while both NOAH and the Alsatian corpus contain parts of the Alemannic Wikipedia, the corpora do not overlap.

- Dutch: UD Dutch Alpino (Bouma and van Noord, 2017; van der Beek et al., 2002) – CC BY-SA 4.0 – github.com/UniversalDependencies/UD_Dutch-Alpino

- Low Saxon: UD Low Saxon LSDC (Siewert et al., 2021) – CC BY-SA 4.0 – github.com/UniversalDependencies/UD_Low_Saxon-LSDC

- Norwegian (Nynorsk): UD Norwegian Nynorsk (Velldal et al., 2017; Solberg et al., 2014) – CC BY-SA 4.0 – github.com/UniversalDependencies/UD_Norwegian-Nynorsk

- Norwegian (Bokmål): UD Norwegian Bokmaal (Øvrelid and Hohle, 2016; Solberg et al., 2014) – CC BY-SA 4.0 – github.com/UniversalDependencies/UD_Norwegian-Bokmaal

- West, East and North Norwegian: dialect transcriptions: LIA Norwegian—Corpus of historical dialect recordings (Øvrelid et al., 2018) – CC BY-NC-SA 4.0 – tekstlab.uio.no/LIA/norsk; treebank: UD Norwegian NynorskLIA (Øvrelid et al., 2018) – CC BY-SA 4.0 – github.com/UniversalDependencies/UD_Norwegian-NynorskLIA. The Trønder data (Lierne/Nordli) from the same dataset are omitted because their sample size is much smaller than those of the other dialect groups. We group the remaining locations as follows: East Norwegian (Ål, Bardu,[8] Eidsberg, Gol), West Norwegian (Austevoll, Farsund/Lista, Giske), North Norwegian (Flakstad, Vardø).

- French: UD French GSD (Guillaume et al., 2019) – CC BY-SA 4.0 – github.com/UniversalDependencies/UD_French-GSD

- Picard: Annotated Corpus for Picard (Martin et al., 2018; Bernhard et al., 2018) – CC BY-SA 4.0 – zenodo.org/record/1485988

- Spanish: UD Spanish AnCora (Taulé et al., 2008) – CC BY 4.0 – github.com/UniversalDependencies/UD_Spanish-AnCora

- Occitan: Annotated Corpus for Occitan (Bras et al., 2018; Bernhard et al., 2018) – CC BY-SA 4.0 – zenodo.org/record/1182949

- Finnish: UD Finnish TDT (Pyysalo et al., 2015; Haverinen et al., 2014) – CC BY-SA 4.0 – github.com/UniversalDependencies/UD_Finnish-TDT

- Finnish dialects: The Finnish Dialect Corpus of the Syntax Archive, Downloadable VRT Version (University of Turku and Institute for the Languages of Finland) – CC-BY-ND 4.0 – urn.fi/urn:nbn:fi:lb-2019092001. We use the dialect regions that are indicated in the corpus: South-Western, South-Eastern, Tavastian, Ostrobothnian, and Savonian dialects, as well as dialects from the transition region between the South-Western area and Tavastia.

B Tagset Conversion

B.1 QCRI Dialectal Arabic Resources

To convert the POS tags of the dialectal Arabic dataset, we use the corpus documentation (Darwish et al., 2018), the documentation of the Farasa tagset (Darwish et al., 2014) (on which the corpus's tagset is based), the documentation for Arabic treebanks in general and UD Arabic PADT in particular,[9], grammars of standard and non-standard Arabic (Ryding, 2005; Brustad, 2000), and Sanguinetti et al.'s (2022) tagging recommendations for user-generated content. Table 3 shows how we converted the tags to UPOS. The PART tag is converted to UPOS PART unless the associated word form is one of the subordinating conjunctions tagged as such (SCONJ) in UD Arabic PADT. Tokens tagged with CASE/NSUFF or PROG_PART are fused with preceding ADJ/NOUN or VERB tokens, when possible. When they appear on their own, they are tagged with X. Additional tags from the extended Farasa tagset that are not used in the treebank are: ABBREV, JUS, VSUFF.

[8] The history of the dialects spoken in and around Bardu is complex, as it is a contact point of East and North Norwegian. For more information, see Jahr (1996).

[9] universaldependencies.org/ar/index.html; universaldependencies.org/treebanks/ar_padt

UPOS	Farasa (extended)
ADJ	(DET+)ADJ(+CASE/NSUFF)
ADP	PREP
ADV	ADV
AUX	FUT_PART
CCONJ	CONJ
DET	DET
NOUN	(DET+)NOUN(+CASE/NSUFF)
NUM	NUM
PART	PART,* NEG_PART
PROPN	MENTION
PRON	PRON
PUNCT	PUNC
SCONJ	PART*
SYM	EMOT, URL
VERB	(PROG_PART+)V
X	FOREIGN, HASH, CASE* NSUFF,* PROG_PART*

Table 3: **POS tag conversion for the non-standard Arabic varieties.** The treatment of tags marked with an asterisk* is explained in the text.

UPOS	Finnish Dialect Corpus
ADJ	a, a:pron, a:pron:dem, a:pron:int, a:pron:rel, num:ord, num:ord_pron, q*
ADP	p:post, p:pre
ADV	adv, adv:pron, adv:pron:dem, adv:pron:int, adv:pron:rel, adv:q, p:adv
AUX	v*, neg
CCONJ	cnj:coord
DET	–
INTJ	intj
NOUN	n
NUM	num:card, num:murto
PART	–
PROPN	n:prop, n:prop:pname
PRON	pron, pron:dem, pron:int, pron:pers, pron:pers12, pron:ref, pron:rel, q*
PUNCT	punct
SCONJ	cnj:rel, cnj:sub
SYM	–
VERB	v*
X	muu

Table 4: **POS tag conversion for the Finnish Dialect Corpus.** Tags marked with an asterisk* are disambiguated with the help of lexical information.

B.2 Finnish Dialect Corpus of the Syntax Archive

The conversion of the Finnish tags is based on documentation for the Finnish Dialect Corpus,[10] on the UPOS documentation,[11] and on the documentation of the UD Finnish TDT corpus.[12] Table 4 shows the correspondences between the two tagsets. UD Finnish TDT does not use DET or PART. Two tags needed to be further disambiguated: *v* (used for auxiliaries and full verbs) and *q* (used for interrogative words). For these entries, we use the lemma to decide which POS a given word belongs to.

C Language Models

We use the following PLMs:

- mBERT (Devlin et al., 2019)[13] – Apache 2.0 – huggingface.co/bert-base-multilingual-cased. mBERT's pretraining data include all of the source languages from our study. It also includes Low Saxon and Occitan.
- XLM-R (Conneau et al., 2020a) – MIT licence – huggingface.co/xlm-roberta-base. XLM-R's pretraining data also include all of the source languages from our study. The documentation does not specify whether the Norwegian pretraining data are written in Bokmål, Nynorsk, or both. XLM-R was not trained on any of our target languages.
- Arabic: AraBERT v. 2 (Antoun et al., 2020) – custom licence[14] – huggingface.co/aubmindlab/bert-base-arabertv2
- German: GBERT (Chan et al., 2020) – MIT licence – huggingface.co/deepset/gbert-base
- Dutch: BERTje (de Vries et al., 2019) – Apache 2.0 – github.com/wietsedv/bertje

[10] kielipankki.fi/aineistot/la-murre/la-murre-annotaatiot; blogs.helsinki.fi/fennistic-info/files/2020/12/2.-Sananmuodot-morfologia-morfosyntaksi.pdf

[11] universaldependencies.org/u/pos/all.html

[12] universaldependencies.org/treebanks/fi_tdt

[13] The article details the architecture. Information on the multilingual version can be found at github.com/google-research/bert/blob/master/multilingual.md

[14] github.com/aub-mind/arabert/blob/master/arabert/LICENSE

- Norwegian (both Bokmål and Nynorsk): NorBERT v. 2 (Kutuzov et al., 2021) – CC0 1.0 – huggingface.co/ltgoslo/norbert2.
- French: CamemBERT (Martin et al., 2020) – MIT licence – camembert-model.fr
- Spanish: BETO (Cañete et al., 2020) – CC BY 4.0 – huggingface.co/dccuchile/bert-base-spanish-wwm-cased
- Finnish: FinBERT v. 1.0 (Virtanen et al., 2019) – CC BY 4.0 – github.com/TurkuNLP/FinBERT

We also use the *Transformers* (Wolf et al., 2020) and *PyTorch Lightning* (Falcon and The PyTorch Lightning team, 2022; Paszke et al., 2019) libraries for Python. We use the following hyperparameters for finetuning the models:

Parameter	Grid search	Used
Batch size	16, 32	32
Learning rate	3e-5, 2e-5	2e-5
Epochs	1, 2, 3	2
Classifier dropout	(0.1)	0.1

Table 5: **Hyperparameters used during the grid search and for the final experiments.**

D Additional Correlations

Src	Target	Monoling.		mBERT		XLM-R	
		ρ	p	ρ	p	ρ	p
Ger.	Als. G.	0.34	0.03	0.62	0.00	0.69	0.00
Ger.	Swiss G.	0.78	0.00	0.79	0.00	0.64	0.00
Ger.	L. Saxon	0.40	0.01	0.73	0.00	0.86	0.00
Dutch	L. Saxon	0.75	0.00	−0.25	0.19	0.68	0.00
Bokm.	East N.	0.30	0.11	−0.64	0.00	−0.79	0.00
Bokm.	North N.	0.29	0.11	−0.51	0.01	−0.72	0.00
Bokm.	West N.	0.22	0.25	−0.76	0.00	−0.80	0.00
Nynor.	East N.	−0.36	0.05	−0.64	0.00	−0.82	0.00
Nynor.	North N.	−0.42	0.02	−0.56	0.00	−0.80	0.00
Nynor.	West N.	−0.62	0.00	−0.71	0.00	−0.81	0.00
French	Picard	0.82	0.00	0.24	0.21	0.52	0.00
French	Occitan	0.66	0.00	−0.79	0.00	0.48	0.01
Spa.	Occitan	0.69	0.00	−0.87	0.00	0.15	0.44
MSA	Egy. A.	−0.36	0.05	0.27	0.14	0.57	0.00
MSA	Gulf A.	−0.56	0.00	−0.41	0.02	−0.02	0.94
MSA	Lev. A.	−0.58	0.00	−0.30	0.11	0.33	0.11
MSA	Mag. A.	−0.23	0.22	−0.26	0.17	0.27	0.20
Fin.	Ost. F.	0.01	0.98	−0.79	0.00	0.44	0.01
Fin.	SE F.	−0.07	0.71	−0.73	0.00	0.40	0.03
Fin.	SW F.	−0.37	0.05	−0.84	0.00	0.35	0.06
Fin.	SW tr.	−0.12	0.52	−0.83	0.00	0.37	0.05
Fin.	Sav. F.	0.01	0.95	−0.77	0.00	0.47	0.01
Fin.	Tav. F.	−0.06	0.75	−0.73	0.00	0.57	0.00

Table 6: **Correlation between seen subword ratio and accuracy.** Spearman's ρ with p-values for all noise levels and random initializations per language pair and PLM. Negative correlations are highlighted in blue, positive ones in yellow. P-values of 0.05 and above have a grey background.

Src	Target	Monoling.		mBERT		XLM-R	
		ρ	p	ρ	p	ρ	p
Ger.	Als. G.	0.89	0.00	0.77	0.00	0.85	0.00
Ger.	Swiss G.	0.81	0.00	0.76	0.00	0.84	0.00
Ger.	L. Saxon	0.86	0.00	0.72	0.00	0.88	0.00
Dutch	L. Saxon	0.74	0.00	0.26	0.16	0.79	0.00
Bokm.	East N.	0.30	0.11	−0.66	0.00	−0.70	0.00
Bokm.	North N.	0.37	0.04	−0.65	0.00	−0.59	0.00
Bokm.	West N.	0.23	0.21	−0.81	0.00	−0.68	0.00
Nynor.	East N.	−0.48	0.01	−0.76	0.00	−0.86	0.00
Nynor.	North N.	−0.52	0.00	−0.61	0.00	−0.77	0.00
Nynor.	West N.	−0.59	0.00	−0.62	0.00	−0.79	0.00
French	Picard	0.51	0.00	0.62	0.00	0.79	0.00
French	Occitan	0.82	0.00	−0.50	0.00	0.77	0.00
Spa.	Occitan	0.51	0.00	−0.53	0.00	0.44	0.01
MSA	Egy. A.	0.17	0.38	0.39	0.03	0.29	0.12
MSA	Gulf A.	0.01	0.96	0.04	0.85	−0.27	0.19
MSA	Lev. A.	0.07	0.72	0.12	0.54	−0.08	0.69
MSA	Mag. A.	0.30	0.11	0.02	0.93	−0.03	0.90
Fin.	Ost. F.	0.23	0.22	0.41	0.02	0.62	0.00
Fin.	SE F.	−0.11	0.55	0.01	0.94	0.76	0.00
Fin.	SW F.	−0.34	0.06	−0.15	0.42	0.69	0.00
Fin.	SW tr.	0.36	0.05	0.49	0.01	0.44	0.01
Fin.	Sav. F.	0.16	0.41	0.28	0.13	0.75	0.00
Fin.	Tav. F.	0.24	0.20	0.50	0.01	0.61	0.00

Table 7: **Correlation between seen word ratio and accuracy.** Spearman's ρ with p-values for all noise levels and random initializations per language pair and PLM. Negative correlations are highlighted in blue, positive ones in yellow. P-values of 0.05 and above have a grey background.

Src	Target	Monoling.		mBERT		XLM-R	
		ρ	p	ρ	p	ρ	p
Ger.	Als. G.	0.87	0.00	0.37	0.05	0.53	0.00
Ger.	Swiss G.	0.67	0.00	0.11	0.57	0.21	0.26
Ger.	L. Saxon	0.90	0.00	0.62	0.00	0.69	0.00
Dutch	L. Saxon	0.53	0.00	−0.50	0.00	0.15	0.44
Bokm.	East N.	0.16	0.41	−0.80	0.00	−0.49	0.01
Bokm.	North N.	0.15	0.44	−0.67	0.00	−0.45	0.01
Bokm.	West N.	0.13	0.49	−0.73	0.00	−0.56	0.00
Nynor.	East N.	−0.83	0.00	−0.65	0.00	−0.10	0.62
Nynor.	North N.	−0.85	0.00	−0.47	0.01	−0.05	0.80
Nynor.	West N.	−0.92	0.00	−0.61	0.00	−0.06	0.74
French	Picard	−0.14	0.45	0.15	0.42	0.33	0.07
French	Occitan	0.45	0.01	−0.83	0.00	0.39	0.03
Spa.	Occitan	0.36	0.05	−0.95	0.00	−0.41	0.03
MSA	Egy. A.	−0.38	0.04	−0.77	0.00	−0.82	0.00
MSA	Gulf A.	−0.37	0.05	−0.95	0.00	−0.89	0.00
MSA	Lev. A.	−0.31	0.10	−0.91	0.00	−0.84	0.00
MSA	Mag. A.	−0.63	0.00	−0.87	0.00	−0.73	0.00
Fin.	Ost. F.	−0.87	0.00	−0.75	0.00	−0.03	0.87
Fin.	SE F.	−0.87	0.00	−0.76	0.00	−0.17	0.38
Fin.	SW F.	−0.81	0.00	−0.71	0.00	−0.08	0.69
Fin.	SW tr.	−0.81	0.00	−0.69	0.00	−0.10	0.60
Fin.	Sav. F.	−0.87	0.00	−0.77	0.00	−0.09	0.63
Fin.	Tav. F.	−0.87	0.00	−0.80	0.00	0.02	0.90

Table 8: **Correlation between TTR ratio and accuracy.** Spearman's ρ with p-values for all noise levels and random initializations per language pair and PLM. Negative correlations are highlighted in blue, positive ones in yellow. P-values of 0.05 and above have a grey background. The TTR ratio stayed below 1 for all cross-dialectal Finnish set-ups (regardless of PLM choice) and above 1 for all others.