Dialect NLP

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

Verena Blaschke
University of Groningen

CLASP research seminar, University of Gothenburg
14 October 2026



university of groningen

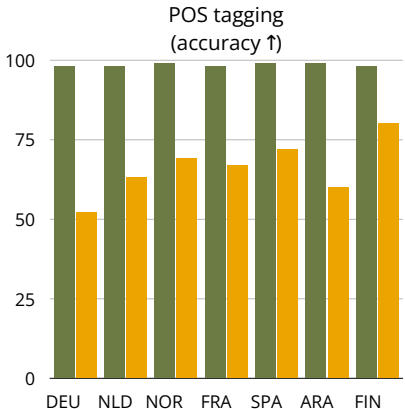
Natural Language Processing
… but *which* languages?

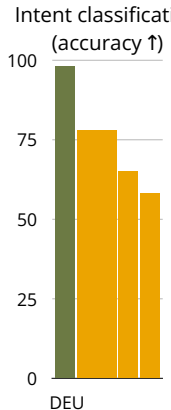
- Many speakers, abundant data, standardization

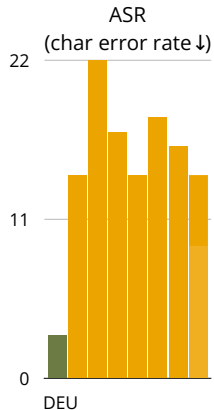
But does everyone use language this way?

- 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

Performance on standard languages and related non-standard varieties







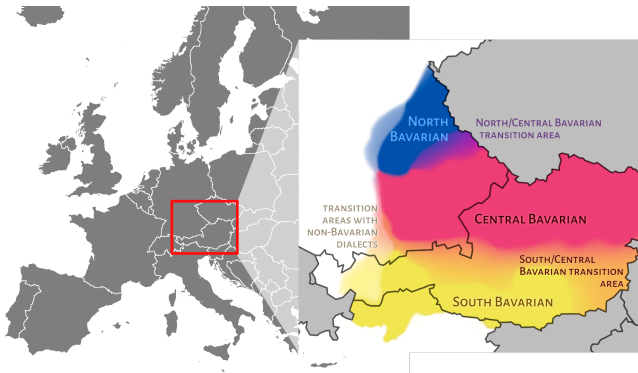
What do I mean with “dialects”?

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
- …

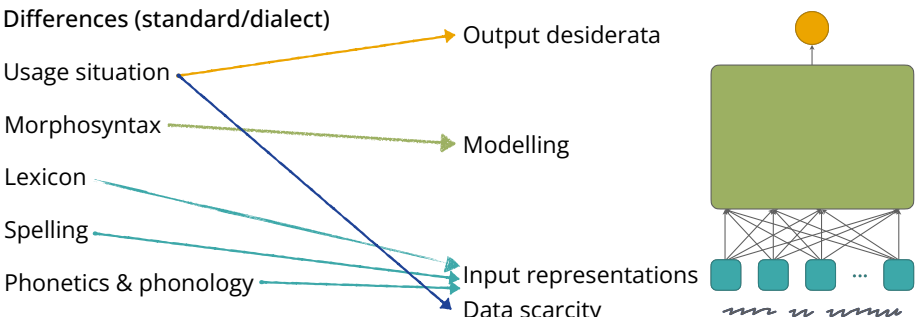
What do I mean with “dialects”?

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



Linguistic differences

BAR	Isches	sunnig	afoure?	I	mechad	wissn,	wia	hoas	dass	des	weard.
DEU	Ist es	sonnig	draußen?	Ich	möchte	wissen,	wie	heiß		es	wird.
	Is it	*sunny*	*outside?*	*I*	*want*	*know*	*how*	*hot*	*that*	*it*	*becomes.*



What difference does the input modality make?

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

ACL 2026

Differences (standard/dialect)

Usage situation

Morphosyntax

Lexicon

Spelling

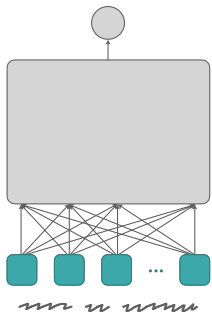
Phonetics & phonology

Output desiderata

Modelling

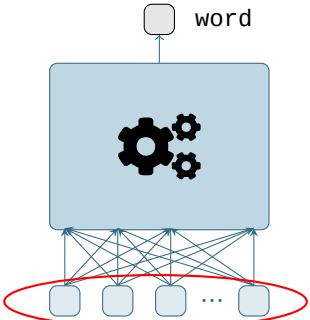
Input representations

Data scarcity



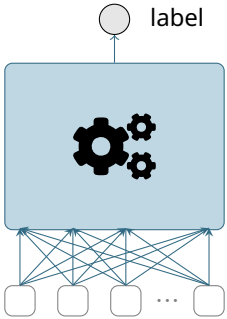
Cross-dialectal transfer

✗ Pretraining



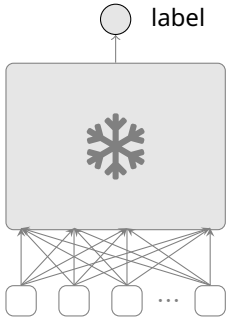
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum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

Non-standard orthographies + tokenization

A few promising approaches

- Changing the tokenization of the (standard-language) fine-tuning data can improve transfer to a related dialect

German	zweisprachig *bilingual*	zwei sprach ig *two language y*
Bavarian	zwaasprochig	z wa as pro chi g
	zwaspråchig	z was pr å chi g
	zwoasprochig	z wo as pro chi g
	zwasprochig	z was pr pro chi g
	…	

Non-standard orthographies + tokenization

A few promising approaches

- Changing the tokenization of the (standard-language) fine-tuning data can improve transfer to a related dialect

German	zweisprachig	zwei sprach ig
“noised” German	zwei**x**prachig	zwei x pra chi g
Bavarian	zwaasprochig	z wa as pro chi g
	zwaspråchig	z was pr å chi g
	…	

Aepli & Sennrich (ACL Findings 2022), Blaschke ea. (VarDial 2023)

Non-standard orthographies + tokenization

A few promising approaches

- Using vision-based encoders
  (no predefined subword tokens necessary)





Muñoz-Ortiz, Blaschke & Plank (COLING 2025)

Using spoken instead of written input?

- Natural modality for many non-standard dialects!
- Language technology for spoken dialectal inputs: popular among speakers of (German) dialects (as we'll see later)
- Spoken inputs often get transcribed before processing, and many datasets for, e.g., intent classification are thus text-only



Lorem ipsum...

Data

Requirements for controlled experiments:

1. Test data are parallel across varieties (standard/non-std.) …
2. … and also parallel across modalities (written/spoken)
3. Training data are parallel across modalities (written/spoken)
4. Labelled for a content classification task

→ **New evaluation dataset: xSID-audio**

- Spoken & written intent classification for German and Bavarian

Training data: Standard German (Speech-)MASSIVE

Additional experiments with Swiss German topic classification data
(in the paper, not in these slides)

MASSIVE (FitzGerald ea., ACL 2023), *Speech-MASSIVE* (Lee ea., Interspeech 2024)

Setups

Text-only

Ist es heute kalt?
Is es heid koid?
Is it cold today?

Speech-only



Cascaded



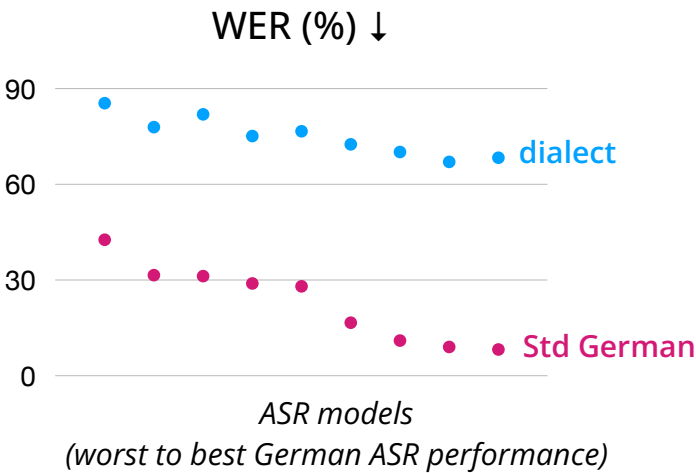
Ist es heute kalt?
Is it cold today?



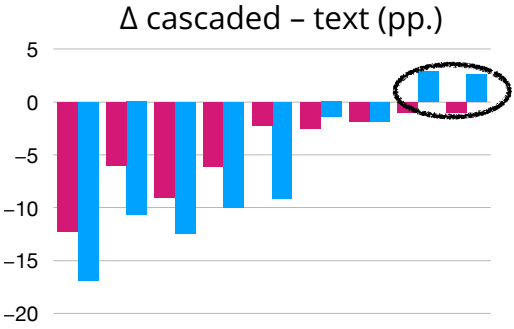
Ist es halt keut? *(sic)*
Is it just [nonce]?

- Train text/speech model on **Standard German**, test also on **dialect**
- Encoder models (some of the speech encoders can be extended with decoders for ASR)
- Accuracy averaged over multiple runs

How well do the text models generalize to ASR data?

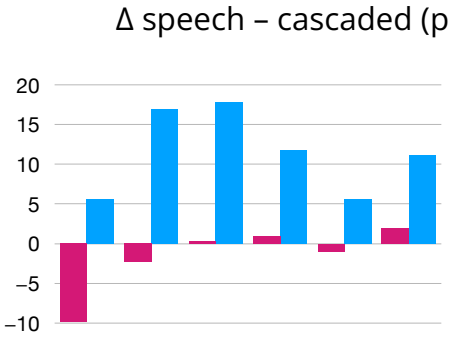


How well do the text models generalize to ASR data?



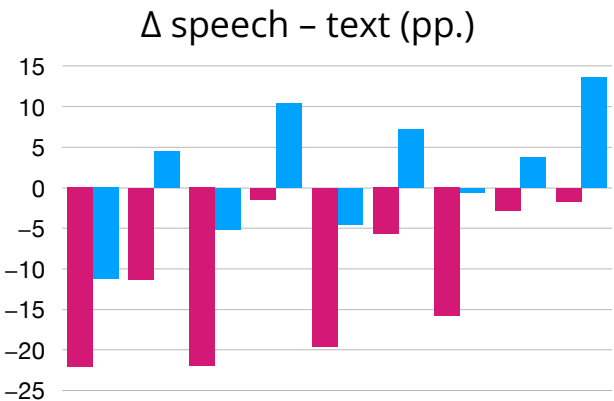
- Correlation between WER and performance Δ:
  $-0.93 \leq r \leq -0.74$
  for data shown here
  ($-0.98 \leq r \leq -0.65$ for all data)
- **Standard German**:
  gold text > cascading
- **Dialect**: cascading works best when the output is de facto normalized to Std German

How to best deal with spoken inputs?



- **Standard German**: barely makes a difference (unless the ASR model is particulary bad)
- **Dialect**: speech performs *much* better

Speech vs. text: which modality is preferable?



- **Standard German**: text > speech
- **Dialect**: speech models often outperform text models

Results for Standard German don't generalize to dialects

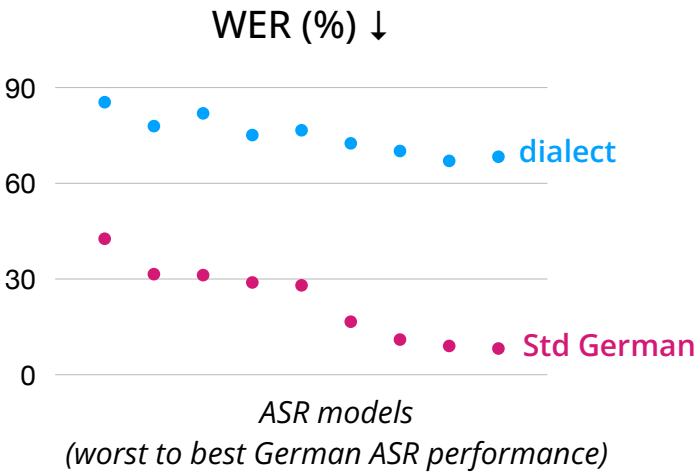


	Standard German	Dialect
Text-only	always text-only > speech	often speech > text-only
Speech-only	speech ≈ cascaded	always speech > cascaded
Cascaded	always text-only > cascaded	cascaded can beat text-only! (when ASR normalizes the data)

Takeaways

- Patterns for the standard language don’t necessarily also hold for closely related non-standard dialects!
  - Likely due to differences in input representations *(ongoing work)*
- We should rethink only constructing text-based datasets for otherwise speech-related tasks!
  - New dataset for studying modality-based differences (xSID-audio)

So, dialect ASR doesn't work well... Why?



So, dialect ASR doesn't work well... Why?

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

Differences (standard/dialect)

Usage situation

Morphosyntax

Lexicon

Spelling

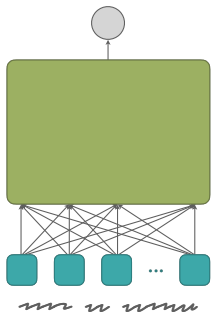
Phonetics & phonology

Output desiderata

Modelling

Input representations

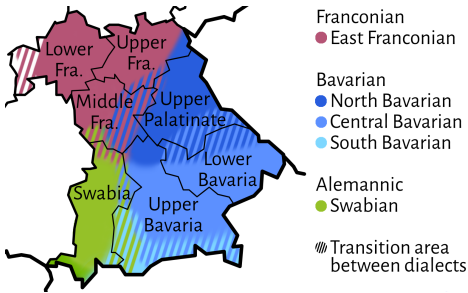
Data scarcity



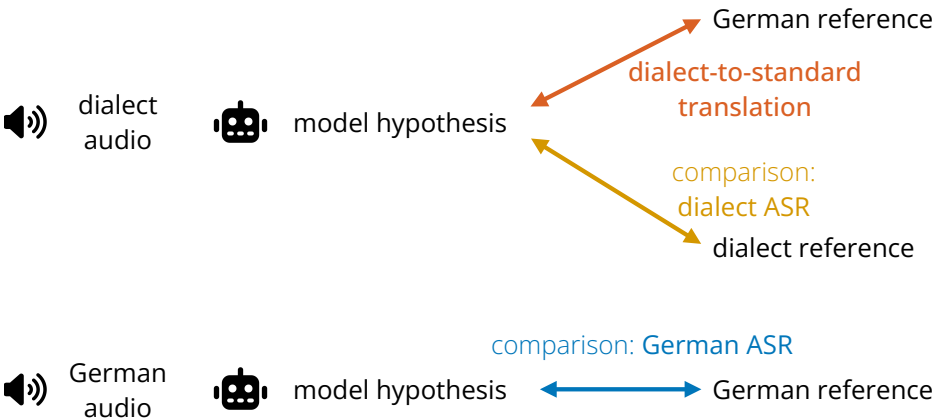
Dataset

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

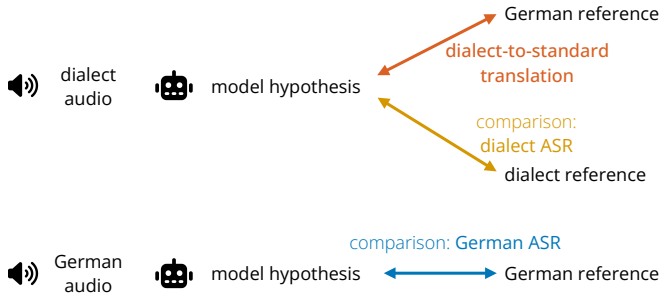
- Dialectal audio recordings from the 7 administrative regions of Bavaria
- 1 dialectal & 1 German transcription per sentence
- German audio split for comparison
- 30+ mins per variety



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

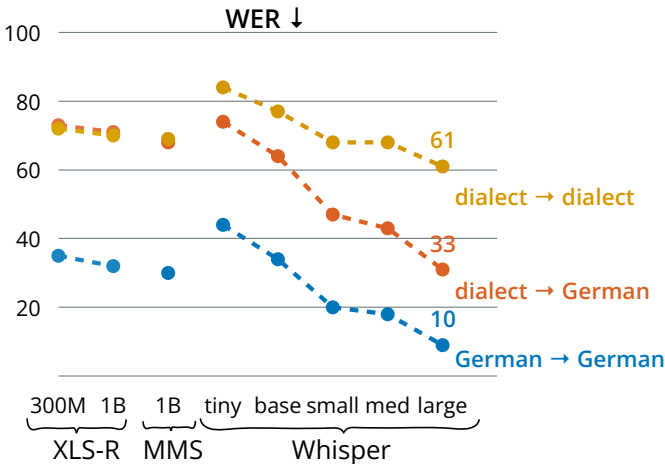
Architectures

- 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 (word error rate)



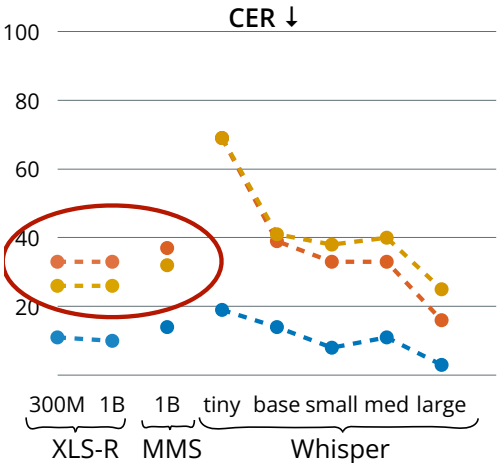
Performance gap
German vs. dialectal audio (but no systematic differences across regions)

Larger models = better

Dialect audio & decoder types

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

Quantitative results (character error rate)



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 = not at all, 5 = totally
- 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

Differences from the standard language in

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

—

- Robustness wrt pronunciation differences

Transcribing dialectal speech is difficult

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!

Perspectives of the speaker communities

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

Differences (standard/dialect)

Usage situation

Morphosyntax

Lexicon

Spelling

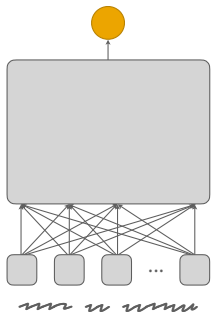
Phonetics & phonology

Output desiderata

Modelling

Input representations

Data scarcity



Why dialect NLP?

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

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

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

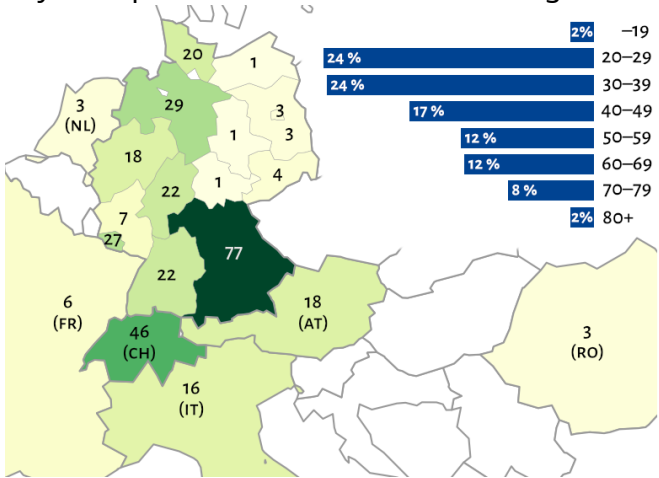
20. Stimmen Sie den folgenden Aussagen zu?

G310

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

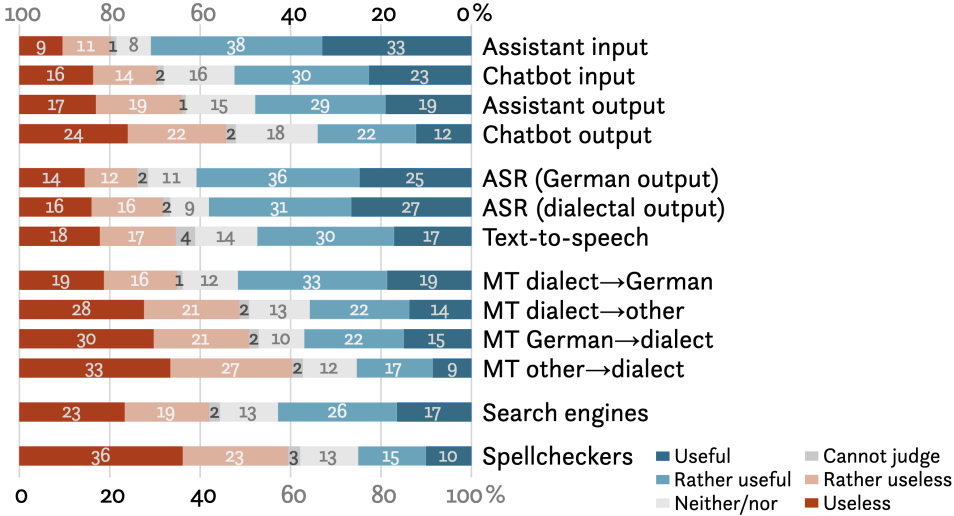
Responses by **327** speakers of German dialects/regional languages



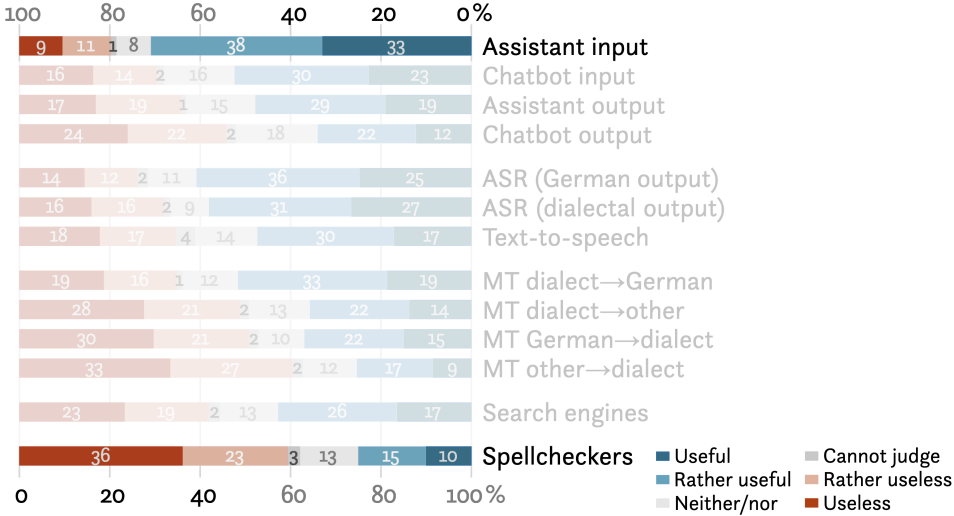
Dialect background and attitudes

- 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

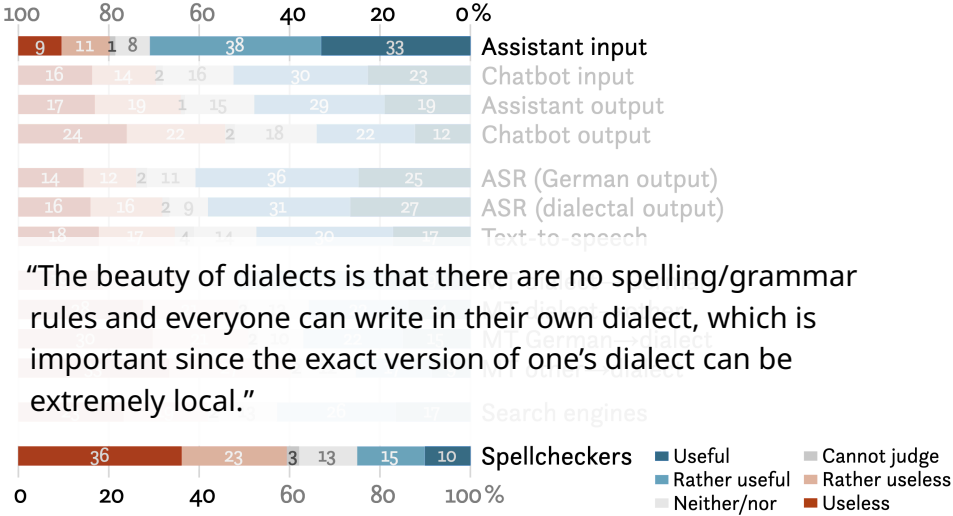
Which dialect LTs are deemed useful?



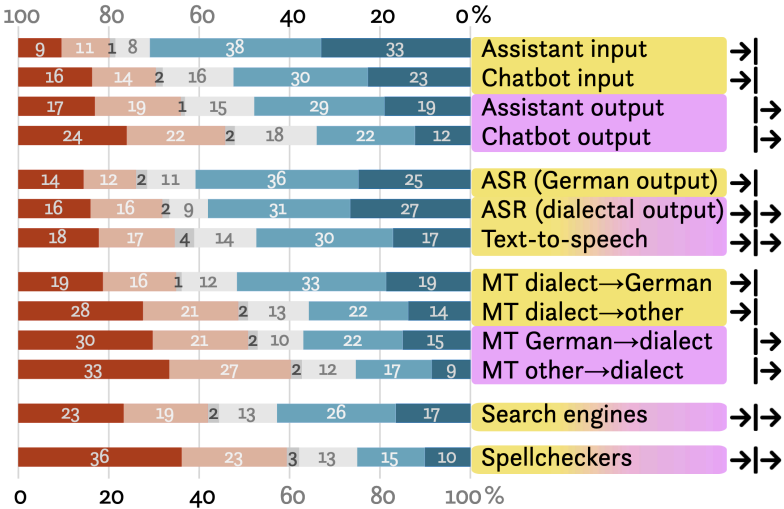
Which dialect LTs are deemed useful?



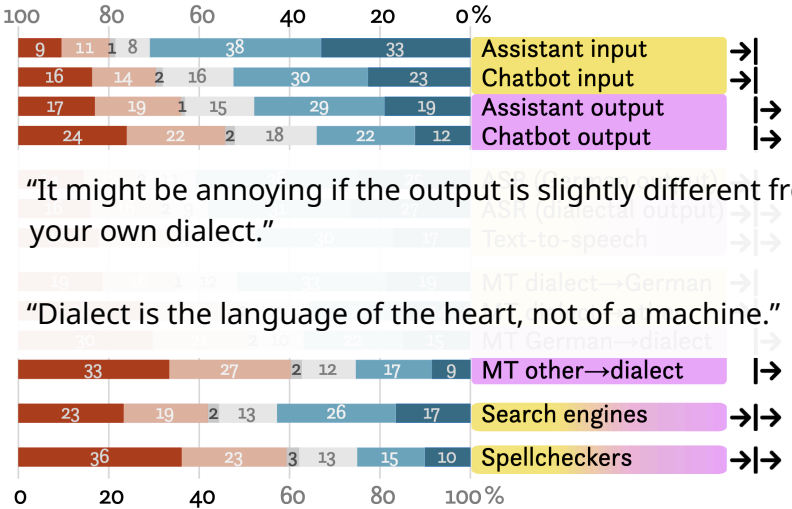
Which dialect LTs are deemed useful?



Dialect input vs. output?



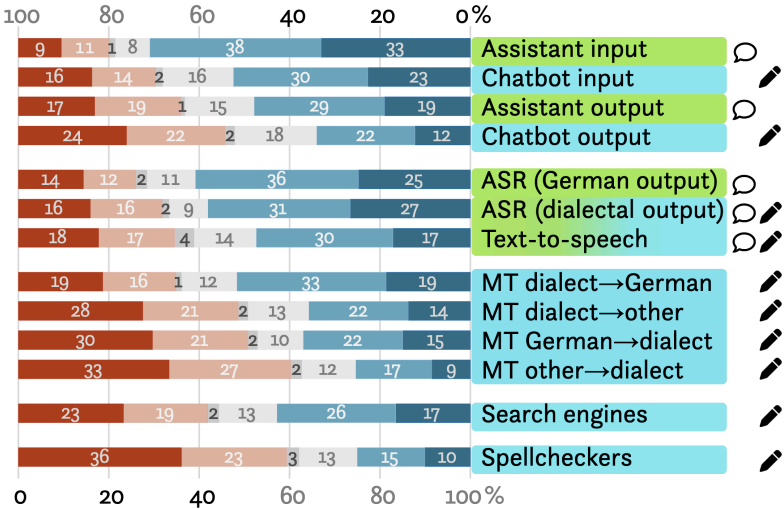
Dialect input vs. output?



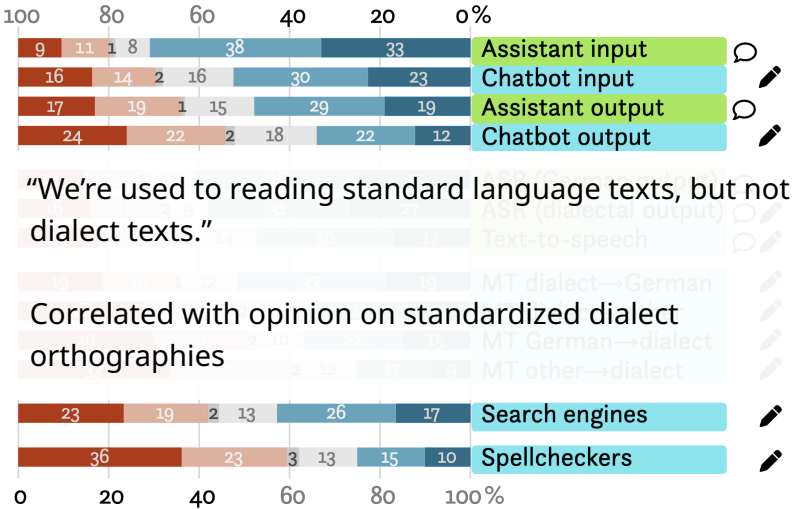
"It might be annoying if the output is slightly different from your own dialect."

"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)



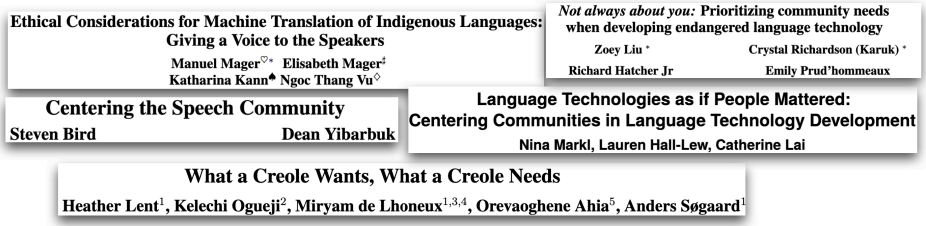
- Low Saxon
  - Recognized as language
  - Preservation efforts
  - 👍 Dialect LTs in general
  - 👍 Orthographies + spellcheckers
- Central/Southern Germany + Austria
  - Partially replaced by regiolects
- Swiss German
  - High prestige, strong diglossia
  - 👎 Orthographies + spellcheckers
  - 👍 Spoken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

Takeaways

- 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



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

Interested in more dialect NLP?

- SIG on Similar Languages, Varieties, and Dialects: aclsigvar.org
- VarDial workshop @ EACL 2027: aclsigvar.org/vardial2027

Thank you for listening!

verenablaschke.github.io v.blaschke@rug.nl