Dialect NLP

How (and why) to process written and spoken dialect data

Verena Blaschke
LMU Munich & MCML

Saarland University, Phonetics Colloquium
04 February 2026

Munich Center for Machine Learning

Natural Language Processing

… but *which* languages?

- Many speakers, abundant data, standardization

But does everyone use language this way?

- Also include minority languages, non-standard varieties
- Tricky for NLP! (sparse, heterogeneous data)
- Dialects are an interesting example of language variation that often is overlooked in NLP

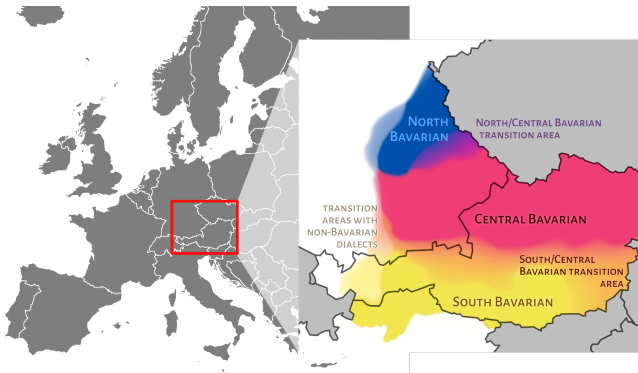
What do I mean with “dialects”?

Many definitions in linguistics, NLP & everyday language

- Any language variety spoken by a (geographically) distinct group of speakers
- Any *non-standard* 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)
- 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"

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!)

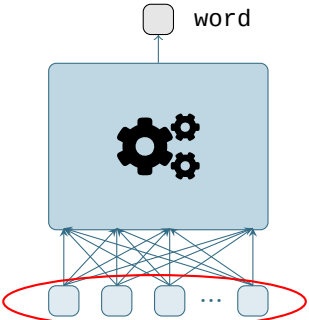
Linguistic differences

Differences from the standard language in

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

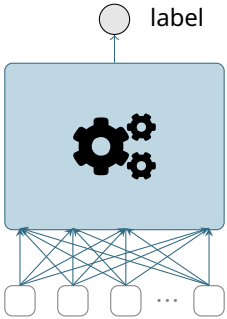
Cross-dialectal transfer

✗ Pretraining



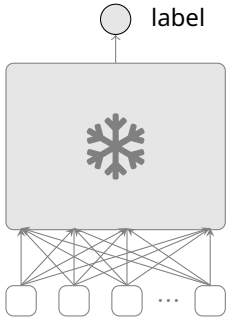
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum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

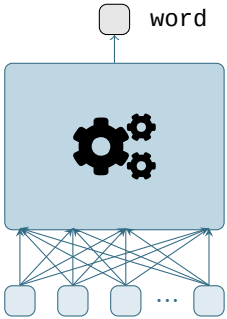
Die	Lammer	hat	ein	recht		sauberes	Wasser
`Die`	`Lamm` `-er`	`hat`	`ein`	`recht`		`sauber` `-es`	`Wasser`

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*

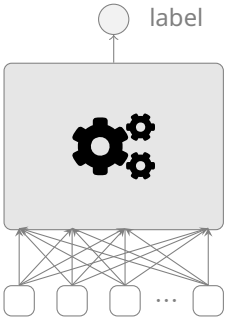
“The Lammer (river) has fairly clean water”

Sentence via `bar.wikipedia.org/wiki/Låmma`
GBERT: Chan et al. (COLING 2020) “German’s next language model”

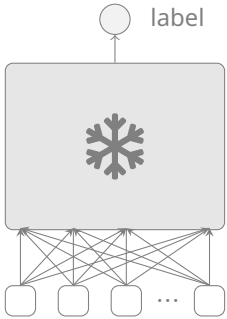
More robust input representations?



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,

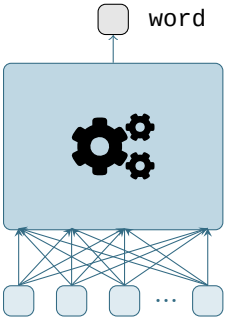


Task-specific input text



Input text in related dialect

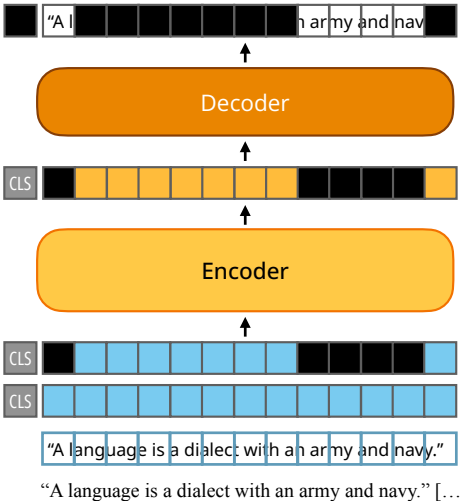
More robust input representations?



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

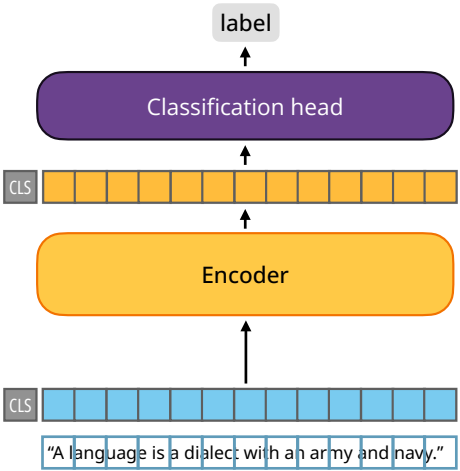
Mask spans

Projection

Render text as image

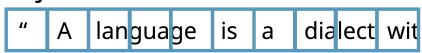
“A language is a dialect with an army and navy.” […]

Pixel models (Rust+, 2023) – fine-tuning



“A language is a dialect with an army and navy.” […]

Text rendering can be adjusted for word-level tasks



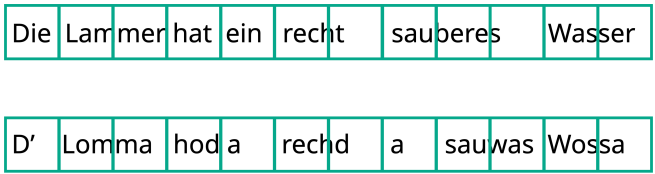
Pixel models – robustness

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

Attack	Sentence
NONE	Penguins are designed to be streamlined
CONFUSABLE	Pe𝛑guins 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



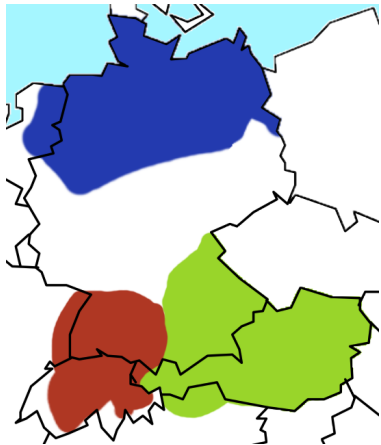
Evaluating Pixel Language Models on Non-Standardized Languages

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

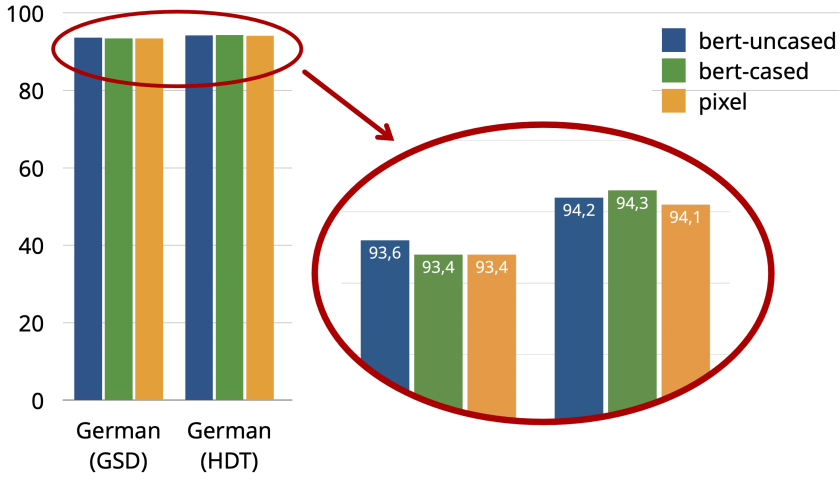
COLING 2025

German Pixel experiments

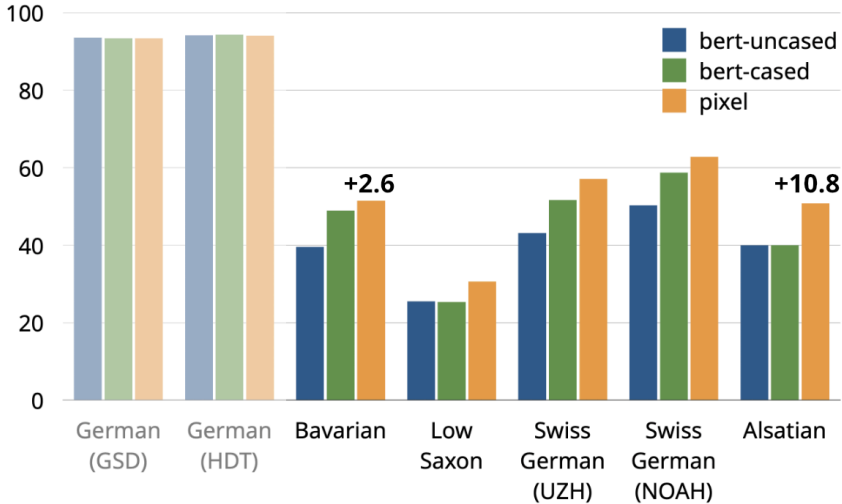
- German Pixel model (new!)
  - Same training data as a German BERT model
- Fine-tune on German, evaluate on dialects/regional languages
- 2 grammatical tasks: POS tagging, parsing
- 2 semantic tasks: intent classification (easy), topic classification (harder)



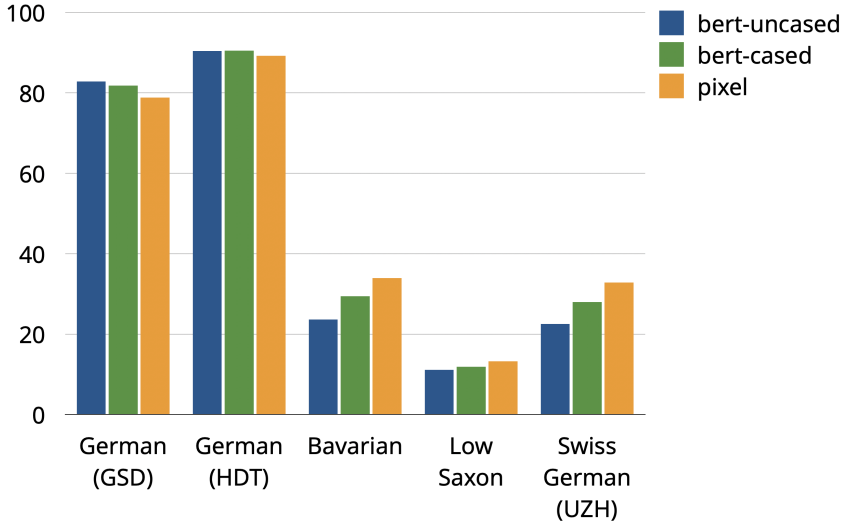
German Pixel: POS tagging (accuracy)



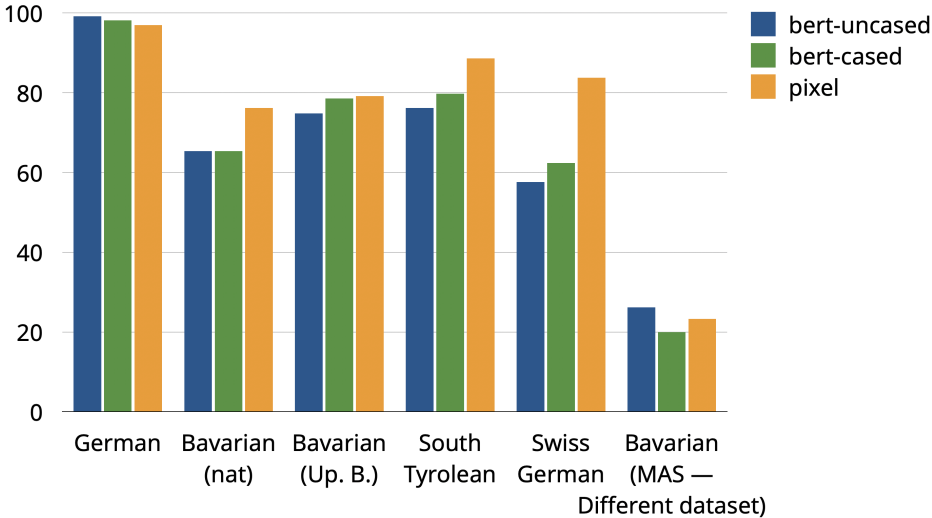
German Pixel: POS tagging (accuracy)



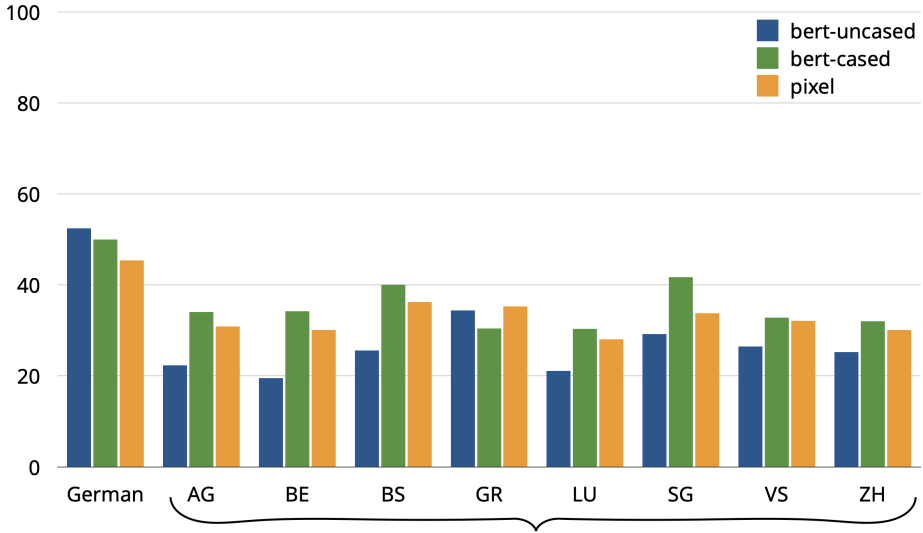
German Pixel: Parsing (LAS)



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

$\rightarrow$ Worthwhile for other settings where tokenizers don't work well?

Other input representations

Sneak-peek (paper under review)

Why not speech, given that dialects are predominantly spoken?

Standard-to-Dialect Transfer Trends Differ across Text and Speech: A Case Study on Intent and Topic Classification in German Dialects

Verena Blaschke Miriam Winkler Barbara Plank

Preprint; under review

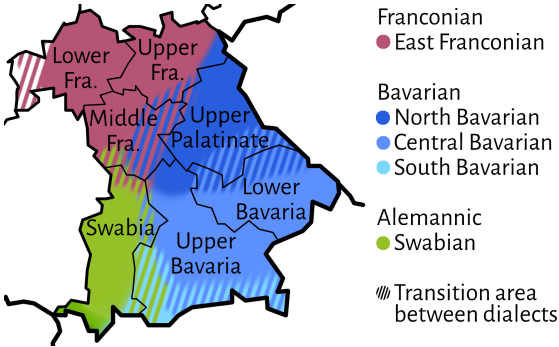
Speech vs. text

- Intent & topic classification
- Fine-tune text/speech encoders on German, test on dialects
- Three set-ups:

Speech vs. text – Findings

- German
  - Text-only > cascaded > speech-only
- Dialects
  - Speech-only > cascaded
  - Speech-only > text-only (mostly)
  - Text-only vs. cascaded: depends heavily on ASR quality!

Transcribing dialect data



A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation

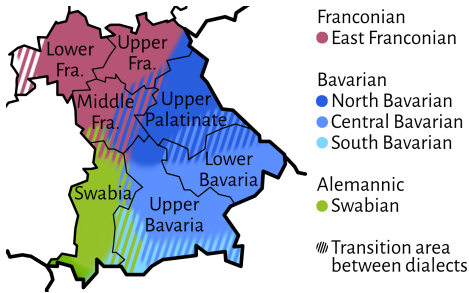
Verena Blaschke[1,2], *Miriam Winkler*[1], *Constantin Förster*[3], *Gabriele Wenger-Glemser*[3], *Barbara Plank*[1,2]

Interspeech 2025

Dataset

Good-night stories for children broadcast on the radio
→ read speech, high-quality recordings

- Dialectal audio recordings from the 7 administrative regions of Bavaria
- 1 dialectal & 1 German transcription per sentence
- German audio split for comparison
- 30+ mins per variety

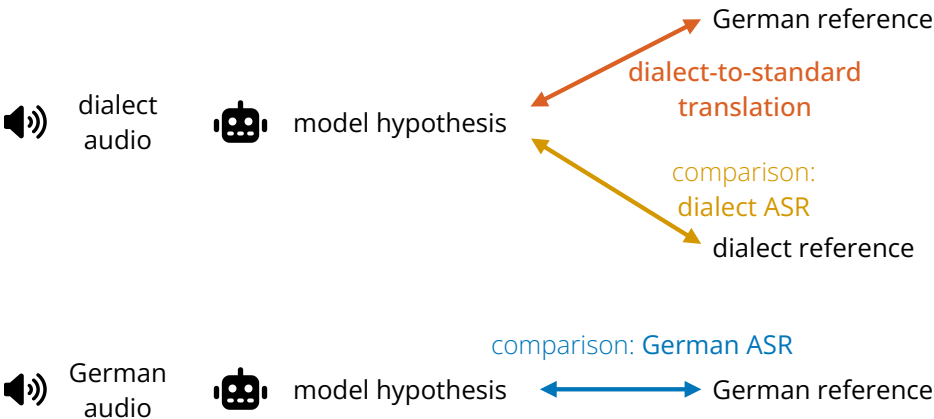


Linguistic differences

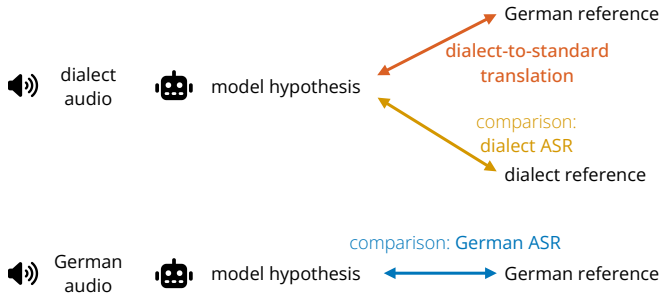
Differences from the standard language in

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

Experiments: Setup



Experiments: Metrics



- CER – spelling differences between standard & dialect
- WER – lexically/structurally similar outputs desired, also for translation
  (in paper additionally BLEU)

Experiments: Models

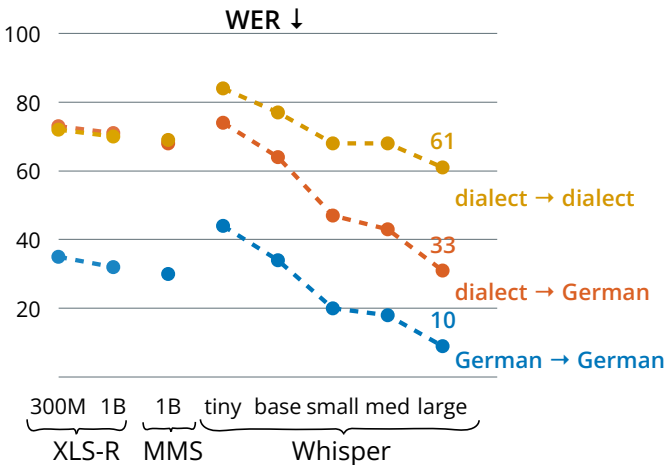
Architectures

- Whisper – language model decoding
- MMS – connectionist temporal classification (CTC)
- XLS-R (fine-tuned for German ASR) – CTC

Multiple sizes (more sizes & fine-tuned versions in paper)

Output language setting: German (no dialects available)

Quantitative results



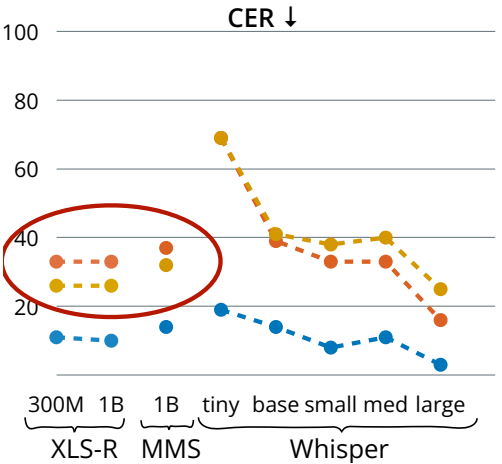
Performance gap
German vs. dialectal audio (but no systematic differences across regions)

Larger models = better

Dialect audio & decoder types

- Whisper outputs: closer to German
- XLS-R & MMS (CTC): similarly distant to both German & dialect

Quantitative results



CTC models:
Output is closer to dialect than German on a character level

dialect → dialect

dialect → German

German → German

Human evaluation

Comparing $\sim$600 of the best model's hypotheses (Whisper large-v3) to the German references:

- Meaning: Is the meaning fully preserved? $\mu = 3.9 \pm 1.1$
- Fluency: Does the output sound like fluent German? $\mu = 3.7 \pm 1.1$
- Likert scale: 1 = worst, 5 = best
- 2–3 annotators / sentence

Moderately correlated w/ automatic metrics: $0.48 \leq |\rho| \leq 0.59$

- Higher when taking the mean of *meaning* and *fluency*: $0.53 \leq |\rho| \leq 0.63 \rightarrow$ interplay

Error analysis

Same ~600 sentences:
- ✔ identical to German reference
- ✔ different, but acceptable
- ✘ different, and wrong

[German]	**Sofort**		**Mathildas**		**Geldstück**	**suchen,**	…
	Immediately		*Mathilda's*		*coin*	*search*	
[Dialect]	Sofort	da	Mathilda	ihr	Geldstückle	sung,	…
		the	*Mathilda*	*her*			
[ASR]	**Sofort**	**der**	**Mathilda**	**ihr**	**Geldstück**	**lesung,**	…
	✔	✔	✔	✔	✔	✘	

Error analysis

Subset (~600 sentences): ✔ identical to German reference

✔ different, but acceptable

✘ different, and wrong

Words/constructions that…

- are identical in German & the dialect: usually correct (86 %) ✔
- differ only in terms of pronunciation/morphology: usually correct (75 %) ✔
- lexically different: usually nonsense (63 %) ✘
- syntactically different: usually like the dialectal structure (acceptability in German varies) ✔ ✘

Common error source: incorrectly recognized word boundaries

Transcribing dialectal speech is difficult

Blaschke, Winkler, Förster, Wenger-Glemser & Plank (Interspeech 2025)

“A multi-dialectal dataset for German dialect ASR …”

Differences from the standard language in

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

—

- Robustness wrt pronunciation differences

Transcribing dialectal speech is difficult

Blaschke, Winkler, Förster, Wenger-Glemser & Plank (Interspeech 2025)
"A multi-dialectal dataset for German dialect ASR …"

Differences from the standard language in

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

—

- Robustness wrt pronunciation differences
- Difficult balance between being both faithful to the audio and in acceptable German – also an evaluation challenge!

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

ACL 2024

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 applied 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, online
- 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.

Questionnaire

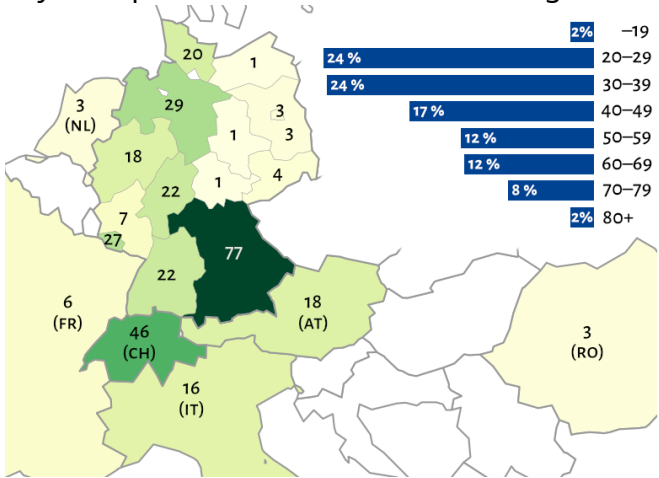
G310

20. Stimmen Sie den folgenden Aussagen zu?

Es sollte Transkriptionsprogramme geben, …	Ja, unbedingt	Eher ja	Weder noch	Eher nein	Nein, das halte ich nicht für sinnvoll	Das kann ich nicht bewerten
... die Audioaufnahmen in meinem Dialekt als geschriebenes Hochdeutsch wiedergeben.	○	○	○	○	○	○
... die Audioaufnahmen in meinem Dialekt als geschriebenen Dialekt wiedergeben.	○	○	○	○	○	○

Dialect background and attitudes

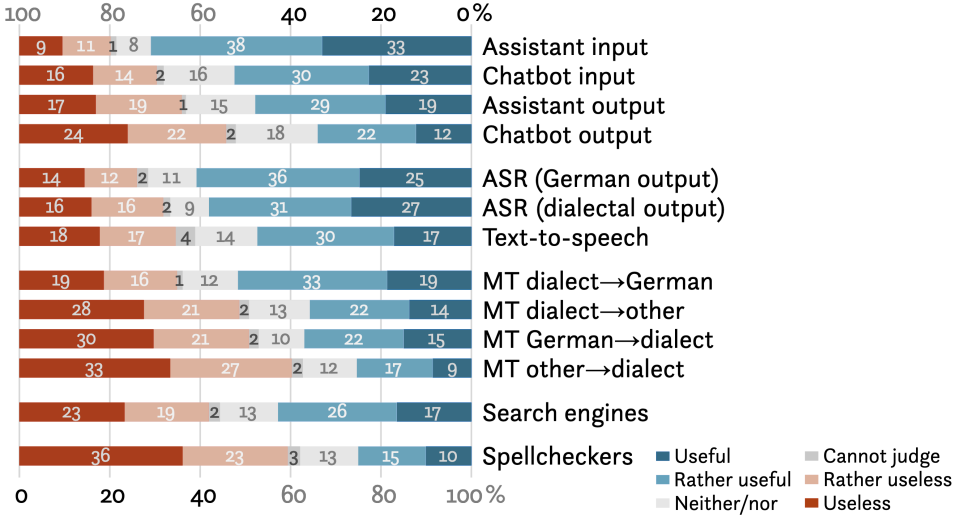
Responses by **327** speakers of German dialects/regional languages



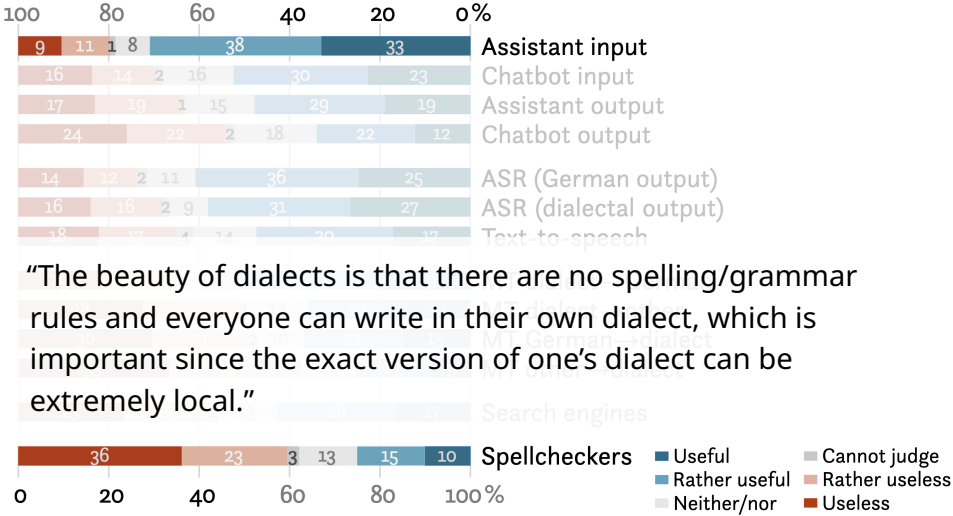
Dialect background and attitudes

- 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?

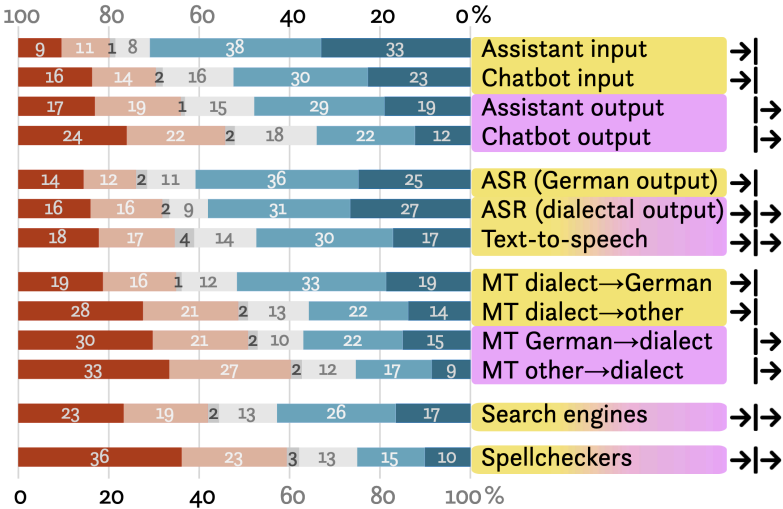


Which dialect LTs are deemed useful?

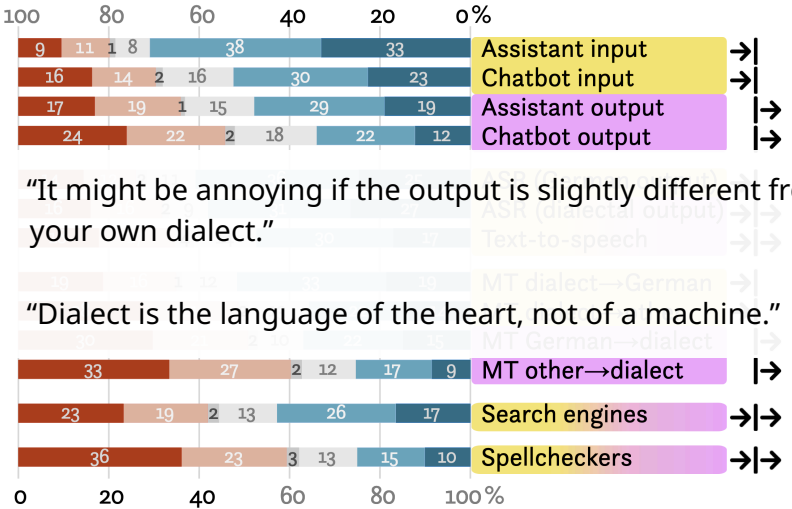


"The beauty of dialects is that there are no spelling/grammar rules and everyone can write in their own dialect, which is important since the exact version of one's dialect can be extremely local."

Dialect input vs. output?



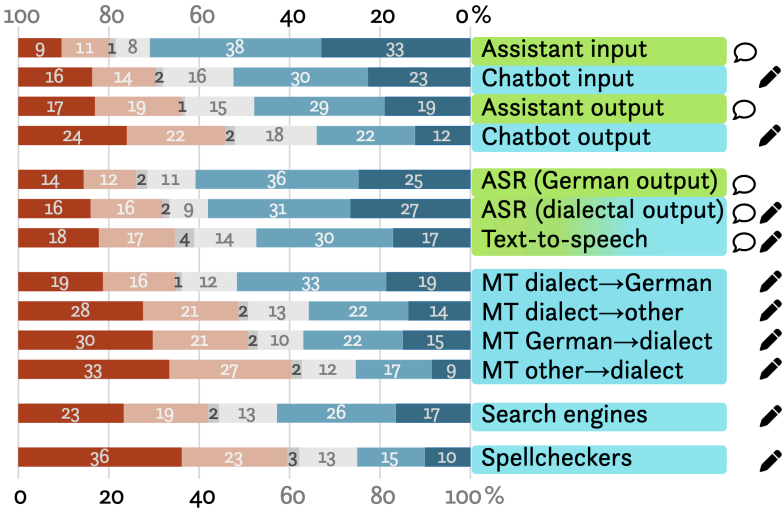
Dialect input vs. output?



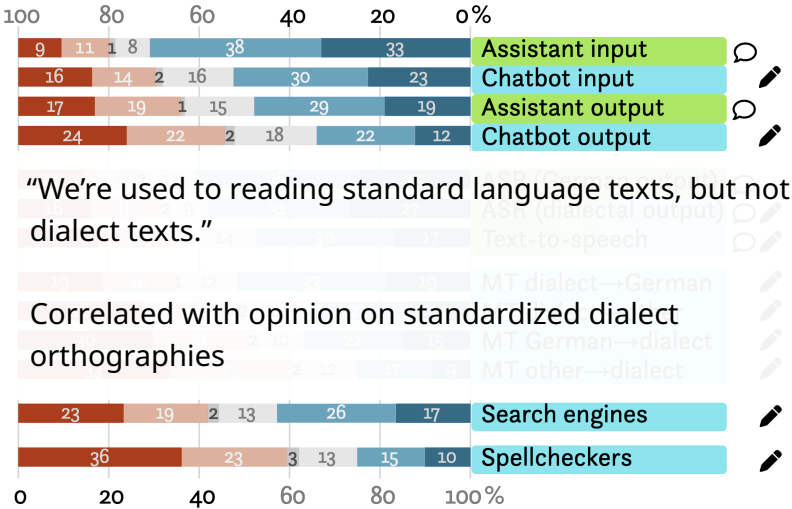
"It might be annoying if the output is slightly different from your own dialect."

"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

Dialect "strength"

- Respondents with especially traditional dialects want more strongly that dialectal output corresponds to their exact variety
- Otherwise almost no notable effect

Age

- Very few statistically significant correlations!
- Young respondents: especially interested in the most popular 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

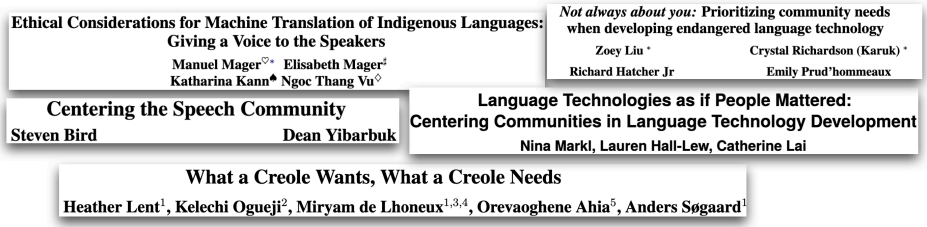
Blaschke, Purschke, Schütze & Plank (ACL 2024)

“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!

Conclusion: Dialect NLP

- Challenges:
  - Data availability & quality
  - Input representations
  - Variation & NLG: evaluation challenge
- Speaker perspectives regarding applied technologies are important – not just in dialect NLP



Thank you for listening!