Beyond “noisy” text:
How (and why) to process dialect data

Verena Blaschke
LMU Munich & MCML

Workshop on noisy and user-generated text (W-NUT)
NAACL
03 May 2025

Munich Center for Machine Learning

Dialect NLP

Linking two lines of research

- NLP robust to noisy inputs
- NLP for underserved language communities

? What: Dialects and language variation

? How: Dialect NLP – (Some) challenges and methods

? For whom: Speaker perspectives

- Recommendations
- Are dialects “noisy” language data?

Dialect NLP

Linking two lines of research

- NLP robust to noisy inputs
- NLP for underserved language communities

? **What: Dialects and language variation**

? How: Dialect NLP – (Some) challenges and methods

? For whom: Speaker perspectives

- Recommendations
- Are dialects “noisy” language data?

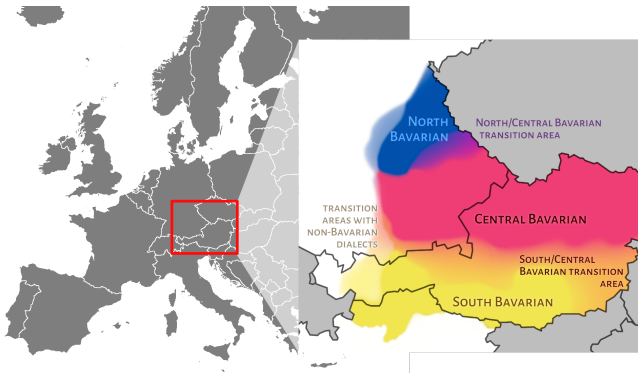
What do I mean with “dialects”?

Many definitions in linguistics, NLP & everyday language

- Any language variety spoken by a (geographically) distinct group of speakers
- National language varieties
- Accents
- …

What do I mean with “dialects”?

- **Non-standardized**
- **Closely related to a standard language**
- Often: continuum standard – dialect
- Often: subdialects



Linguistic differences

Differences from the standard language

- Pronunciation ($\rightarrow$ spelling)

[German]	**Sie**	**haben**	**keine**	Beine	
[Bavarian]	**Se**	**hom**	**koane**	Haxn	ned
	They	*have*	*no*	*legs*	*not*

"They [=fish] have no legs"

Linguistic differences

Differences from the standard language

- Pronunciation ($\rightarrow$ spelling)
- Lexicon

[German]	Sie	haben	keine	**Beine**	
[Bavarian]	Se	hom	koane	**Haxn**	ned
	They	*have*	*no*	*legs*	*not*

"They [=fish] have no legs"

Linguistic differences

Differences from the standard language

- Pronunciation ($\rightarrow$ spelling)
- Lexicon
- Grammar: morphology, syntax

[German]	Sie	haben	**keine**	Beine	
[Bavarian]	Se	hom	**koane**	Haxn	**ned**
	They	*have*	**no**	*legs*	**not**

"They [=fish] have no legs"

Linguistic differences

Differences from the standard language

- Pronunciation ($\rightarrow$ spelling)
- Lexicon
- Grammar: morphology, syntax
- Usage context
    - Dialect speakers typically also write (+ speak?) the standard

[German]	Sie	haben	keine	Beine	
[Bavarian]	Se	hom	koane	Haxn	ned
	They	*have*	*no*	*legs*	*not*

“They [=fish] have no legs”

Linguistic differences

Differences from the standard language

- Pronunciation ($\rightarrow$ spelling)
- Lexicon
- Grammar: morphology, syntax
- Usage context
    - Dialect speakers typically also write (+ speak?) the standard

[German]	Sie	haben	keine	Beine	
[Bavarian]	**Se**	**hom**	koane	Haxn	**ned**
	They	*have*	*no*	*legs*	*not*
	De	**ham**	koane	Haxn	_
	Dei	**hobm**	koane	Haxn	_

"They [=fish] have no legs"

When do people use dialects?

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

Losses da gud gehn in Albuquerque und viel Schbass bei de Konferenz!

Why dialect NLP?

- Annotate data for linguists, research variation
- Sparse & heterogeneous data for ML
- Downstream: systems for more robustly processing non-standard data
- (and more!)

Data

Challenges regarding dialect corpora

- Availability
- Quality
- Written representations

A Survey of Corpora for Germanic Low-Resource Languages and Dialects

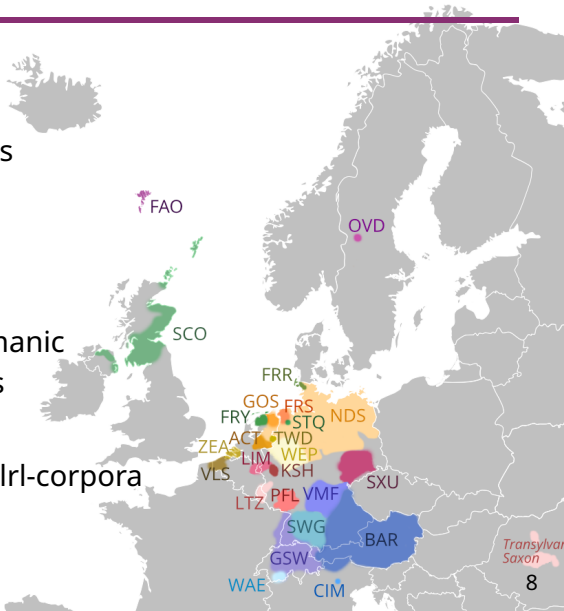
Verena Blaschke **Hinrich Schütze** **Barbara Plank**

Datasets for Germanic low-resource varieties

- Accessible for research
- Computer-friendly formats
- Annotated + unannotated
- High-quality data
  (e.g., no OCR issues!)

➔ 100+ datasets for 35 Germanic dialects + small languages

github.com/mainlp/germanic-lrl-corpora



Annotations

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

- Morphosyntax (POS tags, dependencies, phrase structure)
- Geolocation, dialect group
- Paraphrases, translations, sentiment, topics, slot and intent detection
    - Rare, but getting more popular
- Mostly: not annotated
    - … and sometimes uncurated

Data quality: Uncurated data

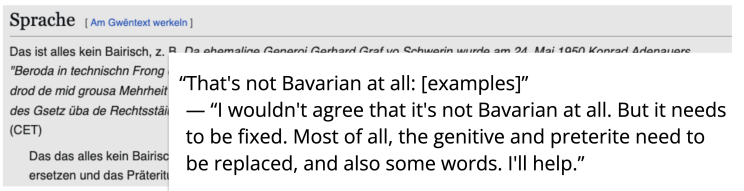
Uncurated LRL data tend to be of rather low quality
– wrong language, bad data cleaning
(Kreutzer+, TACL 2022; Abadji+, LREC 2022)

OSCAR corpus (fixed subsequently)

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



Brooks/Hern, The Guardian, 2020
bar.wikipedia.org/wiki/Dischkrian:Bundeswehr

How to represent a primarily spoken language?

- Normalized text (closely related standard language)

Etter litt godsnakk kom tre av kyrne …	NB Tale
[After some coaxing, three of the cows came …]	*Norwegian*
können sie ihre jugendzeit beschreiben	ArchiMob
[Can you describe your youth?]	*Swiss German*

How to represent a primarily spoken language?

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

””{t@4 l”it g””u:snAkk k”Om t4”e: ”A:v C”y:n‘@ …	
Etter litt godsnakk kom tre av kyrne …	NB Tale
[After some coaxing, three of the cows came …]	*Norwegian*
chönd sii iri jugendziit beschriibe	
können sie ihre jugendzeit beschreiben	ArchiMob
[Can you describe your youth?]	*Swiss German*

How to represent a primarily spoken language?

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

Nu leyt em de böyse vynd disse nacht … UD LSDC
[Now, this night, the wicked enemy let them…] *Low Saxon*

How to represent a primarily spoken language?

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

Nu leit em de baise Find düse Nacht …
Nu leyt em de böyse vynd disse nacht … UD LSDC
[Now, this night, the wicked enemy let them…] *Low Saxon*

How to represent a primarily spoken language?

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

$\rightarrow$ A tool that works for one type of written representation doesn't necessarily work for the others too

Recommendations

Blaschke, Schütze & Plank (NoDaLiDa 2023)

"A survey of corpora for Germanic low-resource languages and dialects"

… for *using* dialect corpora

- Check the quality!
- Suitable written representation for your purposes?
- Data scarcity – overlaps between (pre-)training, dev, test data?
- Data outside traditional NLP venues

… for *creating* dialect corpora

- Document the transcription guidelines / orthographies
- Share metadata like corpus size, data sources, license
- Archives for long-term storage (CLARIN, LRE Map, Zenodo)

Overview

! What: Dialects and language variation

? **How: Dialect NLP – (Some) challenges and methods**

? For whom: Speaker perspectives

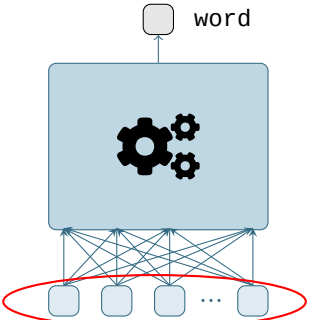
Linguistic differences

Differences from the standard language in

- **Pronunciation ($\rightarrow$ spelling)**
- Lexicon
- Morphology
- Syntax
- Usage context

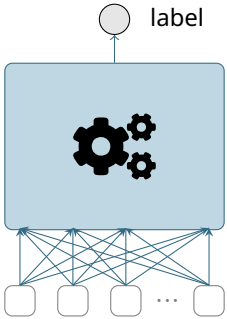
Cross-dialectal transfer

✗ Pretraining



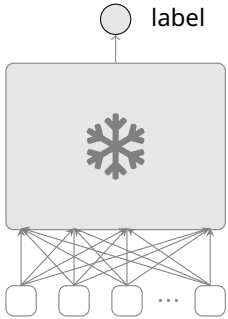
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum 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,

✗ Finetuning



Task-specific input text

✓ Transfer



Input text in related dialect

Non-standard orthographies + tokenization

Subword tokenization with GBERT

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

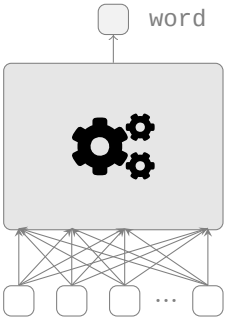
D’	Lomma	hod	a	rechd	a	sauwas	Wossa
`D` `'`	`Lom` `-ma`	`ho` `-d`	`a`	`rech` `-d`	`a`	`sau` `-was`	`Wo` `-ssa`
The	*Lammer*	*has*	*a*	*fairly*	*a*	*clean*	*water*

“The Lammer (river) has fairly clean water”

Sentence via `bar.wikipedia.org/wiki/Låmma`
GBERT: Chan+, COLING 2020, “German’s next language model”

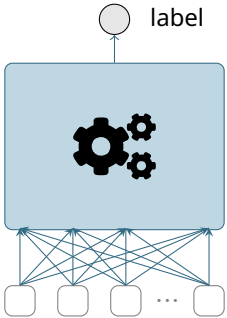
More robust input representations?

Diff. architecture

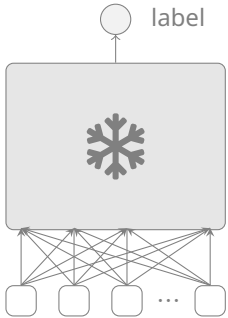


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum 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,

Change finetuning



Task-specific input text



Input text in related dialect

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 finetuning words with “noise”

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

Character-level “noise”

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 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

Verena Blaschke **Hinrich Schütze** **Barbara Plank**

Noise levels

0 %: The WNUT workshop focuses on core NLP tasks over user-generated text, such as that found on social media, web forums, online reviews, digital health records, or language learner essays.

15 %: The WNUT workshop focus**q**es on co**f**e NLP tasks over user-generated text, such as that found on social media, web for_ms, online reviews, digital health re**t**cords, or language learner essays.

...

95 %: _he WNU**P** wo**c**kshop focus_s on**q** c**i**ore _LP tas**o**s o**;**ver user-genera**f**ted text, su**k**ch as**v** t_at **p**found o_ soc**h**al medi**r**, w**a**eb forum**g**s, o_line rev**r**ews, di**o**ital heal**a**h records, **l**or langua**'**ge lea**m**ner ess**s**ays.

Dialect POS tagging

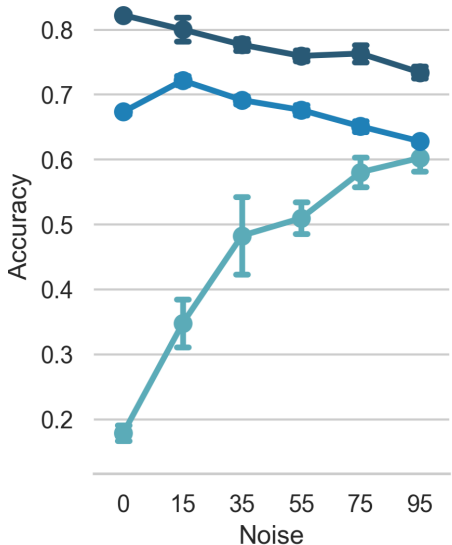
- Part-of-speech tagging
- Transfer from closely related standard languages to...
  - 3 dialects / regional languages of Germany
  - 3 Norwegian dialects
  - 2 regional languages of France
  - 6 Finnish dialects
  - 4 Arabic varieties

  → *Consistent performance drops (standard/dialect)*
- Monolingual BERTs/RoBERTas vs. XLM-R vs. mBERT

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

  → *Optimal choice varies across languages/models*

How much noise to add?



Finnish → Savonian Finnish

Nynorsk → North Norwegian
German → Low Saxon

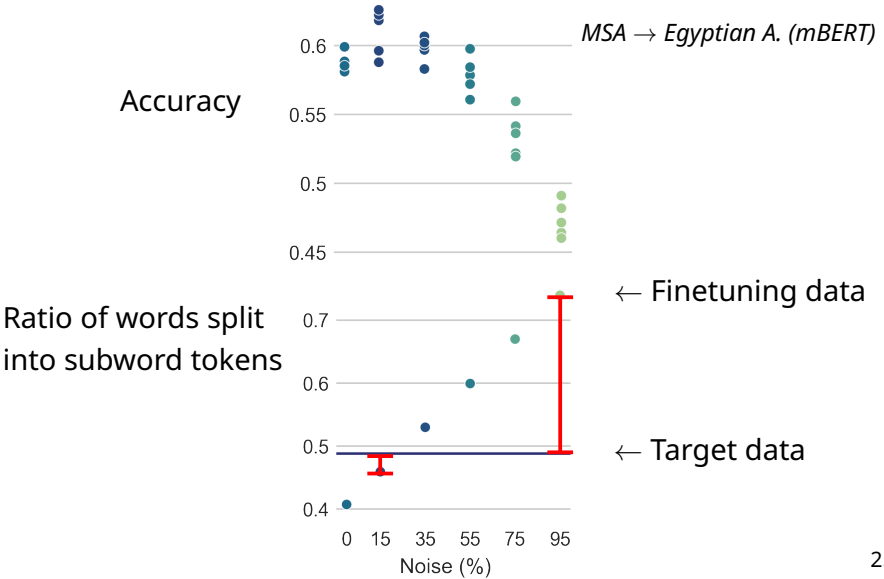
(Monolingual PLMs)

What explains this?

The more similar the word-splitting rates are, the better the results!

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

Noise injection & subword tokenization



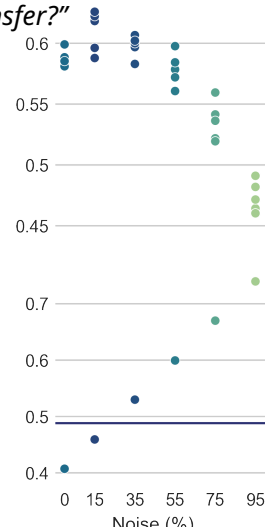
Recommendation

Blaschke, Schütze & Plank (VarDial @ EACL 2023)

"Does manipulating tokenization aid cross-lingual transfer?"

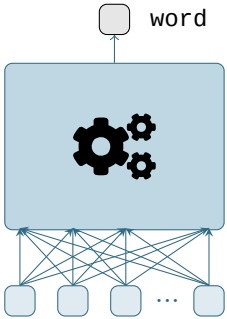
Spearman's rank correlation of > 0.8 for most models/languages (often > 0.95)!

- Don't want to tune noise level as a hyperparameter?
  $\rightarrow$ Compare the *split word ratios* for different noise levels
  + pick the noise level with the smallest difference
- Otherwise: start low + increase noise until dev accuracy drops



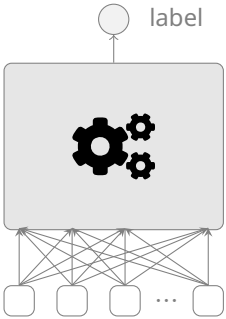
More robust input representations?

Diff. architecture

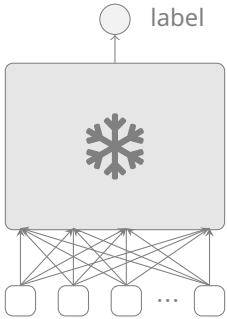


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum 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,

Change finetuning



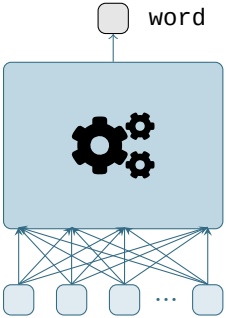
Task-specific input text



Input text in related dialect

More robust input representations?

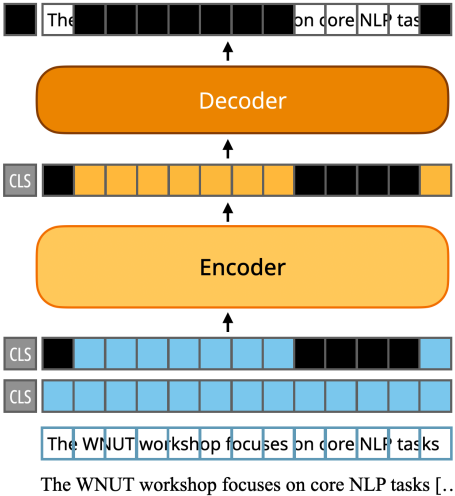
Diff. architecture



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,

"Language modelling with pixels"
Rust, Lotz, Bugliarello, Salesky,
de Lhoneux & Elliott (ICLR 2023)

Pixel models (Rust+, 2023) – pretraining



Decode masked pixels

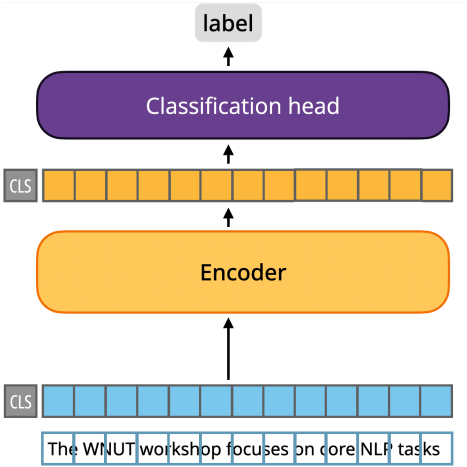
Encode

Mask spans

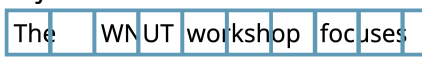
Projection

Render text as image

Pixel models (Rust+, 2023) – finetuning



Text rendering can be adjusted for word-level tasks



Pixel models – robustness

(English) Pixel generally more robust against orthographic attacks than BERT

Attack	Sentence
NONE	Penguins are designed to be streamlined
CONFUSABLE	Peπguתns are designed to be streamlined
SHUFFLE (INNER)	Pegnuins are dnesiged to be sieatrnmled
SHUFFLE (FULL)	ngePnius rae dsgednei to be etimaslernd
DISEMVOWEL	Pngns r dsgnd to be strmlnd
INTRUDE	Pe‘nguins a{re d)esigned t;o b*e stre<amlined
KEYBOARD TYPO	Penguinz xre dwsigned ro ne streamllned
NATURAL NOISE	Penguijs ard design4d ti bd streamlinfd
TRUNCATE	Penguin are designe to be streamline
SEGMENTATION	Penguinsaredesignedtobestreamlined
PHONETIC	Pengwains’s ar dhiseind te be storimlignd

Table via *Rust+ “Language modelling with pixels” (ICLR 2023)*

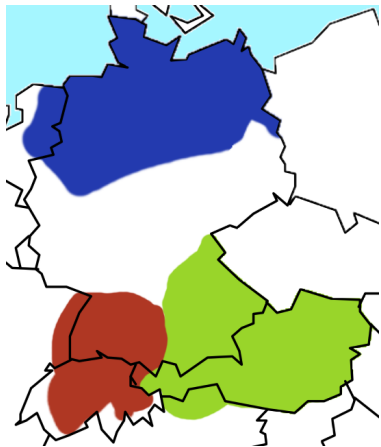
Pixel models – robustness

Evaluating Pixel Language Models on Non-Standardized Languages

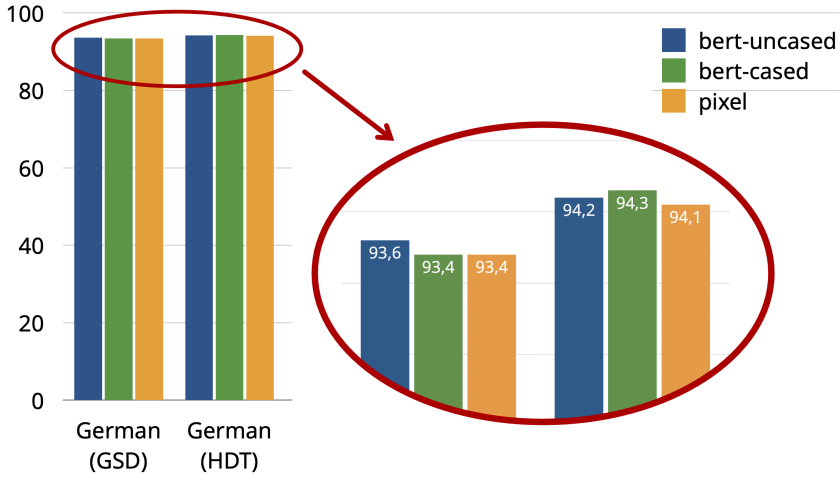
Alberto Muñoz-Ortiz **Verena Blaschke** **Barbara Plank**

German Pixel experiments

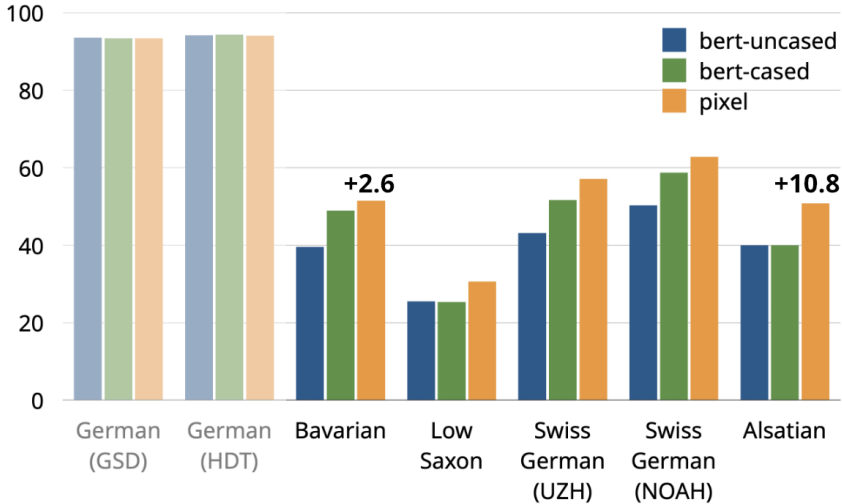
- German Pixel model (new!)
  - Same training data as a German BERT model
- Finetune on German, evaluate on dialects/regional languages
- 2 grammatical tasks: POS tagging, parsing
- 2 semantic tasks: intent classification (easy), topic classification (harder)



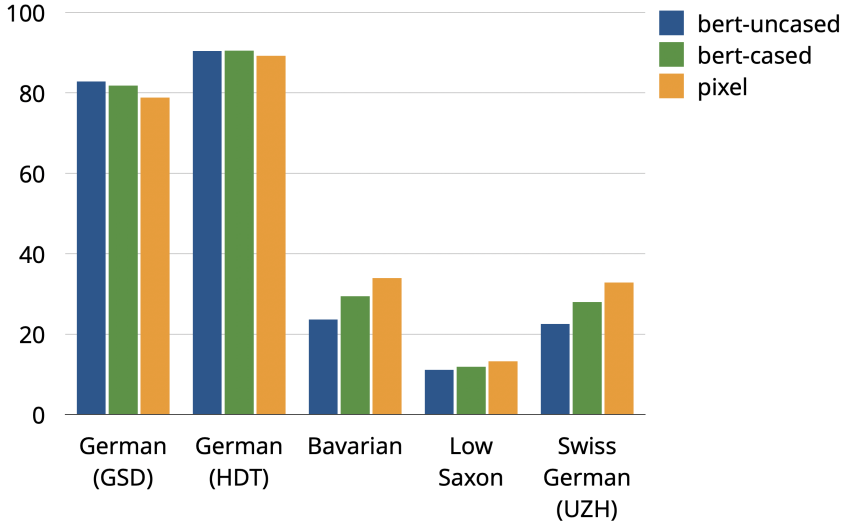
German Pixel: POS tagging (accuracy)



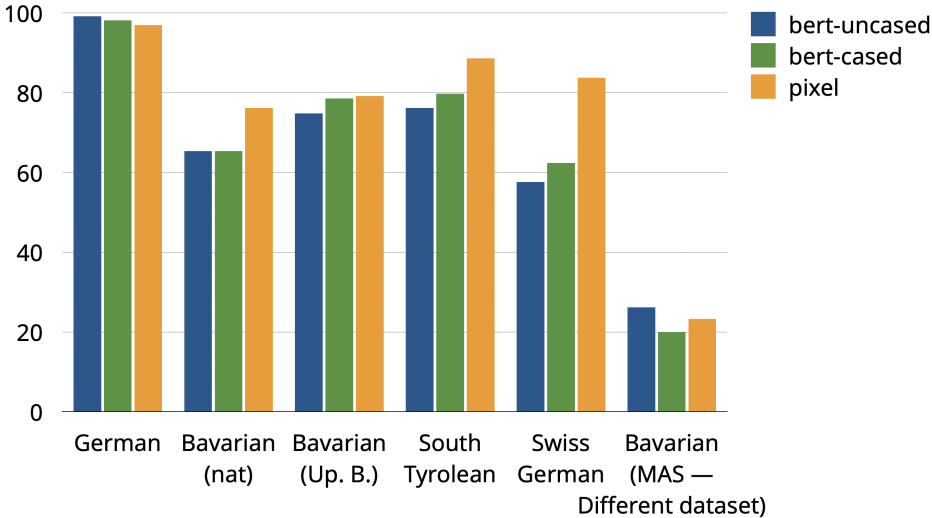
German Pixel: POS tagging (accuracy)



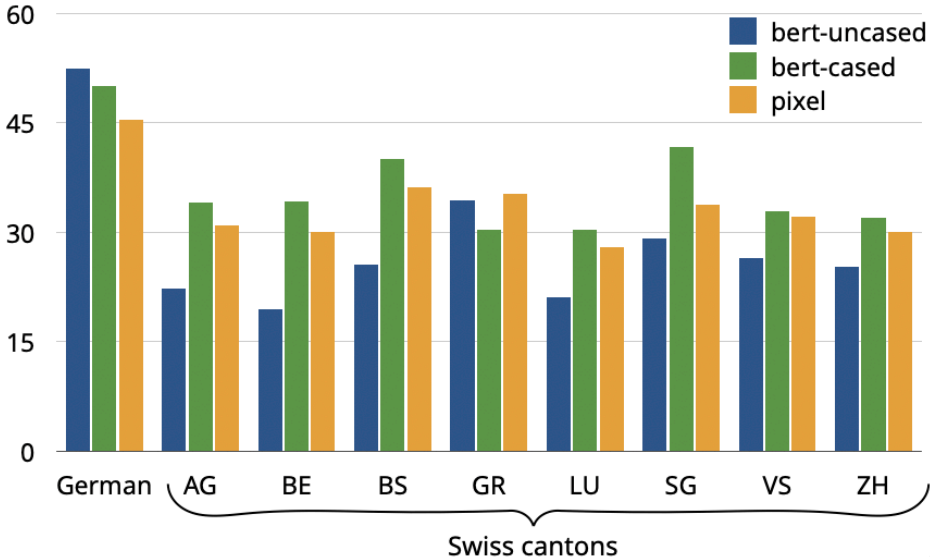
German Pixel: Parsing (LAS)



German Pixel: Intent classification (accuracy)



German Pixel: Topic classification (accuracy)



Pixel: Trade-off

Muñoz-Ortiz, Blaschke & Plank (COLING 2025)

“Evaluating pixel language models on non-standardized languages”

- More compute needed
- On par with or worse than BERT in monolingual settings (+ where std language performance is bad)
- Cross-dialectal settings / settings with less predictable spelling might be the place to shine

→ Worthwhile for other “noisy” settings?

Overview

- ! What: Dialects and language variation
- ! How: Dialect NLP – (Some) challenges and methods
- ? **For whom: Speaker perspectives**

Linguistic differences

Differences from the standard language in

- Pronunciation ($\rightarrow$ spelling)
- Lexicon
- Morphology
- Syntax
- **Usage context**

Why dialect NLP?

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

- Linguistics
- ML research
- Applied reasons
  - Industry perspective
  - Speaker perspective

What Do Dialect Speakers Want?
A Survey of Attitudes Towards Language Technology for German Dialects

Verena Blaschke Christoph Purschke Hinrich Schütze Barbara Plank

Motivation

Language technology (LT) – applied NLP systems

- Machine translation (MT)
- (Written) chatbots
- (Spoken) virtual assistants
- Transcription (ASR)
- Speech synthesis (TTS)
- Search engines
- Spellcheckers

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

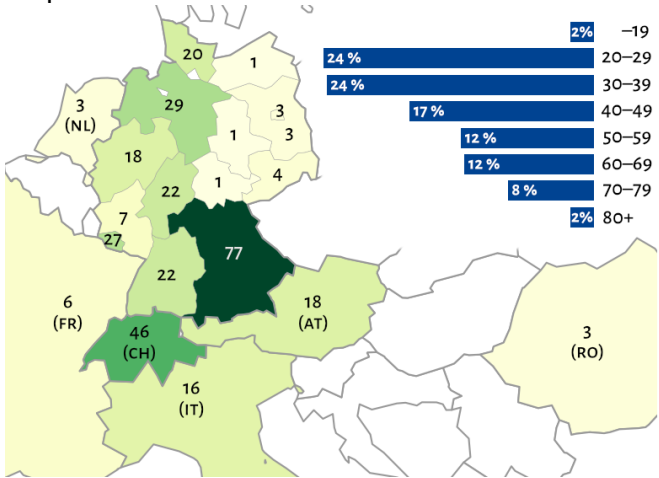
Speech-to-text systems transcribe spoken language. They are for instance used for automatically generating subtitles or in the context of dictation software.

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

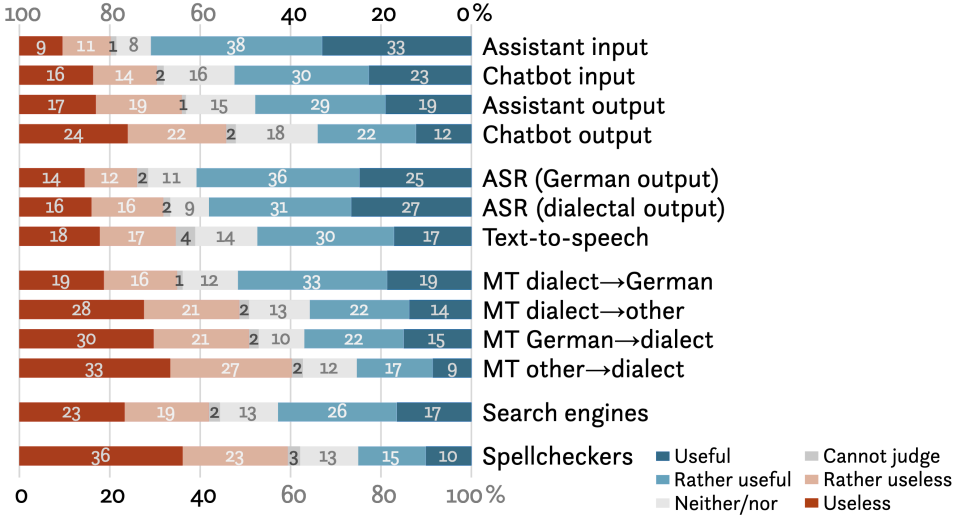
441 respondents – **327** of whom speak a German dialect and finished the questionnaire



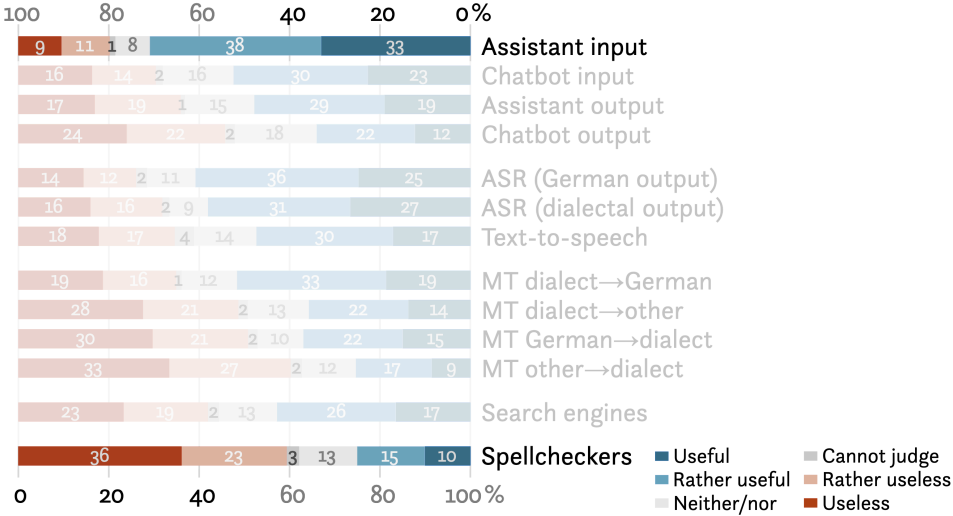
Dialect background and attitudes

- 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

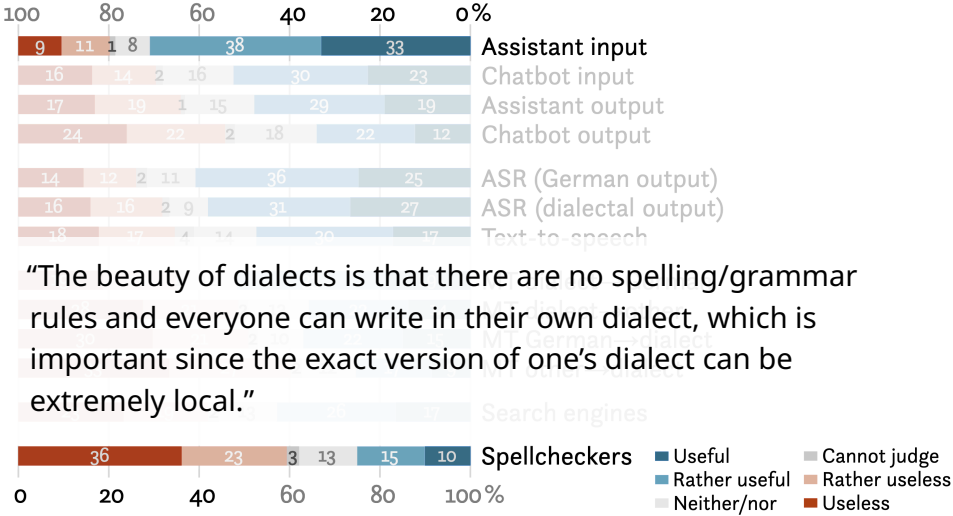
Which dialect LTs are deemed useful?



Which dialect LTs are deemed useful?

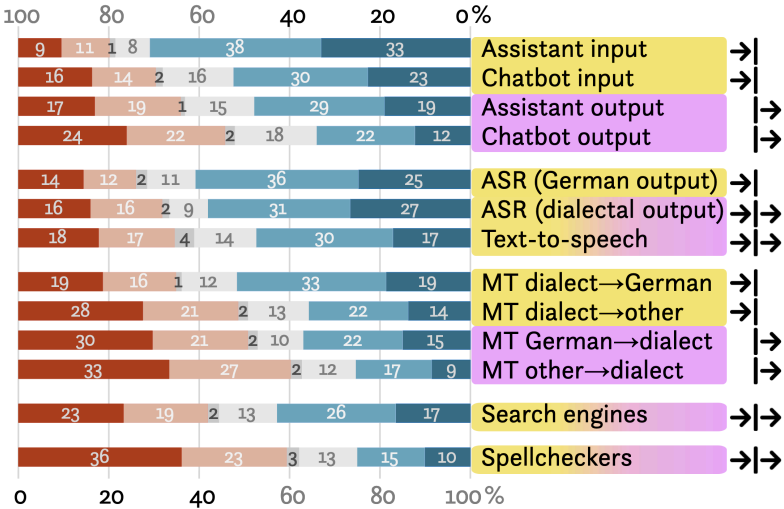


Which dialect LTs are deemed useful?

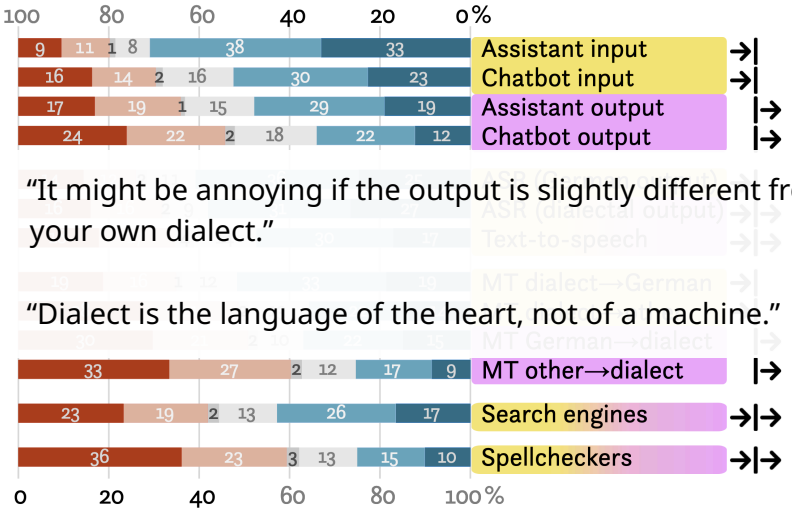


"The beauty of dialects is that there are no spelling/grammar rules and everyone can write in their own dialect, which is important since the exact version of one's dialect can be extremely local."

Dialect input vs. output?



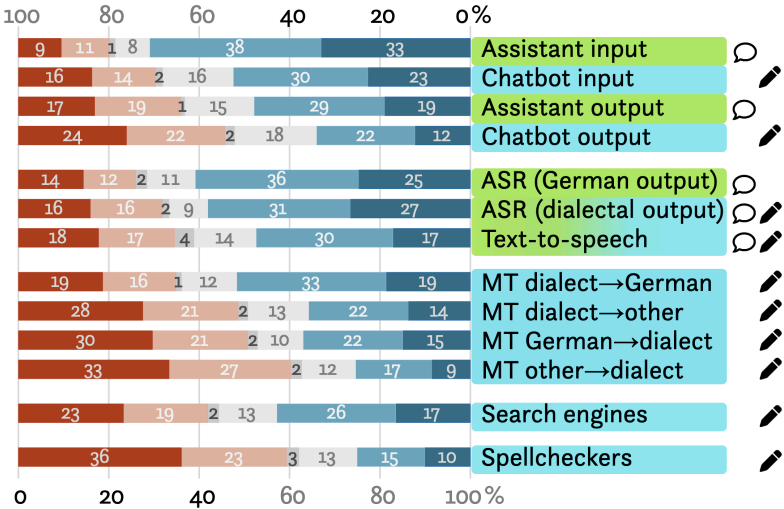
Dialect input vs. output?



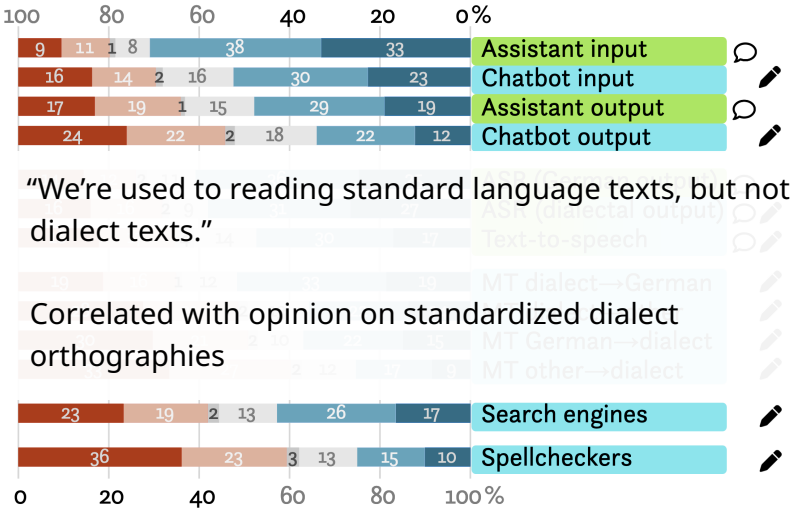
"It might be annoying if the output is slightly different from your own dialect."

"Dialect is the language of the heart, not of a machine."

Spoken vs. written dialect?



Spoken vs. written dialect?



"We're used to reading standard language texts, but not dialect texts."

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? (region)

- Low Saxon
  - Recognized as language
  - Linguistically more distant
  - Preservation efforts
  - 👍 Dialect LTs in general
  - 👍 Orthographies + spellcheckers
- Central/Southern Germany + Austria
  - Partially replaced by regiolects
- Swiss German
  - High prestige
  - Strong diglossia
  - 👎 Orthographies + spellcheckers
  - 👍 Spoken dialectal input

Takeaways

Blaschke, Purschke, Schütze & Plank (ACL 2024)

“What do dialect speakers want?”

- 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!

Personal takeaway: New projects with speech

Under submission

	D’	Lomma	hod	a	rechd	a	sauwas	Wossa
	Die	Lammer	hat	ein	recht		sauberes	Wasser

Phonetic + grammatical + lexical differences

Personal takeaway: New projects with speech

Work in progress

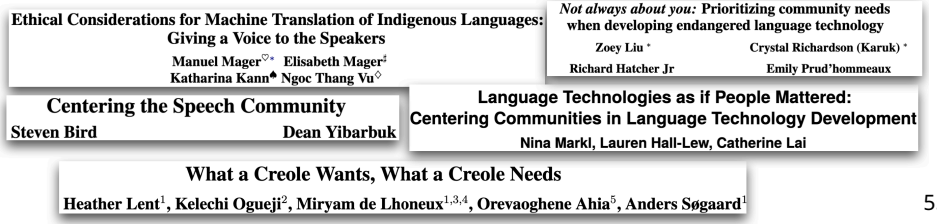
[ENG] Will it snow today?

[DEU] Wird es heute schneien? → Intent: weather query

[BAR] Werds heid schneim?

Conclusion

- ! What: Dialects and language variation
- ! How: Dialect NLP – (Some) challenges and methods
  - Part of a longer line of work on robustness
  - Still not solved, still new methods!
  - Exchange between research communities
- ! For whom: Speaker perspectives
  - Applied technologies & speaker communities
  - The next big topic? (not just in dialect NLP)



Dialects = “noisy” language?

- ! What: Dialects and language variation
- ! **How**: Dialect NLP – (Some) challenges and methods
- ! For whom: Speaker perspectives

→ For processing dialects, it can be helpful to treat language variation as “noise” (to some extent),

Dialects = “noisy” language?

- **What**: Dialects and language variation
- How: Dialect NLP – (Some) challenges and methods
- **For whom**: Speaker perspectives

→ For processing dialects, it can be helpful to treat language variation as “noise” (to some extent),
but dialects are more than just noise
– and that’s important for downstream NLP applications!

Dialects = “noisy” language?

- ! What: Dialects and language variation
- ! How: Dialect NLP – (Some) challenges and methods
- ! For whom: Speaker perspectives

→ For processing dialects, it can be helpful to treat language variation as “noise” (to some extent),
but dialects are more than just noise
– and that’s important for downstream NLP applications!

Papers & resources:
dialect-erc.github.io
verenablaschke.github.io




Thank you to my collaborators!
Thanks for listening!

On the 2026 postdoc job market

Appendix

LT ranking

	All		Non-activists only	
Rank	**LTs**	**Mean**	**LTs**	**Mean**
1	Assistant (in)	3.75	Assistant (in)	3.80
2	ASR (std out)	3.46	ASR (std out)	3.48
3	ASR (dial out)	3.38	Chatbot (in)	3.25
4	Chatbot (input)	3.29	ASR (dial out)	3.24
5	MT (dial→std)	3.17	Assitant (out)	3.01
6	Assitant (out)	3.14	MT (dial→std)	3.00
7	TTS	3.13	TTS	2.99
8	Search engine	2.94	Search engine	2.69
9	Chatbot (out)	2.76	Chatbot (out)	2.59
10	MT (dial→other)	2.73	MT (dial→other)	2.59
11	MT (std→dial)	2.71	MT (std→dial)	2.53
12	MT (other→dial)	2.39	MT (other→dial)	2.17
13	Spellchecker	2.38	Spellchecker	2.08