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

LMU Munich, MCML

ACL | July 5, 2026

Written dialects & tokenization

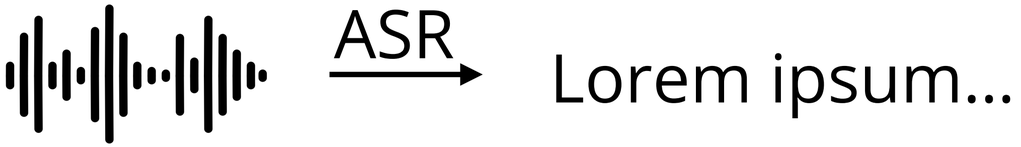
Example: tokenization with mBERT

Std German	zweisprachig	zwei ##sprach ##ig
	bilingual	*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 ##pro ##chi ##g

… but what about speech models with continuous inputs?

Spoken inputs

- Natural modality for many non-standard dialects!
  - Language technology for spoken dialectal inputs: popular among speakers of (German) dialects *[1]*
- Spoken inputs often get transcribed before processing, and many datasets for, e.g., intent classification are thus text-only



[1] What Do Dialect Speakers Want? A Survey of Attitudes Towards Language Technology for German Dialects (Blaschke et al., ACL 2024)

Data

Requirements:

1. Test data are parallel in a standard language & a related non-standard dialect …
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 dataset: xSID-audio

- Spoken + written intent classification for German & Bavarian
- Evaluation only

- Training data: German (Speech-)MASSIVE *[2,3]*

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

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages (FitzGerald et al., ACL 2023)
Speech-MASSIVE: A Multilingual Speech Dataset for SLU and Beyond (Lee et al., Interspeech 2024)

Setups

Text-only

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

Speech-only

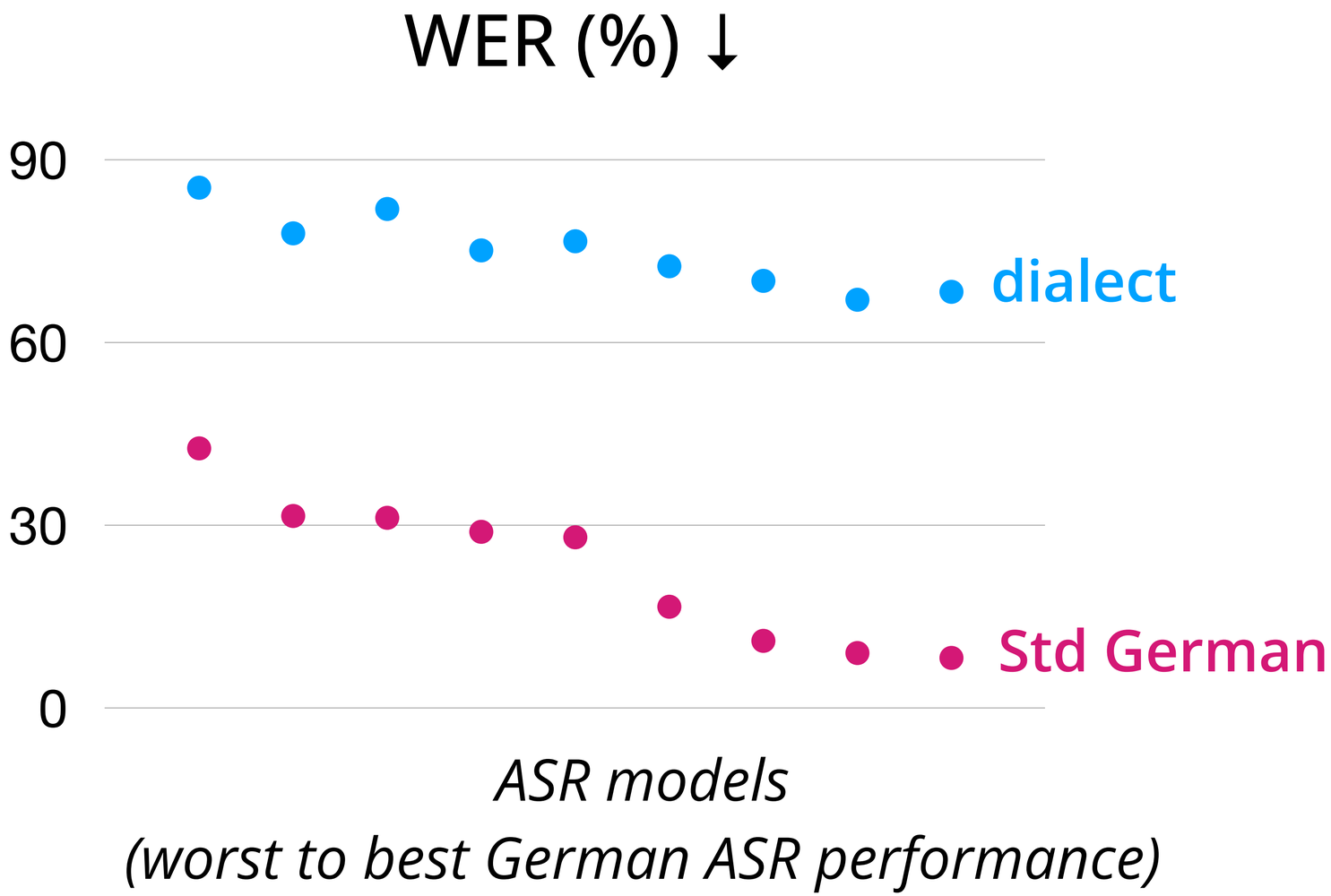
Cascaded


ASR
Ist es heute kalt?
Is it cold today?

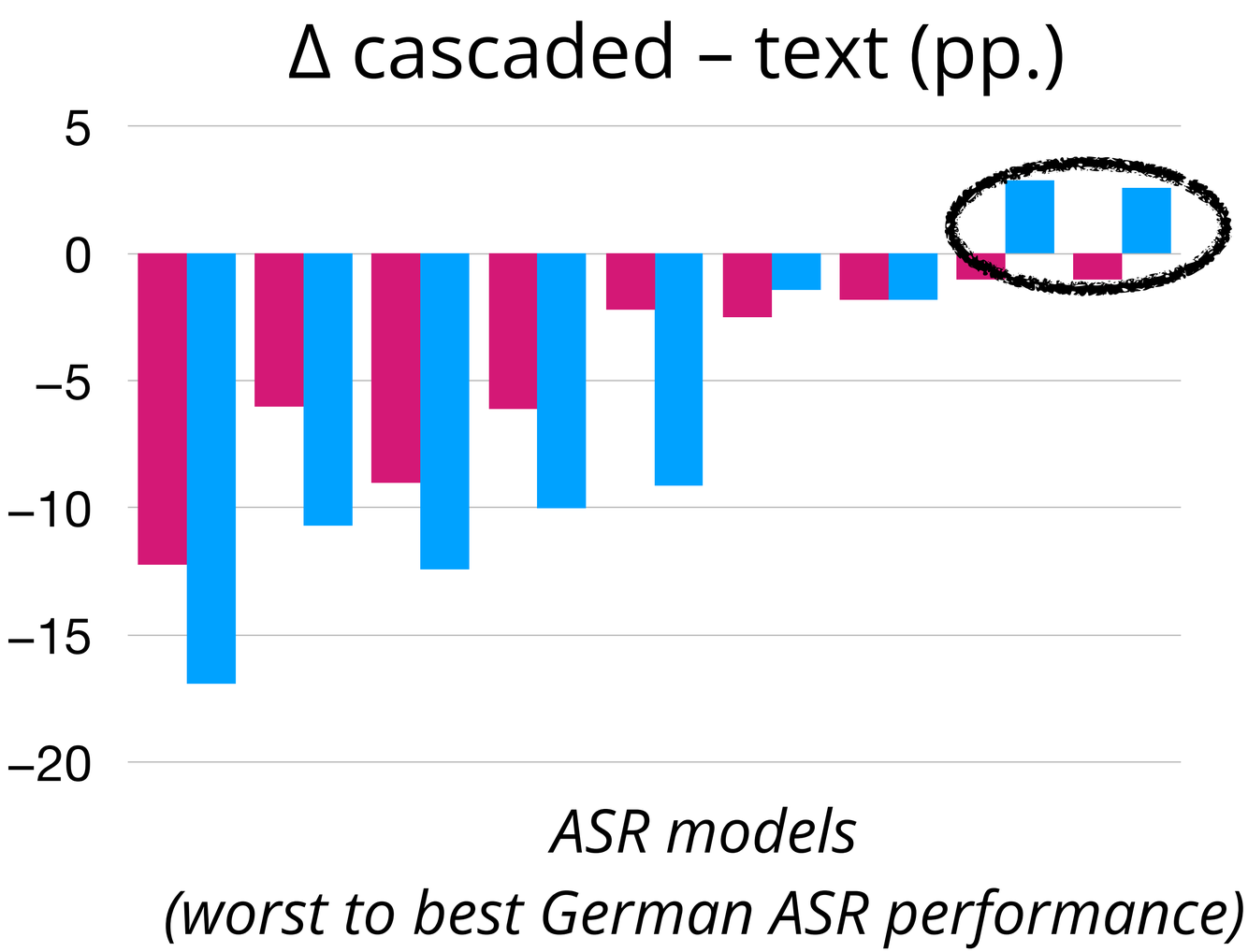

ASR
Ist es halt keut? *(sic)*
Is it just [nonce]?

- Train text/speech encoder model on **Std 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?

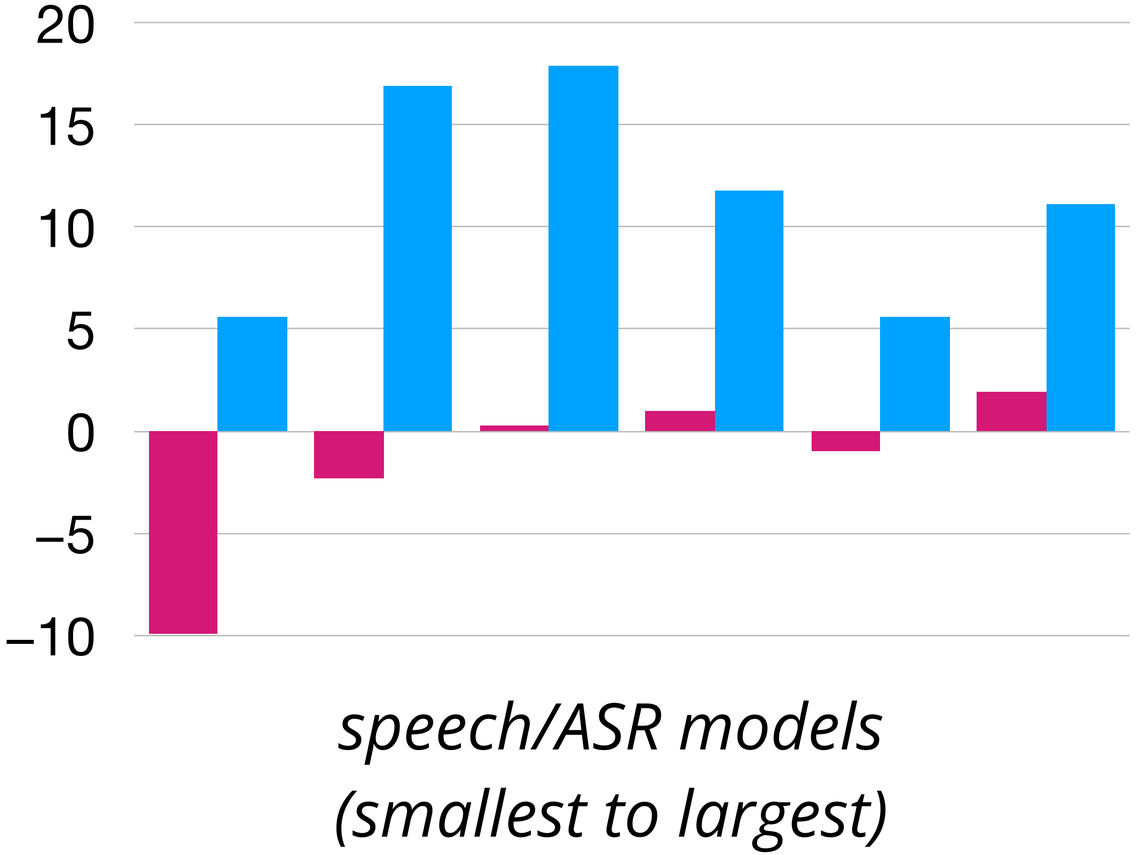


How well do the text models generalize to ASR data?



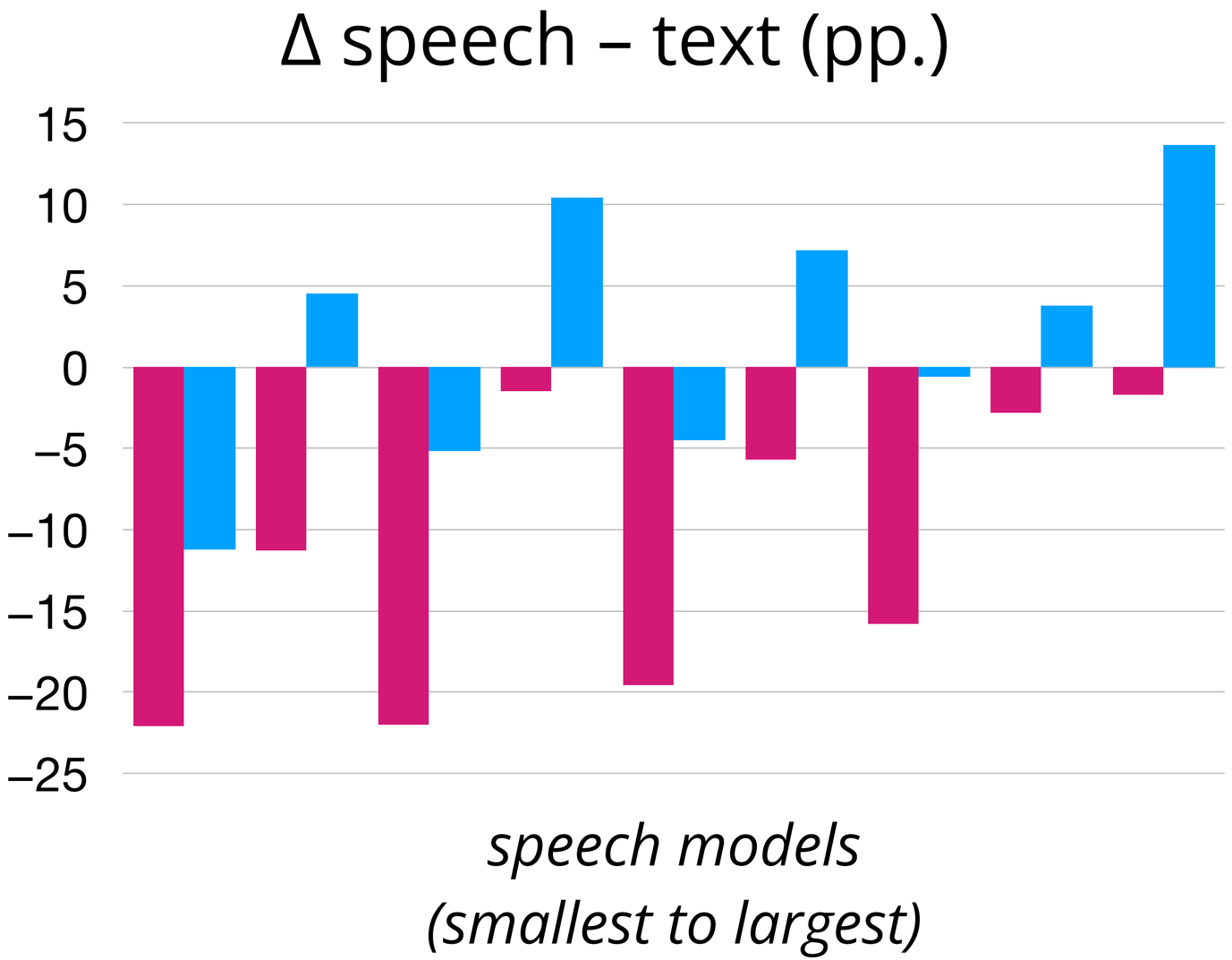
- Correlation between WER and performance Δ:
  $-0.93 \leq r \leq -0.74$
  for data shown here
  (–0.98 ≤ r ≤ –0.65 for all data)
- **Std German:**
  gold text > cascading
- **Dialect:** cascading works best when the output is de facto normalized to Std German

How to best deal with spoken inputs?



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

Speech vs text: which modality is preferable?



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

Results for Std German don’t generalize to dialects

		Std German	Dialect
Text-only	Ist es heute kalt? Is es heid koid? *Is it cold today?*	always text-only > speech	often speech > text
Speech-only		speech ≈ cascaded	always speech > cascaded
Cascaded	 ASR → Ist es heute kalt? *Is it cold today?*  ASR → Ist es halt keut? *(sic)* *Is it just [nonce]?*	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)*
- Rethink only constructing text-based datasets for otherwise speech-related tasks!
  - New dataset for studying modality-based differences (xSID-audio)