Dialect NLP

How (and why) to process non-standard language varieties

Verena Blaschke
MaiNLP lab, LMU Munich

Talk @ NLPnorth, ITU Copenhagen
September 23, 2024

Overview

? What: What do we mean with “dialect” in this context?

? Why: Why dialect NLP?

? How: Challenges and approaches

What do I mean with “dialects”?

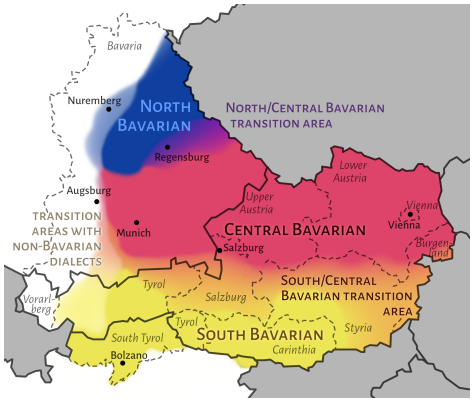
Many definitions in linguistics, NLP & everyday language

- Any language variety spoken by a(n especially geographically) distinct group of speakers
- National language varieties (e.g., Portuguese in Portugal vs. Brasil)
- Accents
- …

What do I mean with “dialects”?

Here:

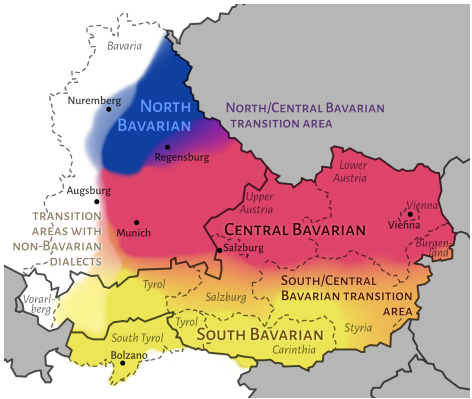
- Non-standardized
- Closely related to a standard language
- Differences from the std language in
  - Pronunciation (spelling)
  - Lexicon
  - Morphology and syntax



What do I mean with “dialects”?

Here:

- Non-standardized
- Closely related to a standard language
- Differences from the std language in
  - Pronunciation (spelling)
  - Lexicon
  - Morphology and syntax
- Often: continuum from standard to dialect
- Di(a)glossia; dialect speakers typically also write (and speak?) the standard



When do people use dialects?

- Spoken language
- Informal written contexts (text messages, social media)
- Some literature, poetry, wikis

Overview

❗ What: What do we mean with “dialect” in this context?

❓ Why: Why dialect NLP?

❓ How: Challenges and approaches

Why dialect NLP?

Why, given the fact that these speakers also speak a/the standard language?

- Linguistics: research language variation

Why dialect NLP?

Why, given the fact that these speakers also speak a/the standard language?

- Linguistics: research language variation
- ML research: sparse and heterogeneous data

Why dialect NLP?

Why, given the fact that these speakers also speak a/the standard language?

- Linguistics: research language variation
- ML research: sparse and heterogeneous data
- Applied reasons
  - Industry perspective (automatic subtitling, voice-based assistants in cars, analyzing social media data, ...)

Why dialect NLP?

Why, given the fact that these speakers also speak a/the standard language?

- Linguistics: research language variation
- ML research: sparse and heterogeneous data
- Applied reasons
  - Industry perspective (automatic subtitling, voice-based assistants in cars, analyzing social media data, ...)
  - Speaker perspective

"What do dialect speakers want? A survey of attitudes towards language technology for German dialects"
Blaschke, Purschke, Schütze & Plank (ACL 2024)

Motivation

Language technology (LT) – applied NLP systems

- Machine translation
- Chatbots
- Virtual assistants
- Transcription (ASR/STT)
- Speech synthesis (TTS)
- Spellcheckers
- Search engines
- ...

There is already some research on NLP for German dialects

Research questions

1. Which dialect technologies do respondents find especially useful?
2. Does this depend on…
    - whether the input or output is dialectal?
    - whether the LT works with speech or text data?
3. How does this reflect relevant sociolinguistic factors?

Questionnaire

- Target audience: speakers of German dialects + regional languages
- 3 weeks
- Word-of-mouth, social media, mailing lists, dialect/heritage societies

Questions

- Part I: about their dialect
- Part II: about attitudes towards LTs for their dialect

Questionnaire

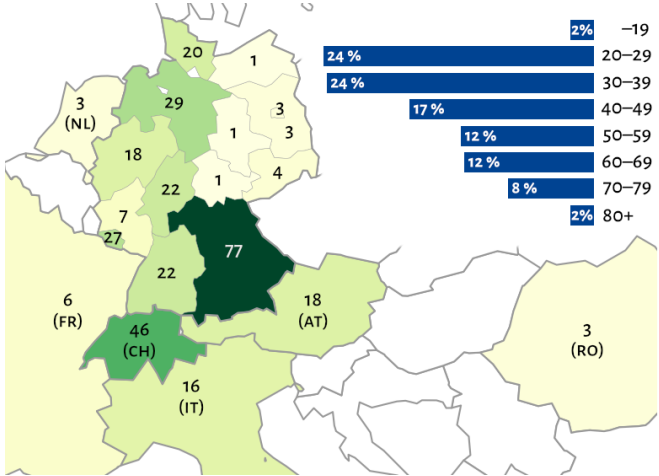
Speech-to-text systems transcribe spoken language. They are for instance used for automatically generating subtitles or in the context of dictation software.

Do you agree with the following statements?
There should be speech-to-text software...

- ...that transcribes audio recorded in my dialect as written Standard German.
- ..that transcribes audio recorded in my dialect as written dialect.

Dialect background and attitudes

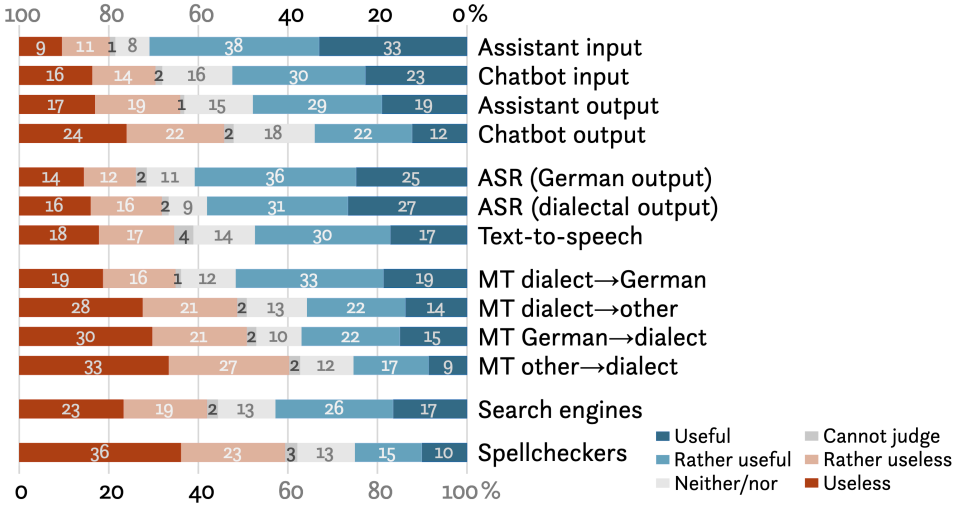
441 respondents – **327** of whom speak a German dialect and finished the questionnaire



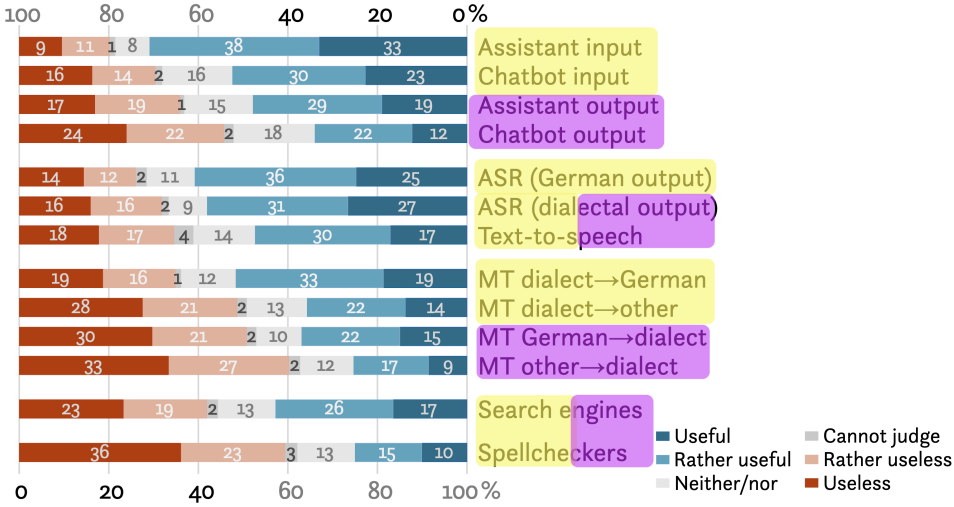
Dialect background and attitudes

- 52 % speak their dialect daily
- 65 % against standardized orthography
- 66 % write their dialect (even if rarely)
- 35 % are actively involved in dialect preservation
  - dialect preservation societies (13 %), teachers, dialectologists, …
  - speaking the dialect in public, with children
- 14 % already familiar with an LT for their dialect

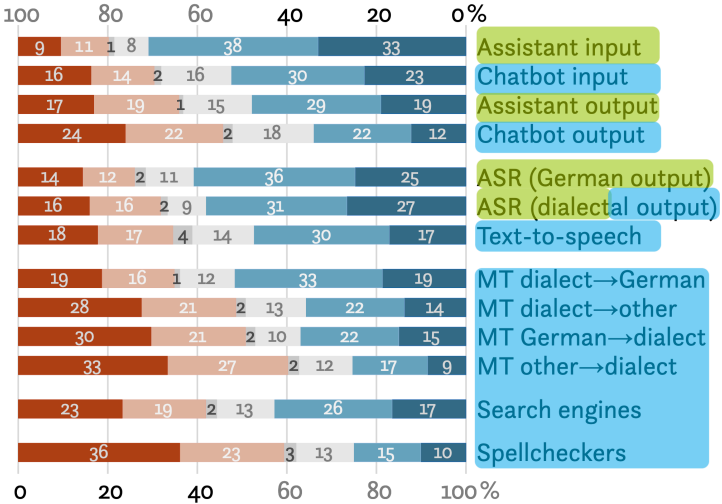
Which dialect LTs are deemed useful?



Dialect input vs. output?



Spoken vs. written dialect?



Correlated with opinion on standardized dialect orthographies

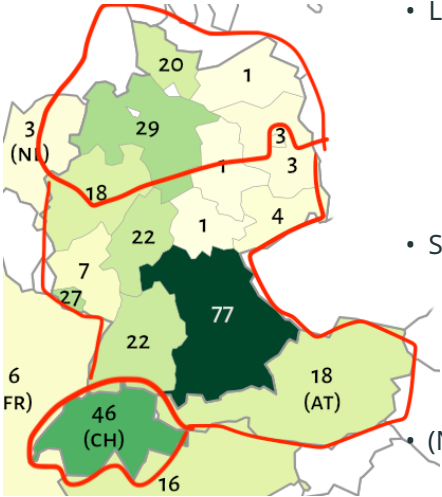
Do attitudes reflect sociolinguistic factors?

“Language activists” (involved in preservation)

- More in favour of dialect LTs involving text than non-activists
- ! Removing the activists’ responses has very little impact on the order of preferred LTs

Do attitudes reflect sociolinguistic factors?

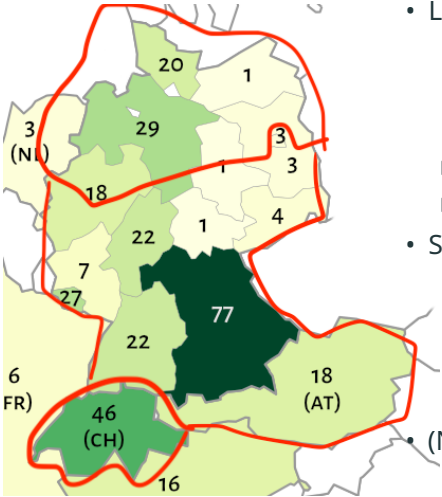
Three large dialect/language regions with different sociolinguistic realities



- Low Saxon
  - Recognized as language
  - Linguistically more distant
  - Preservation efforts
- Swiss German
  - High prestige
  - Spoken dialect, written Std German
- (Non-Low-Saxon) Germany + Austria
  - Partially replaced by regiolects

Do attitudes reflect sociolinguistic factors?

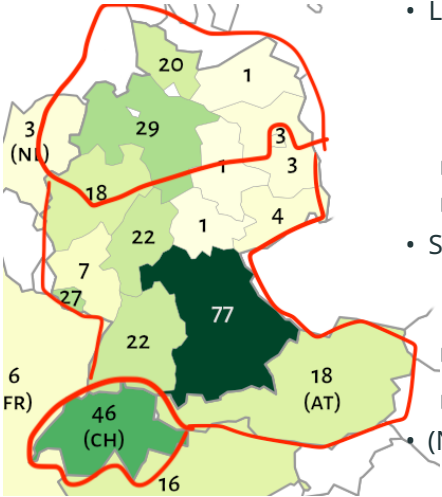
Three large dialect/language regions with different sociolinguistic realities



- Low Saxon
  - Recognized as language
  - Linguistically more distant
  - Preservation efforts
  - 👍 Dialect LTs in general
  - 👍 Orthographies + spellcheckers
- Swiss German
  - High prestige
  - Spoken dialect, written Std German
- (Non-Low-Saxon) Germany + Austria
  - Partially replaced by regiolects

Do attitudes reflect sociolinguistic factors?

Three large dialect/language regions with different sociolinguistic realities



- Low Saxon
  - Recognized as language
  - Linguistically more distant
  - Preservation efforts
  - 👍 Dialect LTs in general
  - 👍 Orthographies + spellcheckers
- Swiss German
  - High prestige
  - Spoken dialect, written Std German
  - 👎 Orthographies + spellcheckers
  - 👍 Spoken dialectal input
- (Non-Low-Saxon) Germany + Austria
  - Partially replaced by regiolects

Takeaways

- Interest in LTs processing dialectal input + speech-based LTs
- Speaker(group)s aren't monoliths!
- Sociolinguistic backgrounds are an important factor (but individual opinions exist too)
- Actively consider the wants & needs of the relevant speaker communities!

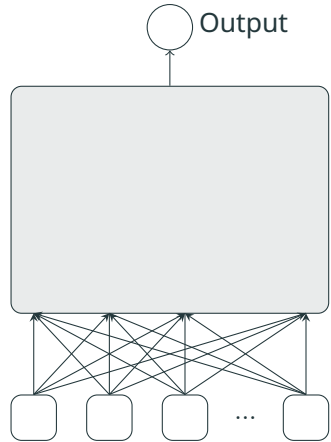
Overview

❗ What: What do we mean with “dialect” in this context?

❗ Why: Why dialect NLP?

❓ How: Challenges and approaches

Transfer learning problems



What tools + why?

Evaluating generated dialect output

Modelling non-standard data

Encoding non-standard data

`Lorem ipsum dolor sit`

Available dialect data

Data

Challenges regarding dialect corpora

- Availability
- Quality
- How do you represent data from unwritten/newly written languages?

"A survey of corpora for Germanic low-resource languages and dialects"
Blaschke, Schütze & Plank (NoDaLiDa 2023)

Which relevant datasets are out there?

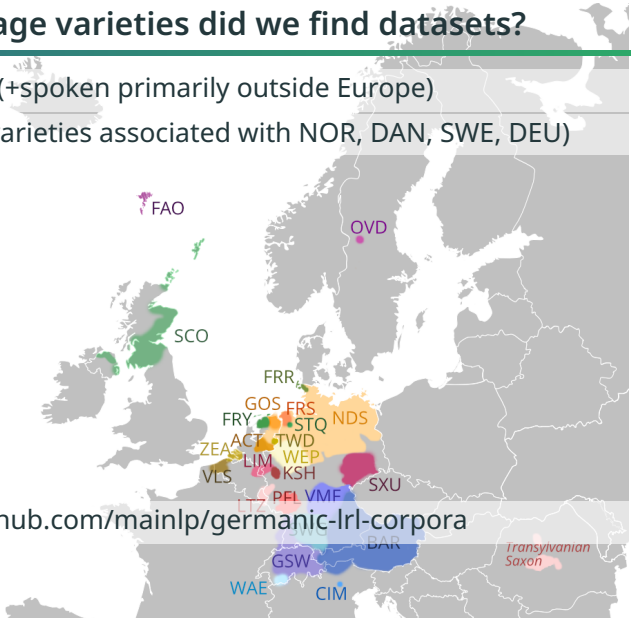
- Germanic dialects and low-resource languages
- (In)directly accessible for research (!!)
- Computer-friendly formats
- High-quality data (e.g., no OCR issues!)

$\rightarrow$100+ corpora!
github.com/mainlp/germanic-lrl-corpora

For which language varieties did we find datasets?

(+spoken primarily outside Europe)

(+non-std varieties associated with NOR, DAN, SWE, DEU)



github.com/mainlp/germanic-lrl-corpora

Annotations

What, if any, high-quality annotations do we find?

- Mostly: not annotated
- Morphosyntax (POS tags, dependencies, phrase structure)
- Geolocation, dialect group
- Rare: paraphrases, translations, sentiment, topics, slot and intent detection

Data quality: Uncurated data

Uncurated LRL data tend to be rather low quality
(wrong language, bad data cleaning)

“West Flemish” QED OPUS corpus

```
<w id="33.28">07,</w>
<w id="33.29">624&amp;amp;</w>
<w id="33.30">lt;</w>
<w id="33.31">br</w>
<w id="33.32">/</w>
<w id="33.33">&amp;amp;</w>
<w id="33.34">gt;</w>
<w id="33.35">Καλά</w>
<w id="33.36">,</w>
<w id="33.37">εντάξει</w>
<w id="33.38">.</w>
```

Data quality: Low-status varieties prone to parodies?

Shock an aw: US teenager wrote huge slice of Scots Wikipedia

Nineteen-year-old says he is 'devastated' after being accused of cultural vandalism

Das das alles kein Bairisch ist, würde ich nicht sagen, Holder. Aber man muss es verbessern. Vor allem muss man den Genitiv ersetzen und das Präteritum, und einige Wörter. Ich werde da mithelfen.

theguardian.com/uk-news/2020/aug/26/shock-an-aw-us-teenager-wrote-huge-slice-of-scots-wikipedia
bar.wikipedia.org/wiki/Dischkrian:Bundeswehr#Sprache

How to represent a (mostly spoken) language?

- Privacy issues concerning audio material
- Wildly different transcription styles & orthographies

How to represent a mostly spoken language?

- Normalized text (closely related standard language)

Etter litt godsnakk kom tre av kyrne … NB Tale

können sie ihre jugendzeit beschreiben ArchiMob

UD LSDC

How to represent a mostly spoken language?

- Normalized text (closely related standard language)
- Phone[m/t]ic transcriptions

Etter litt godsnakk kom tre av kyrne … NB Tale
’’’{t@4 l’’it g’’’u:snAkk k’’Om t4’’e: ’’A:v C’’y:n‘@ …

können sie ihre jugendzeit beschreiben ArchiMob
chönd sii iri jugendziit beschriibe

UD LSDC

How to represent a mostly spoken language?

- Normalized text (closely related standard language)
- Phone[m/t]ic transcriptions
- (More or less widely spread) orthographies

Etter litt godsnakk kom tre av kyrne … NB Tale
’’’{t@4 l’’it g’’’u:snAkk k’’Om t4’’e: ’’A:v C’’y:n‘@ …

können sie ihre jugendzeit beschreiben ArchiMob
chönd sii iri jugendziit beschriibe

Nu leyt em de böyse vynd disse nacht … UD LSDC

How to represent a mostly spoken language?

- Normalized text (closely related standard language)
- Phone[m/t]ic transcriptions
- (More or less widely spread) orthographies
- Ad-hoc spellings

Etter litt godsnakk kom tre av kyrne … NB Tale
””{t@4 l”it g””u:snAkk k”Om t4”e: ”A:v C”y:n‘@ …

können sie ihre jugendzeit beschreiben ArchiMob
chönd sii iri jugendziit beschriibe

Nu leyt em de böyse vynd disse nacht … UD LSDC
Nu leit em de baise Find düse Nacht …

How to represent a mostly spoken language?

- Normalized text (closely related standard language)
- Phone[m/t]ic transcriptions
- (More or less widely spread) orthographies
- Ad-hoc spellings

Etter litt godsnakk kom tre av kyrne … NB Tale
’’’{t@4 l’’it g’’’u:snAkk k’’Om t4’’e: ’’A:v C’’y:n‘@ …

können sie ihre jugendzeit beschreiben ArchiMob
chönd sii iri jugendziit beschriibe

Nu leyt em de böyse vynd disse nacht … UD LSDC
Nu leit em de baise Find düse Nacht …

→ If you build a tool that works for one type of written representation, it doesn’t necessarily work for the others too

Recommendations

... for *using* dialect corpora

- Check the quality!
- Is the written representation suitable for your purposes?
- Check whether your (pre-)training, dev, and test data are truly from independent sources (datasets overlap!)
- Look also outside traditional NLP venues

Recommendations

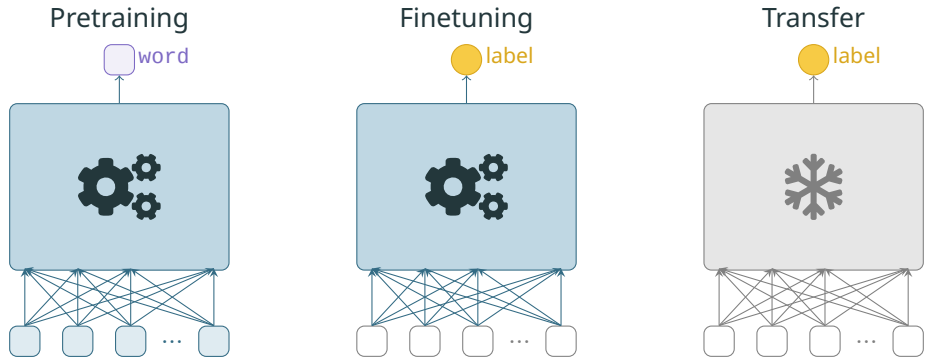
… for *using* dialect corpora

- Check the quality!
- Is the written representation suitable for your purposes?
- Check whether your (pre-)training, dev, and test data are truly from independent sources (datasets overlap!)
- Look also outside traditional NLP venues

… for *creating* dialect corpora

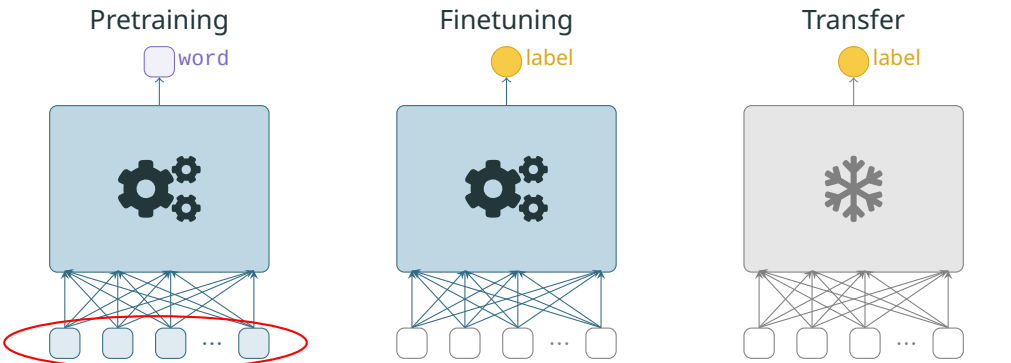
- Document the transcription guidelines / orthographies
- Use archives geared towards long-term storage (CLARIN, LRE Map, Zenodo)
- Share basic metadata like corpus size, data sources, annotation procedure + license information

Pretrain – finetune – transfer



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut

Pretrain – finetune – transfer



Non-standard orthographies + tokenization

Subword tokenization with GBERT

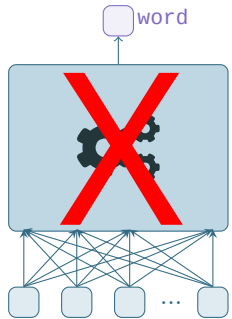
Die	Lammer	hat	ein	recht		sauberes	Wasser
`Die`	`Lamm` `-er`	`hat`	`ein`	`recht`		`sauber` `-es`	`Wasser`

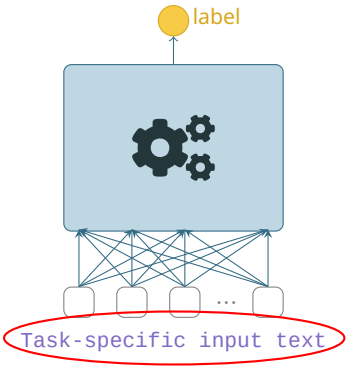
D’	Lomma	hod	a	rechd	a	sauwas	Wossa
`D` `'`	`Lom` `-ma`	`ho` `-d`	`a`	`rech` `-d`	`a`	`sau` `-was`	`Wo` `-ssa`

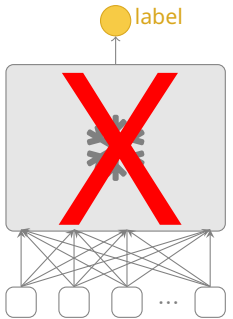
“The Lammer (river) has fairly clean water”

Sentence via `bar.wikipedia.org/wiki/Låmma`
GBERT: Chan/Schweter/Möller, COLING 2020, “German’s Next Language Model”

How to (easily) make models more robust?







Character-level “noise”

Die	Lammer	hat	ein	recht		sauberes	Wasser
Die	Lamm –er	hat	ein	recht		sauber –es	Wasser
D’	Lomma	hod	a	rechd	a	sauwas	Wossa
D '	Lom –ma	ho –d	a	rech –d	a	sau –was	Wo –ssa
D(e	Lammer	hat	ein	recht		saubenes	Wasser
D (e	Lamm –er	hat	ein	recht		sau –ben –es	Wasser

Inject 15% of words with “noise”

Aepli/Sennrich, ACL Findings 2022 “Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise”

Character-level “noise”

D (e Lamm -er hat ein recht sau -ben -es Wasser

Inject 15% of words with “noise” (Aepli/Sennrich 2022)

How many words should we modify this way?

“Does manipulating tokenization aid cross-lingual transfer?
A study on POS tagging for non-standardized languages”
Blaschke, Schütze & Plank (VarDial 2024)

Dialect POS tagging

- Part-of-speech tagging
- Transfer from closely related standard languages to…
  - 3 dialects / regional languages of Germany
  - 3 Norwegian dialects
  - 2 regional languages of France
  - 6 Finnish dialects
  - 4 Arabic varieties

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

- Noise: Modify {0, 15, 35, 55, 75, 95}% of words

Dialect POS tagging

- Part-of-speech tagging
- Transfer from closely related standard languages to…
  - 3 dialects / regional languages of Germany
  - 3 Norwegian dialects
  - 2 regional languages of France
  - 6 Finnish dialects
  - 4 Arabic varieties

  $\rightarrow$ Consistent performance drops (standard/dialect)
- Monolingual BERTs/RoBERTas vs. XLM-R vs. mBERT

- Noise: Modify {0, 15, 35, 55, 75, 95}% of words

Dialect POS tagging

- Part-of-speech tagging
- Transfer from closely related standard languages to…
  - 3 dialects / regional languages of Germany
  - 3 Norwegian dialects
  - 2 regional languages of France
  - 6 Finnish dialects
  - 4 Arabic varieties

  $\rightarrow$ Consistent performance drops (standard/dialect)
- Monolingual BERTs/RoBERTas vs. XLM-R vs. mBERT

  $\rightarrow$ Optimal choice varies across languages
- Noise: Modify {0, 15, 35, 55, 75, 95}% of words

Dialect POS tagging

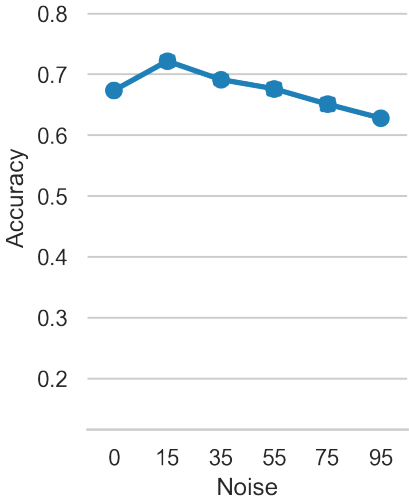
- Part-of-speech tagging
- Transfer from closely related standard languages to…
  - 3 dialects / regional languages of Germany
  - 3 Norwegian dialects
  - 2 regional languages of France
  - 6 Finnish dialects
  - 4 Arabic varieties

  $\rightarrow$ Consistent performance drops (standard/dialect)
- Monolingual BERTs/RoBERTas vs. XLM-R vs. mBERT

  $\rightarrow$ Optimal choice varies across languages
- Noise: Modify {0, 15, 35, 55, 75, 95}% of words

  $\rightarrow$ Optimal choice varies across languages/models

How much noise to add?

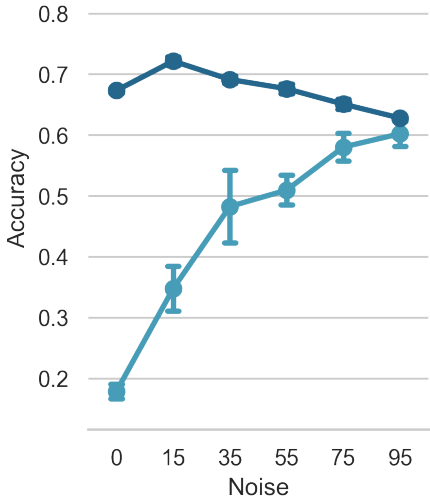


Nynorsk → North Norwegian

(Monolingual PLMs)

Blaschke/Schütze/Plank, VarDial 2023, “Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages”

How much noise to add?

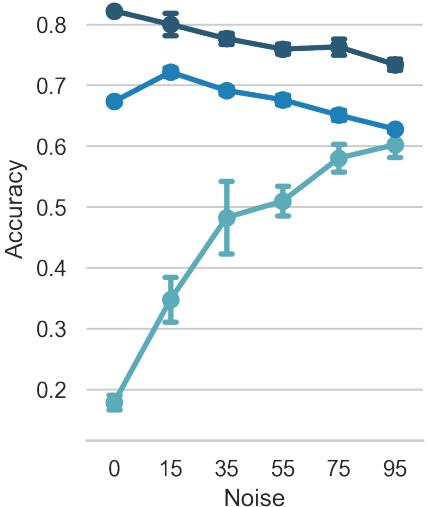


Nynorsk → North Norwegian
German → Low Saxon

(Monolingual PLMs)

Blaschke/Schütze/Plank, VarDial 2023, “Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages”

How much noise to add?



Finnish → Savonian Finnish

Nynorsk → North Norwegian
German → Low Saxon

(Monolingual PLMs)

Blaschke/Schütze/Plank, VarDial 2023, “Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages”

What explains this?

The more similar the word-splitting rates are, the better the results!
Spearman's rank correlation of > 0.8 for most models/languages!

Die	Lammer	hat	ein	recht		sauberes		Wasser
Die	Lamm –er	hat	ein	recht		sauber –es		Wasser
D'	Lomma	hod	a	rechd	a	sauwas		Wossa
D '	Lom –ma	ho –d	a	rech –d	a	sau –was		Wo –ssa
D(e	Lammer	hat	ein	recht		saubenes		Wasser
D (e	Lamm –er	hat	ein	recht		sau –ben –es		Wasser

Recommendation

- Don’t want to tune the noise level as a hyperparameter?
  $\rightarrow$ Cheaply compare the *split word ratios* 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

Summary

❗ What: What do we mean with “dialect” in this context?

❗ Why: Why dialect NLP?

- Also consider the speaker community!
- Especially popular: dialectal input; speech

❗ How: Challenges and approaches

- Finding data
- Tokenization of non-standardized text