Does manipulating tokenization aid cross-lingual transfer?

A study on POS tagging for non-standardized languages

Verena Blaschke, Hinrich Schütze & Barbara Plank

VarDial @ EACL 2023

Center for Information and Language Processing, LMU Munich
Munich Center for Machine Learning

“We speak Alemannic dialects.”

Wir	sprechen	alemannische	Mundarten	.
`Wir`	`sprechen`	`al`, `#emann`, `##ische`	`Mund`, `##arten`	`.`

“We speak Alemannic dialects.”

M’r	redd	alemànnischi	Mundàrte	.
`M, ’, r`	`red, ##d`	`al, ##em, ##à, ##nn, ##isch, ##i`	`Mund, ##à, ##rte`	`.`
Wir	sprechen	alemannische	Mundarten	.
`Wir`	`sprechen`	`al, #emann, ##ische`	`Mund, ##arten`	`.`

"We speak Alemannic dialects."

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

Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise

Noëmi Aepli[1] and **Rico Sennrich**[1,2]

Do the previous findings on noise injection hold for new data/models?

Do different languages and PLMs have different ideal noise levels?

How does noise injection affect the subword tokenization differences between source and target data?

Zero-shot transfer for POS tagging

Aepli & Sennrich	This work
Alemannic	German → 2 Alemannic varieties
	Dutch, German → Low Saxon
Faroese	Nynorsk, Bokmål → 3 Norwegian dialects
Old French	French → Picard
	Spanish, French → Occitan
2 Karelian lects	Finnish → 6 Finnish dialects
	Modern Standard Arabic → 4 Arabic varieties

Zero-shot transfer for POS tagging

Aepli & Sennrich	This work
Alemannic	German → 2 Alemannic varieties
	Dutch, German → Low Saxon
Faroese	Nynorsk, Bokmål → 3 Norwegian dialects
Old French	French → Picard
	Spanish, French → Occitan
2 Karelian lects	Finnish → 6 Finnish dialects
	Modern Standard Arabic → 4 Arabic varieties

Monolingual BERTs/RoBERTas vs. mBERT vs. XLM-R

Zero-shot transfer for POS tagging

Aepli & Sennrich	This work
Alemannic	German → 2 Alemannic varieties
	Dutch, German → Low Saxon
Faroese	Nynorsk, Bokmål → 3 Norwegian dialects
Old French	French → Picard
	Spanish, French → Occitan
2 Karelian lects	Finnish → 6 Finnish dialects
	Modern Standard Arabic → 4 Arabic varieties

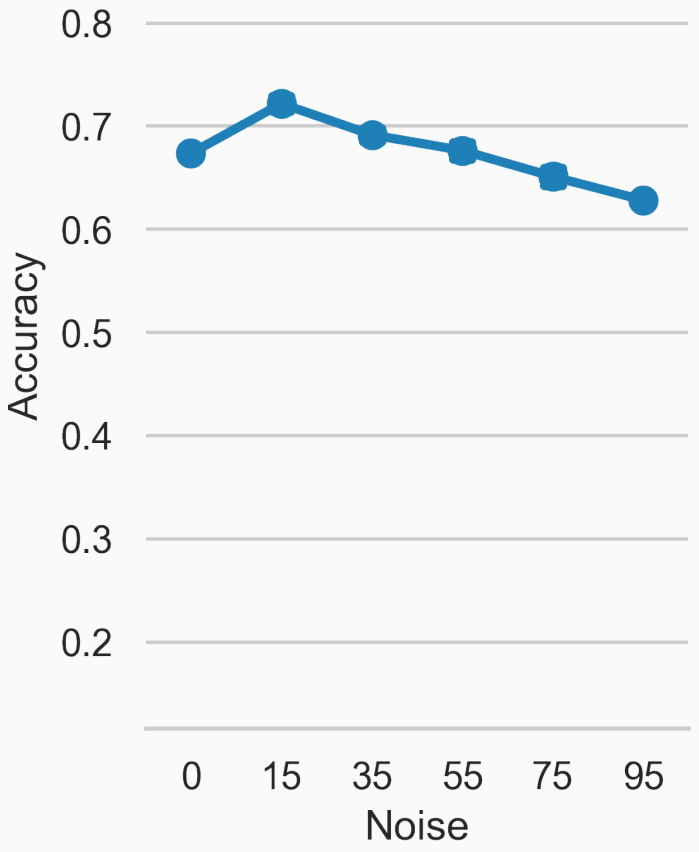
Monolingual BERTs/RoBERTas vs. mBERT vs. XLM-R

Noise levels: *0* / *15* / 35 / 55 / 75 / 95 %

- Cross-lingual performance drop (language-dependent)

- Cross-lingual performance drop (language-dependent)
- PLM choice matters

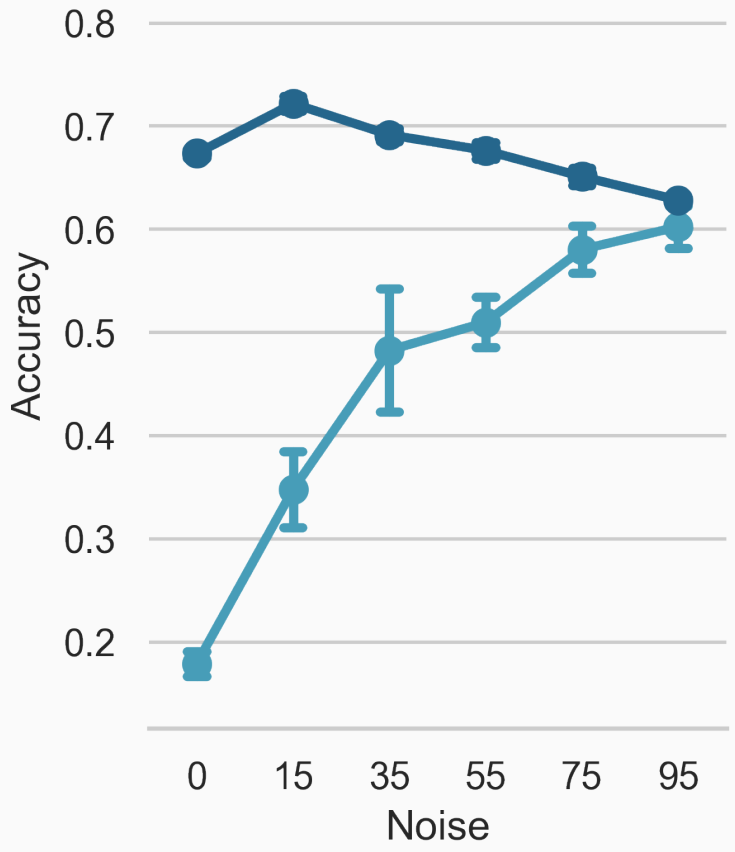
- One best noise level per set-up



Nynorsk → North Norwegian

(Monolingual PLMs)

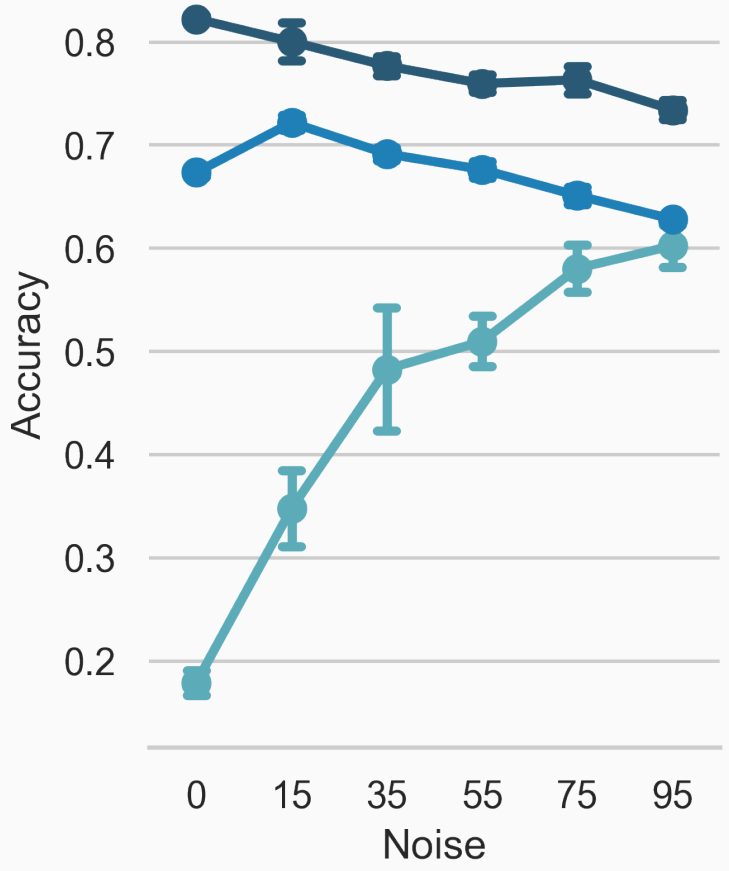
- One best noise level per set-up
- No universal best noise level



Nynorsk → North Norwegian
German → Low Saxon

(Monolingual PLMs)

- One best noise level per set-up
- No universal best noise level



Finnish → Savonian Finnish

Nynorsk → North Norwegian
German → Low Saxon

(Monolingual PLMs)

How does noise injection affect the subword tokenization differences between source and target data?

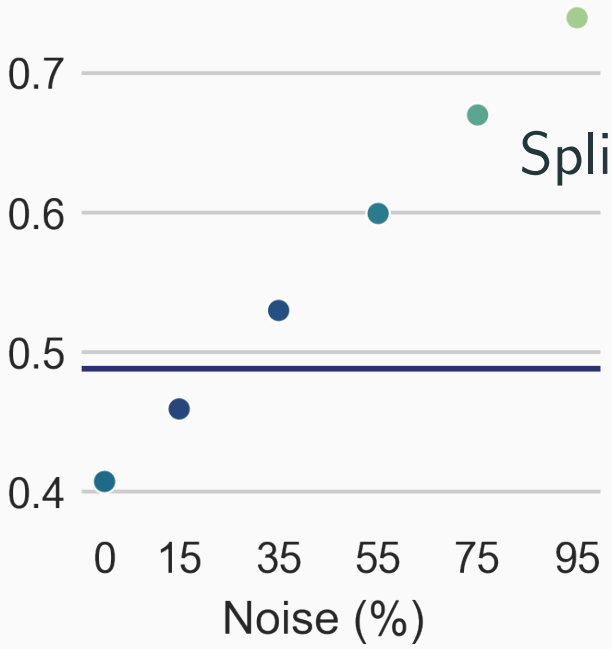
(*Seen subwords*, ...)

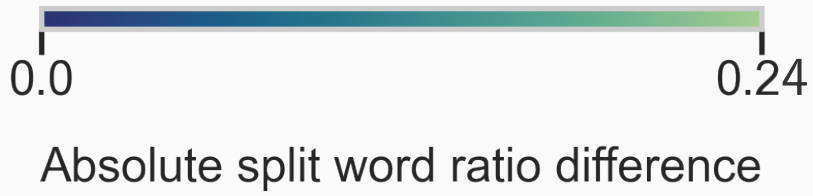
Split word ratio difference:

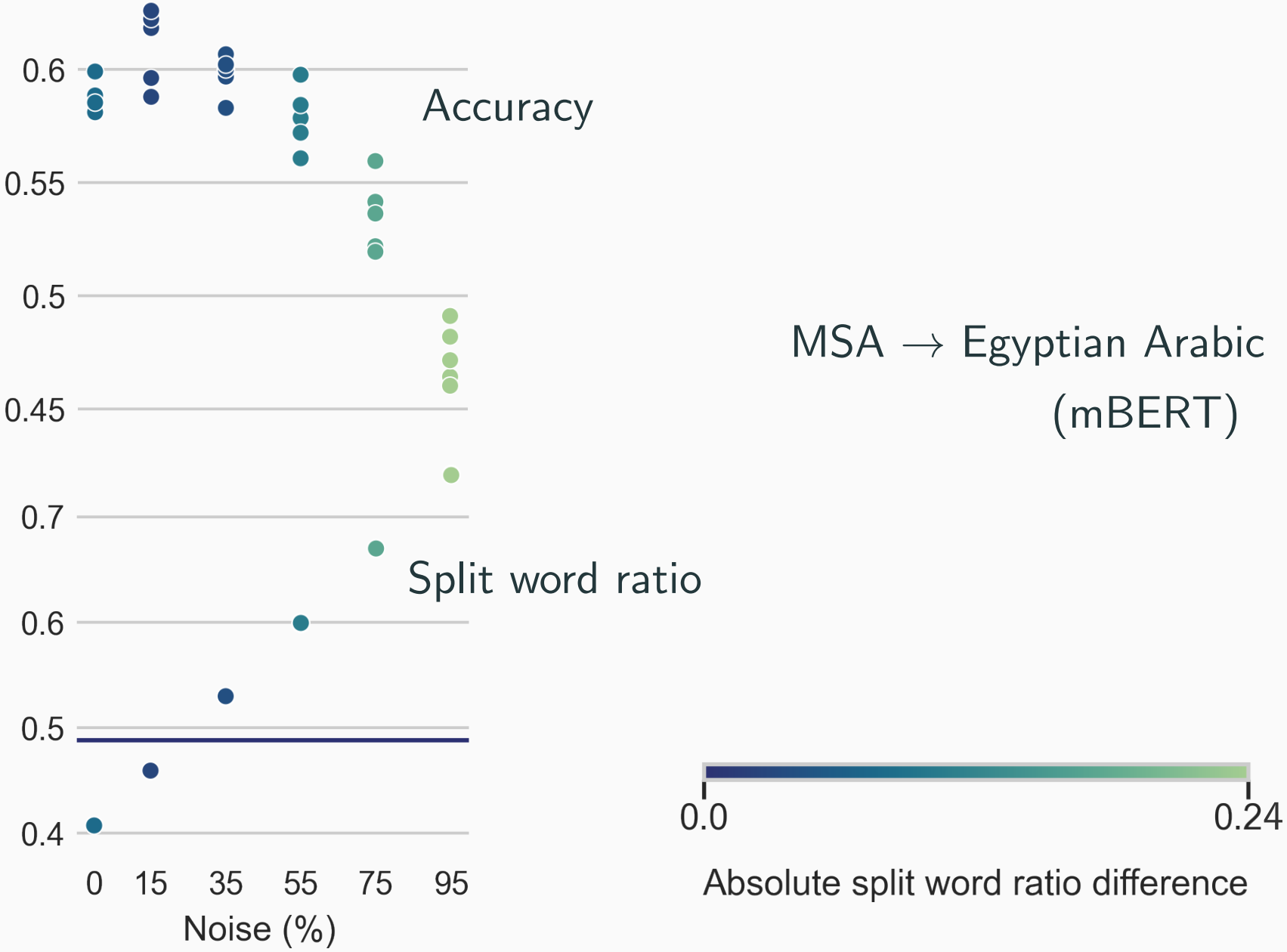
$$\left| \frac{\#\text{ words split in source data}}{\#\text{ words in source data}} - \frac{\#\text{ words split in target data}}{\#\text{ words in target data}} \right|$$

MSA → Egyptian Arabic

(mBERT)






Accuracy
0.6
0.55
0.5
0.45
0.7
Split word ratio
0.6
0.5
0.4
0
15
35
55
75
95
Noise (%)
MSA → Egyptian Arabic
(mBERT)
0.0
0.24
Absolute split word ratio difference

Recommendation:

- Don't want to tune the noise level as a hyperparameter? Cheaply calculate the *split word ratio differences* for different noise levels and pick the noise level with the lowest difference
- Otherwise, find the best noise level by starting low and increasing the noise until the dev accuracy starts dropping

See you at the
VarDial poster session!

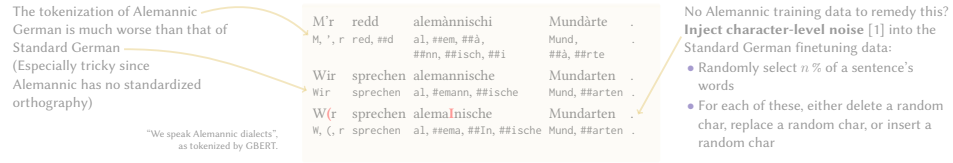
github.com/
mainlp/noisydialect

Does manipulating tokenization aid cross-lingual transfer?

A study on POS tagging for non-standardized languages

Verena Blaschke, Hinrich Schütze & Barbara Plank

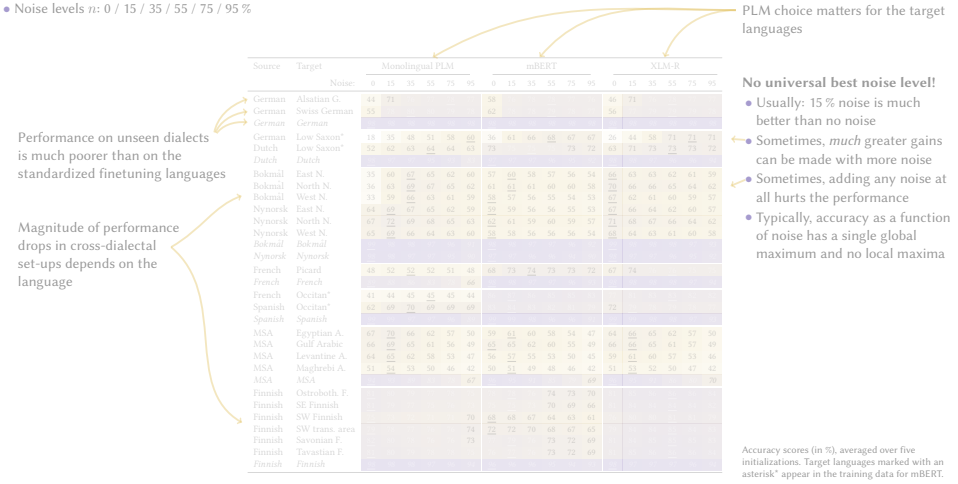
Center for Information and Language Processing, LMU Munich
Munich Center for Machine Learning
verena.blaschke@cis.lmu.de



To what extent does noise injection help?

- POS tagging via zero-shot transfer
- UPOS-tagged data from 3 language families (IE, Afro-Asiatic, Uralic)
- Monolingual BERTs/RoBERTas (source language), mBERT, XLM-R
- Noise levels n: 0 / 15 / 35 / 55 / 75 / 95 %

→ Code incl. tagset conversion scripts at github.com/mainlp/noisydialect



Does noise injection make the representation of the finetuning & target data more similar?
→ *Split word ratio difference* correlates best with accuracy: the (absolute) difference between the ratios of words split into subword tokens in the source and target data

The higher the noise level, the larger the *split word ratio* of the source data (→ rising sequences of dots)

Transfer from MSA to Egyptian Arabic. *Top:* Accuracy scores per language model and noise level. *Bottom:* Split word ratios per language model and noise level for the source data (dots) and the target data (dark blue lines).

Generally: the smaller the *split word ratio difference* between source and target (= the darker the dots), the higher the accuracy
- (Strong, but not perfect, correlations)

Recommendation:
- If you don't want to tune noise level as a hyperparameter: compute the *split word ratio differences* for different noise levels & **pick the noise level with the lowest difference**
- Otherwise, search for the best noise level: start low & **increase the noise level until the dev accuracy starts dropping**

Reference

[1] N. Aepli & R. Sennrich (2022). Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. *Findings of the Association for Computational Linguistics.*

This research is supported by the European Research Council (ERC) Consolidator Grant DIALECT 101043235 and in parts Advanced Grant NonSequeToR 740516.

 N. Aepli & R. Sennrich (2022). Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. *Findings of the Association for Computational Linguistics: ACL 2022*.

This research is supported by the European Research Council (ERC) Consolidator Grant DIALECT 101043235 and in parts Advanced Grant NonSequeToR 740516.