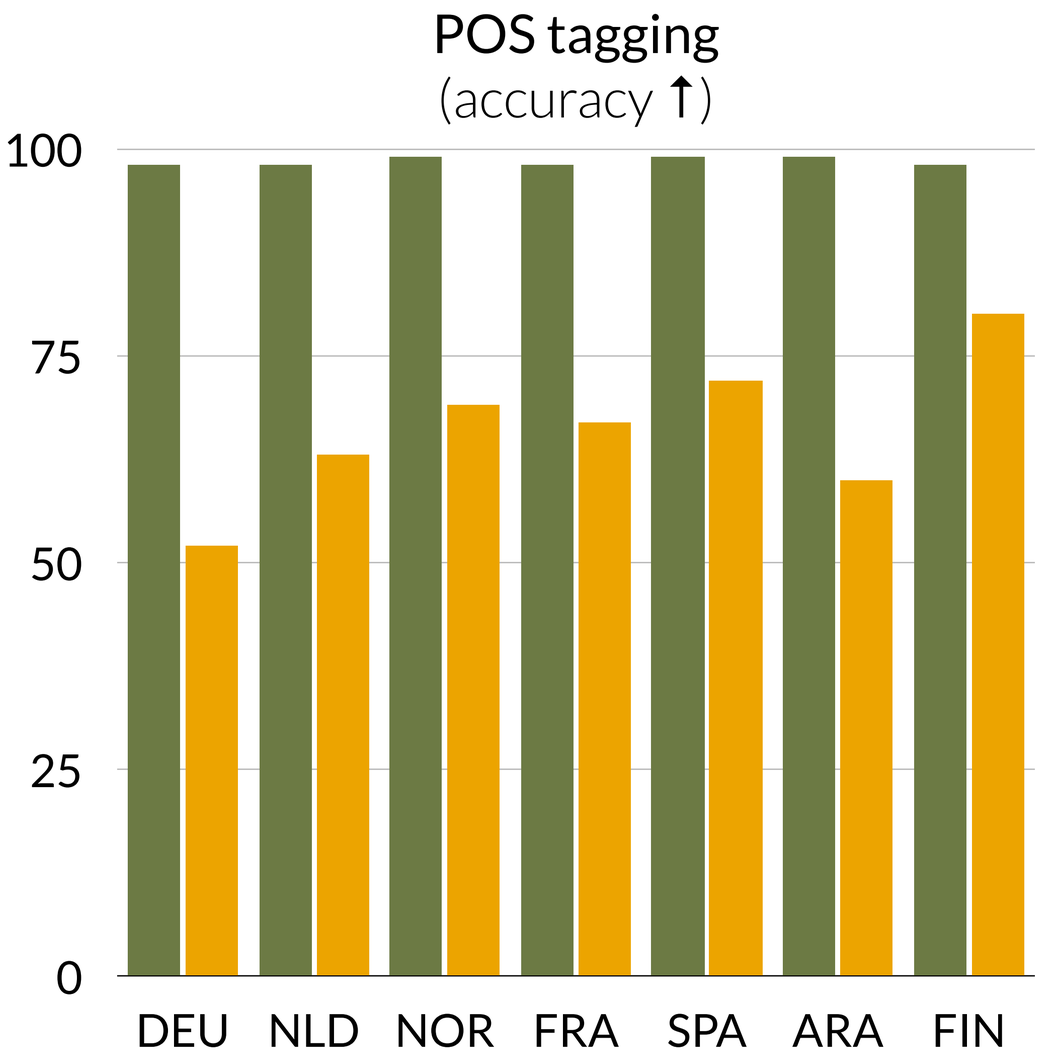
Challenges and Opportunities for Bavarian NLP and Speech Processing

Barbara Plank & Verena Blaschke
LMU Munich

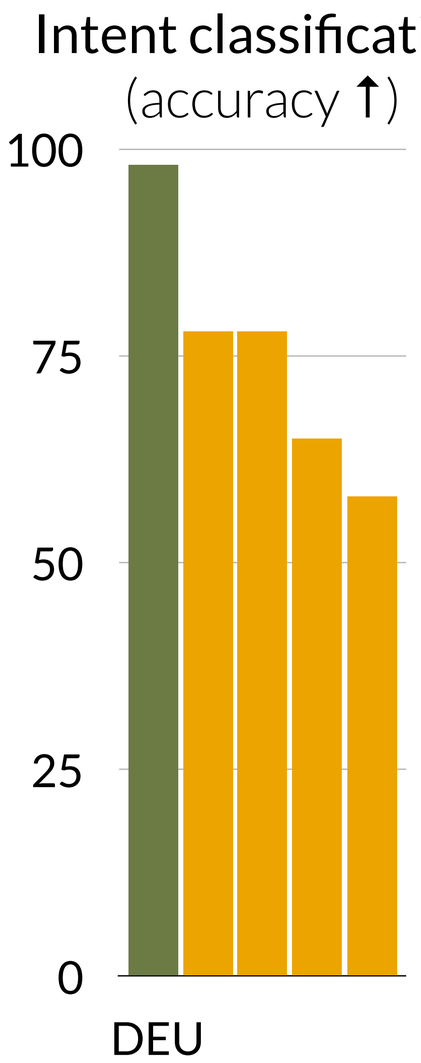
June 30, 2026

Embracing Variability in Natural Language Processing @ ICLaVE 2026
Lausanne, Switzerland

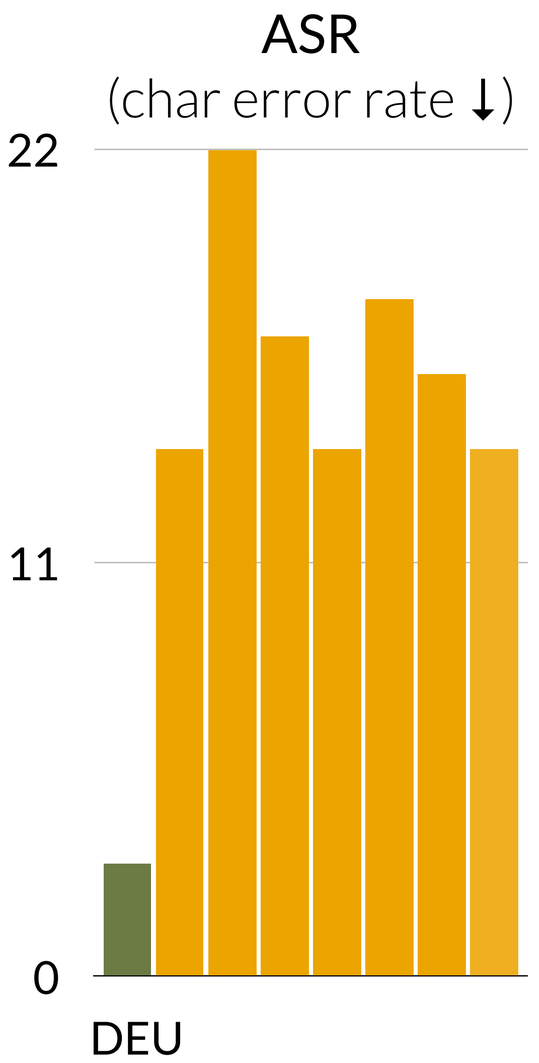
Performance on standard languages and related non-standard varieties



VarDial 2023



VarDial 2025a

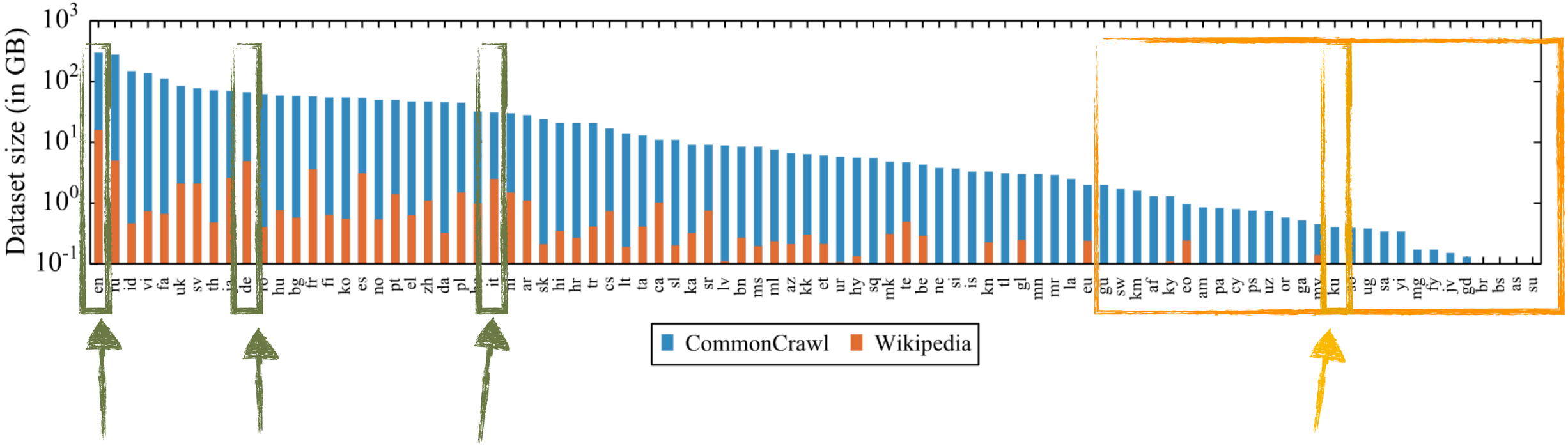


Interspeech 2025

The “monolithic” view on languages in NLP

“high-resource”
(and typically “**standard**" varieties)

versus “low-resource”, a.k.a. the “long tail”



English German Italian

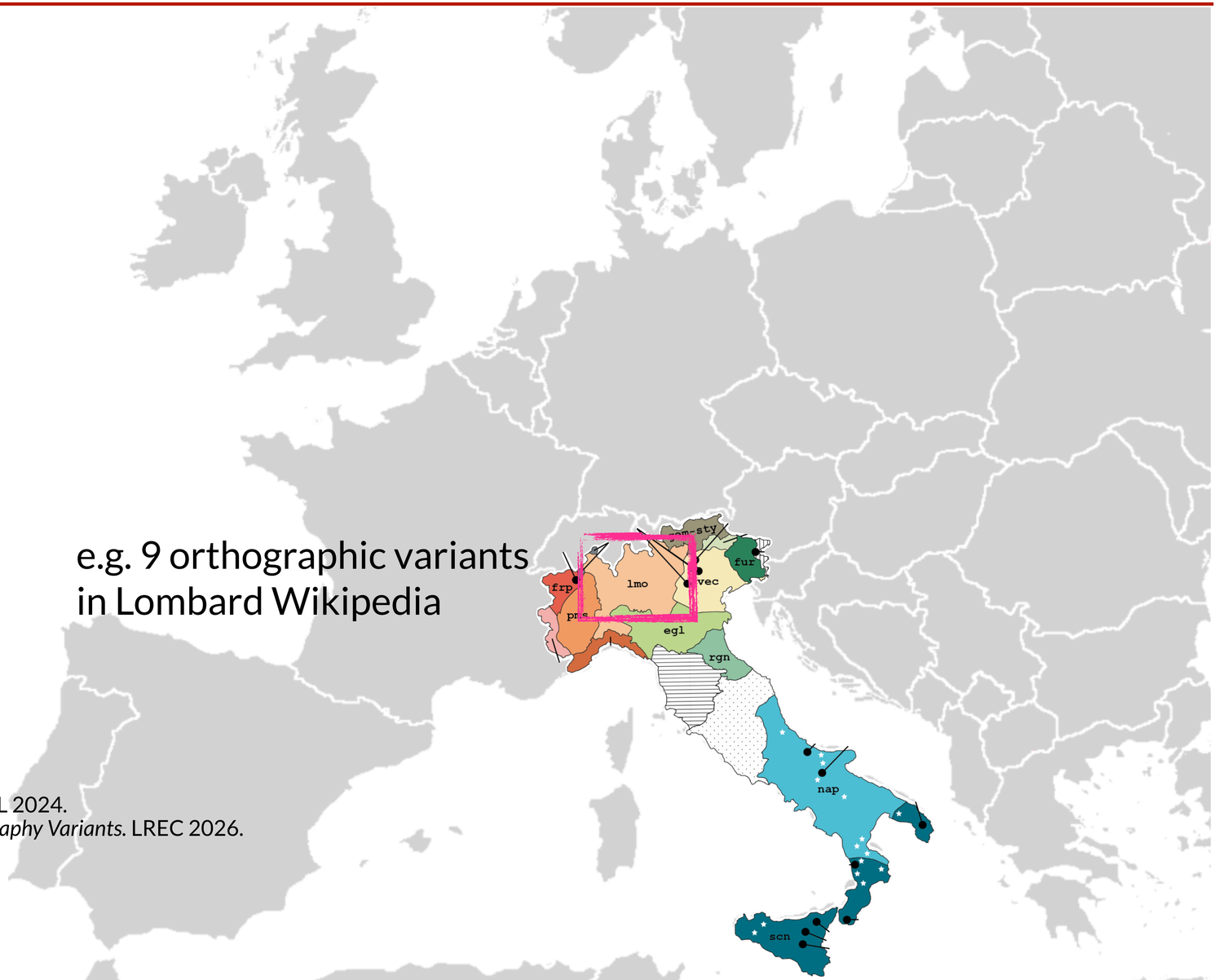
Kurdish

Amount of training data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for the **XLM-R language model**.

https://arxiv.org/pdf/1911.02116.pdf

Language Variation is Everywhere ❤️

- The spectrum of variation
  - Examples:
    - Italian

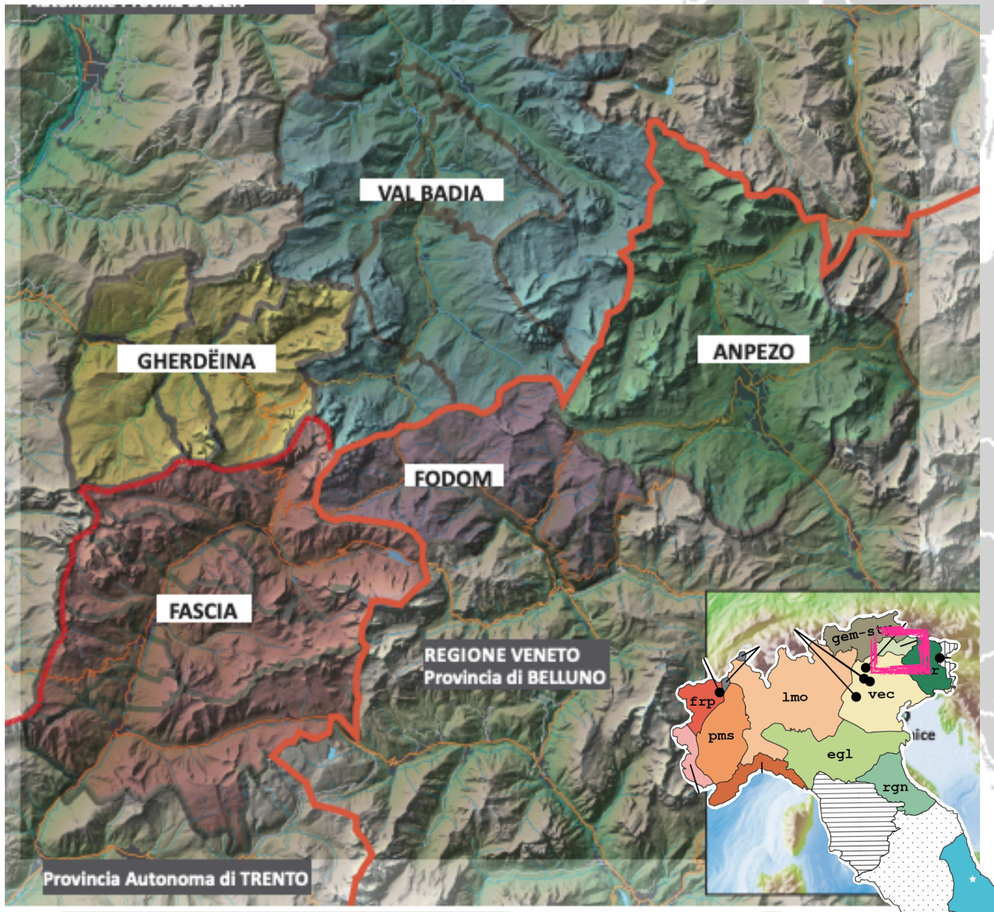


e.g. 9 orthographic variants
in Lombard Wikipedia

Ramponi. *Language Varieties of Italy: Technology Challenges and Opportunities*. TACL 2024.
Signoroni & Rychly. *LombardoGraphia: Automatic Classification of Lombard Orthography Variants*. LREC 2026.

Language Variation is Everywhere ❤️

- The spectrum of variation
  - Examples:
    - Italian



Ladin (Rhaeto-Romance), 5 varieties

Ramponi. *Language Varieties of Italy: Technology Challenges and Opportunities*. TACL 2024.
Signoroni & Rychly. *LombardoGraphia: Automatic Classification of Lombard Orthography Variants*. LREC 2026.
Frontull et al. 2025. Bringing Ladin to FLORES+. In WMT 2025.

Language Variation is Everywhere ❤️

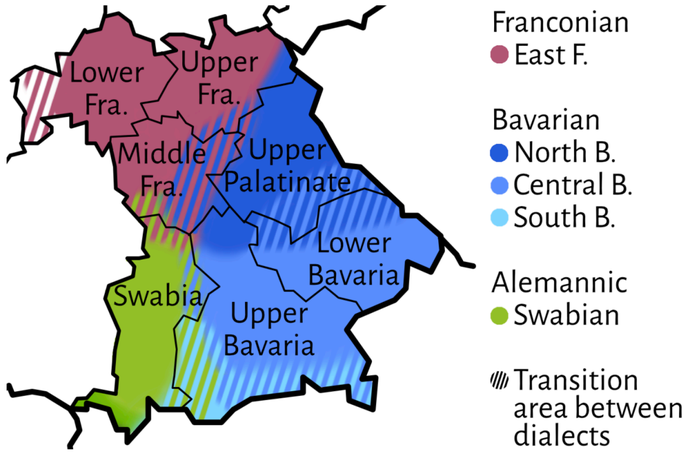


Figure 1: *The dialects and administrative regions included in Betthupferl (dialect division after [5]).*

- **Opportunity**: Away from ***Monoliths*** to embracing the Continuum of Variation at the ❤️ of NLP

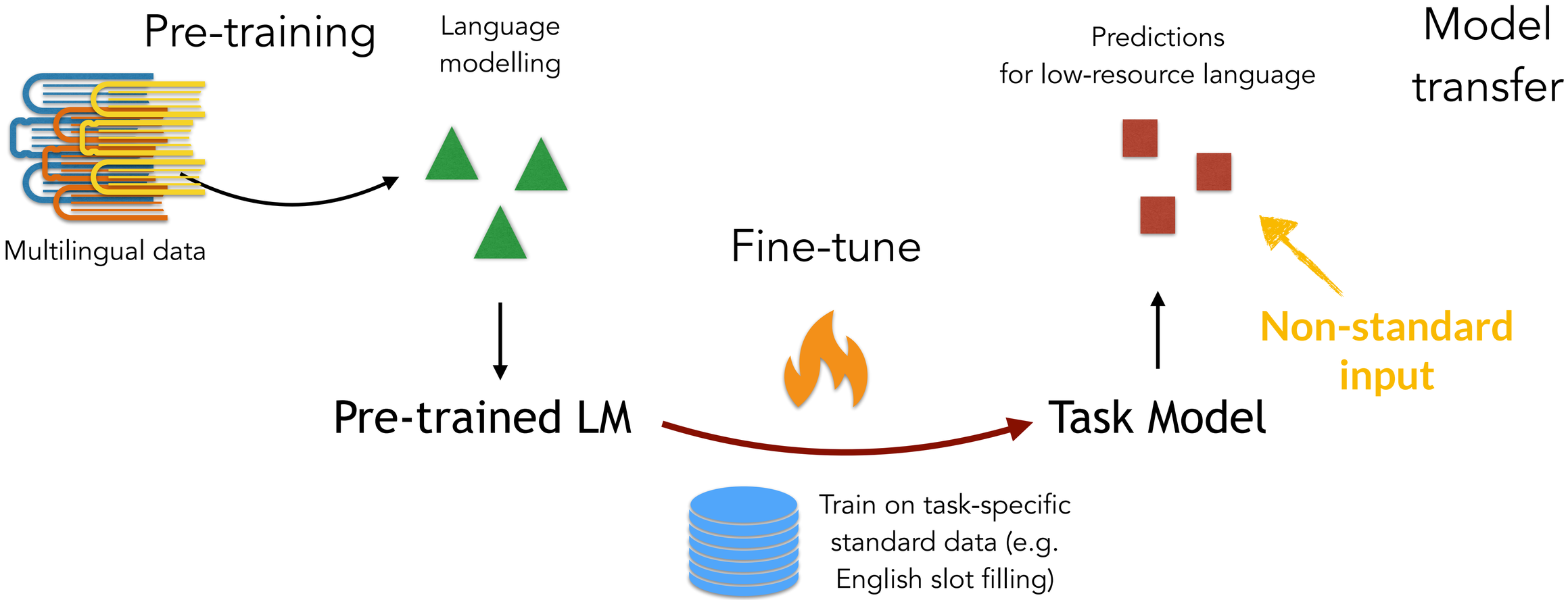
Ramponi. *Language Varieties of Italy: Technology Challenges and Opportunities*. TACL 2024.
Signoroni & Rychly. *LombardoGraphia: Automatic Classification of Lombard Orthography Variants*. LREC 2026
Frontull et al. 2025. Bringing Ladin to FLORES+. In WMT 2025.
Blaschke, Kovačić, Peng, Schütze, Plank. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. LREC-COLING 2024.
Blaschke, Winkler, Förster, Wenger-Glemser, Plank. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. Interspeech 2025.

Time to go **beyond the standard**
&
embrace dialectal variation at the ❤️ of NLP

What can we do about it and what are the challenges?

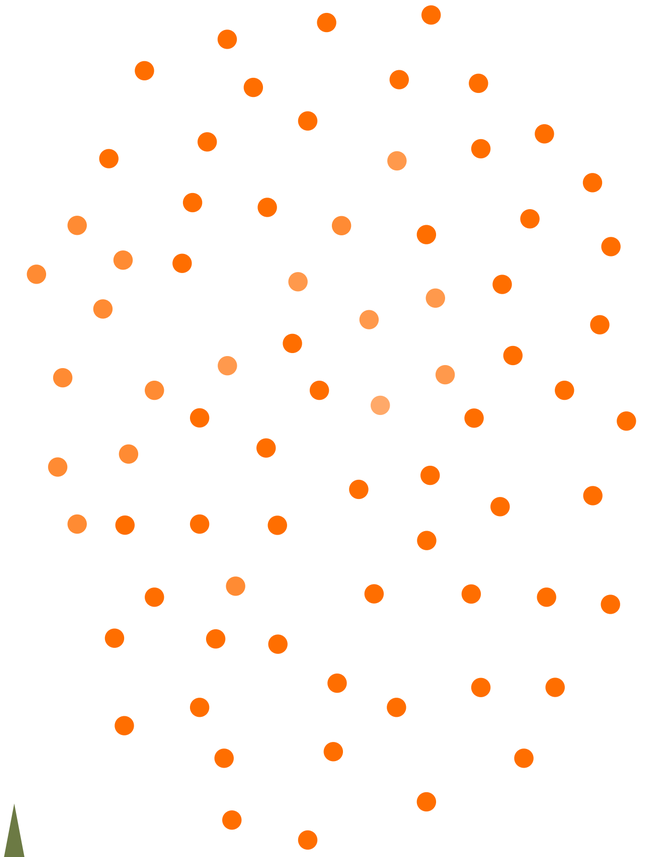
Cross-lingual Transfer: Training on a standard high-res. language

- Direct transfer: Applying a multilingual model trained on a source language (e.g. high-resource like English) directly without further adaptation



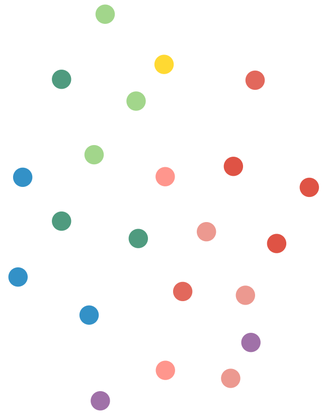
Dialect transfer (generalisation) gap

German
standard



Model

At test time: Models struggle when encountering unseen varieties (e.g. new structures)



South Tyrolean
Tyrolean
Austrian
Bavarian
Franconian
Swabian
Alemannic
Low Saxon

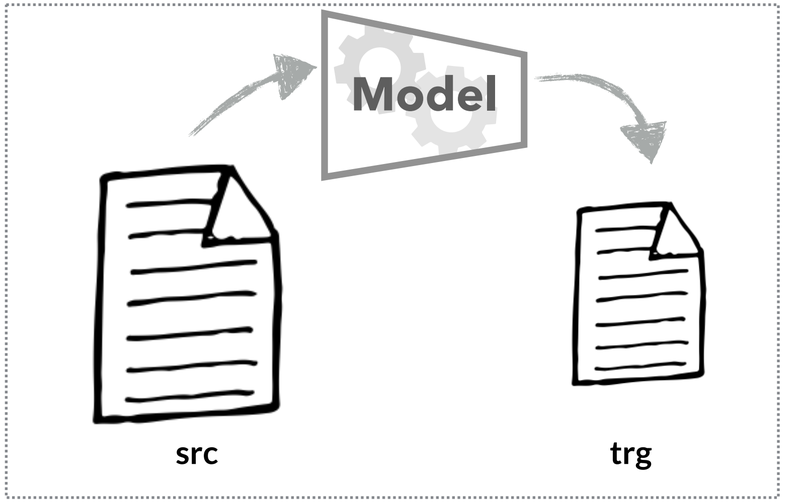
Training: since dialectal data is scarce, we train NLP models on standard variety data

From Transfer Learning to Dialect Transfer

- **Transfer Learning:** a generic approach to cross-lingual learning. Two broad approaches:

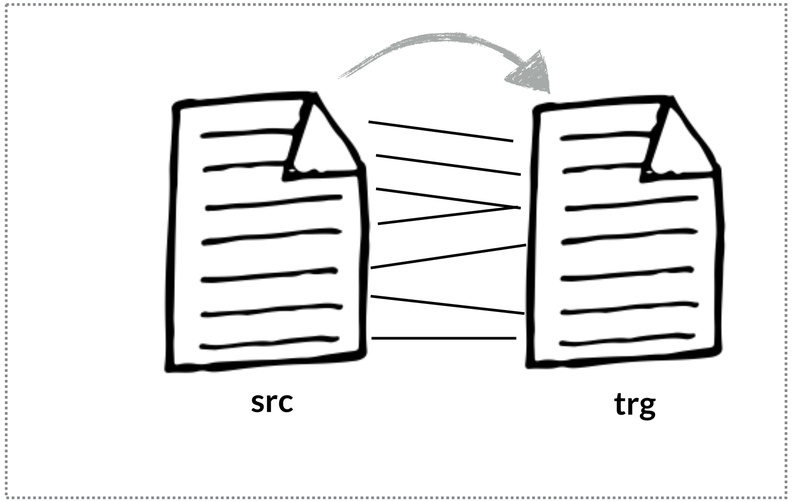
Model adaptation:

- Apply or modify the model (or its inputs)



Data adaptation:

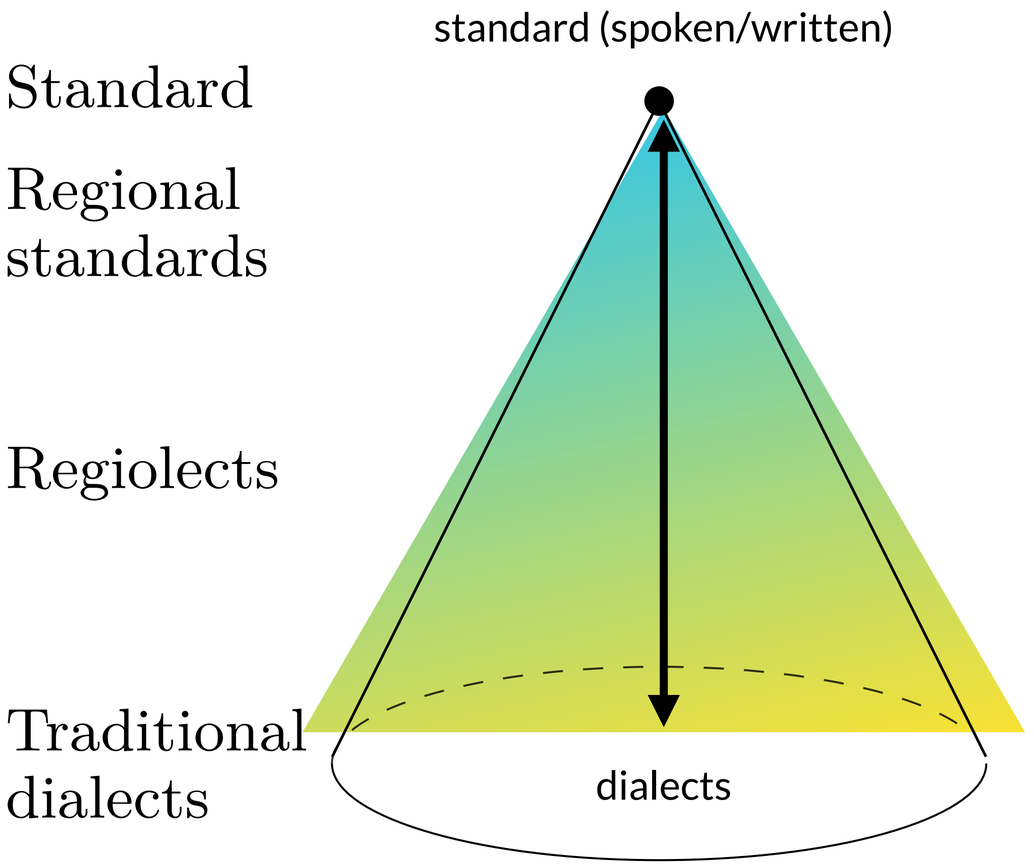
- Find ways to create new (proxy) training data



- **Question:** What makes transfer particularly challenging for dialects?

What makes dialects challenging?

- Linguistic differences
- Data challenges
- Representation & modelling challenges
- Evaluation obstacles



The cone model of dialect-standard variation based on Auer.

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.*

Differences (standard/dialect)

Usage situation

Morphosyntax

Lexicon

Spelling

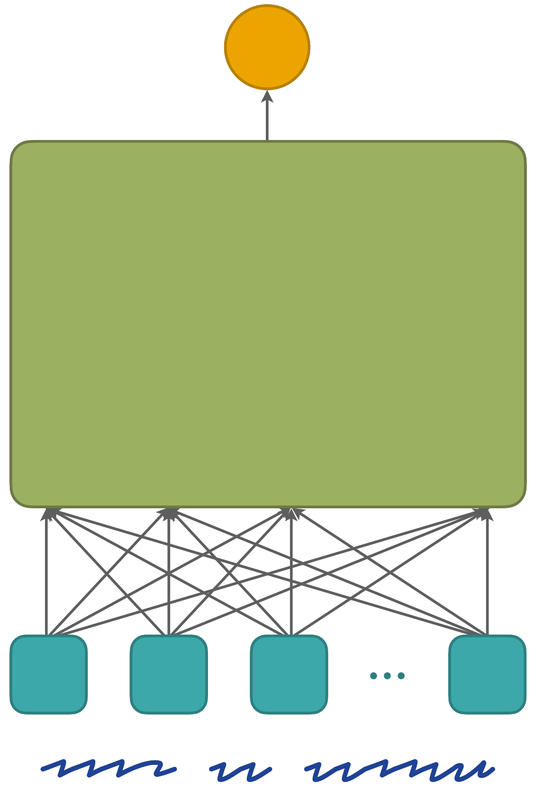
Phonetics & phonology

Output desiderata

Modelling

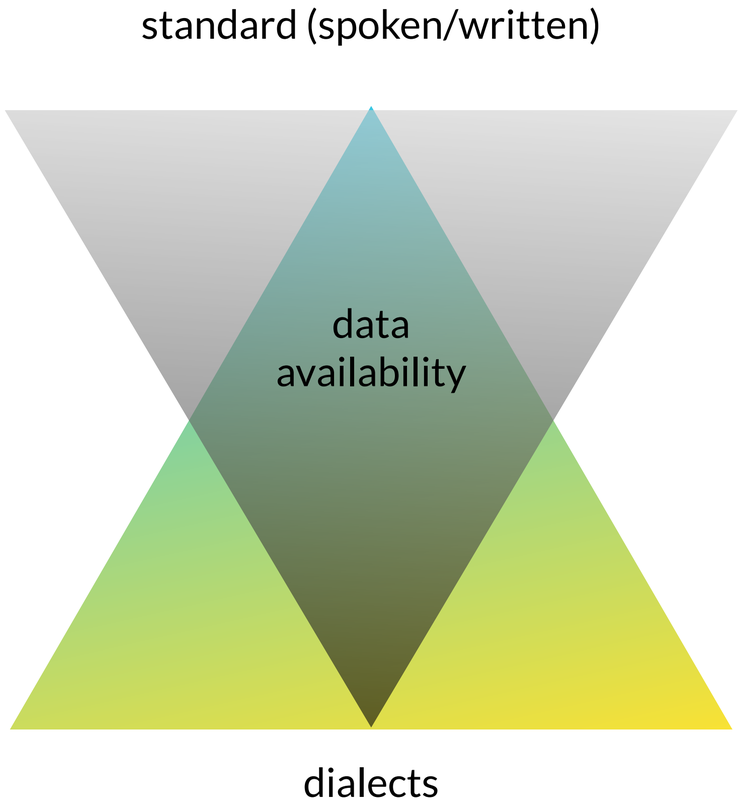
Input representations

Data scarcity



What makes dialects challenging?

- Linguistic differences
- Data challenges
- Representation & modelling challenges
- Evaluation obstacles



Lack of (accessible) datasets

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

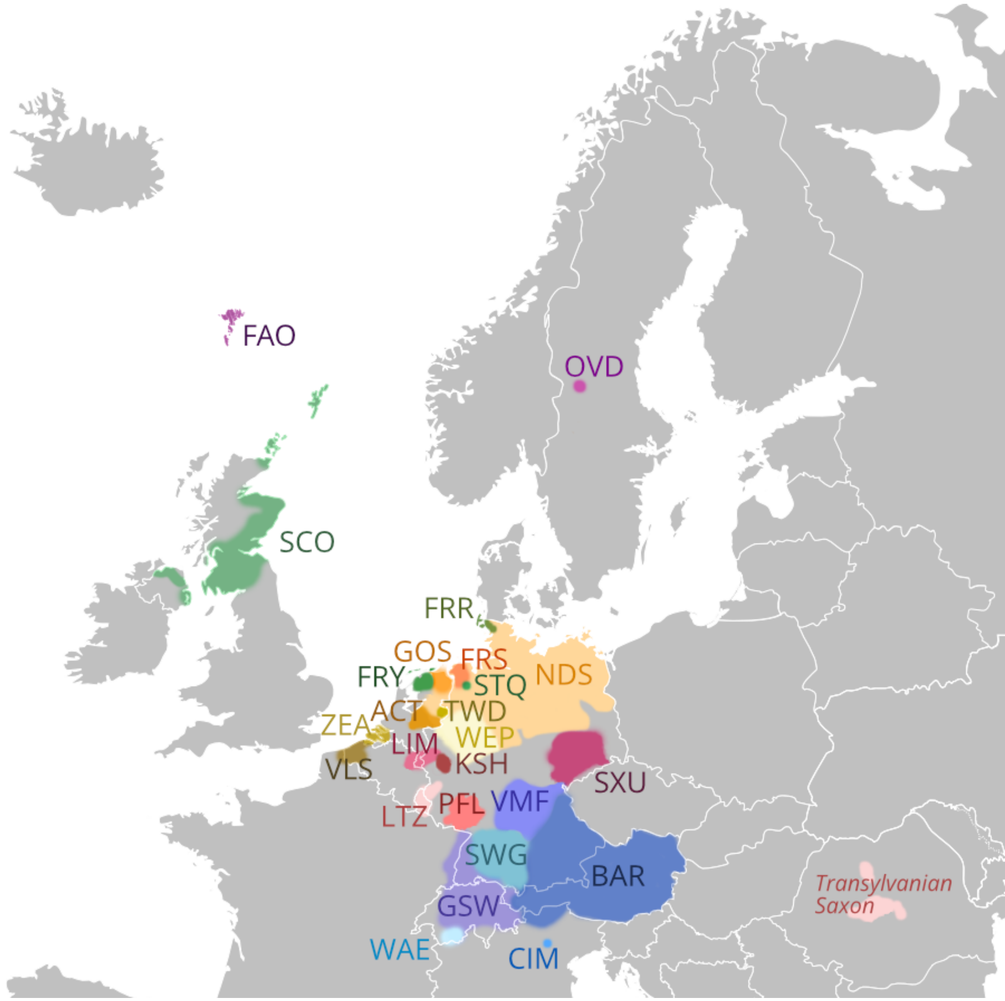
Verena Blaschke **Hinrich Schütze** **Barbara Plank**

140+ datasets
(~30 languages / dialect groups)

- Sentence level (or longer)
- Written or spoken
- With or without further annotations

- Findable; licenses allowing re-use
- Long-term storage + accessibility
- Two communities: variationists & NLP researchers
  — data exchange

github.com/mainlp/germanic-lrl-corpora



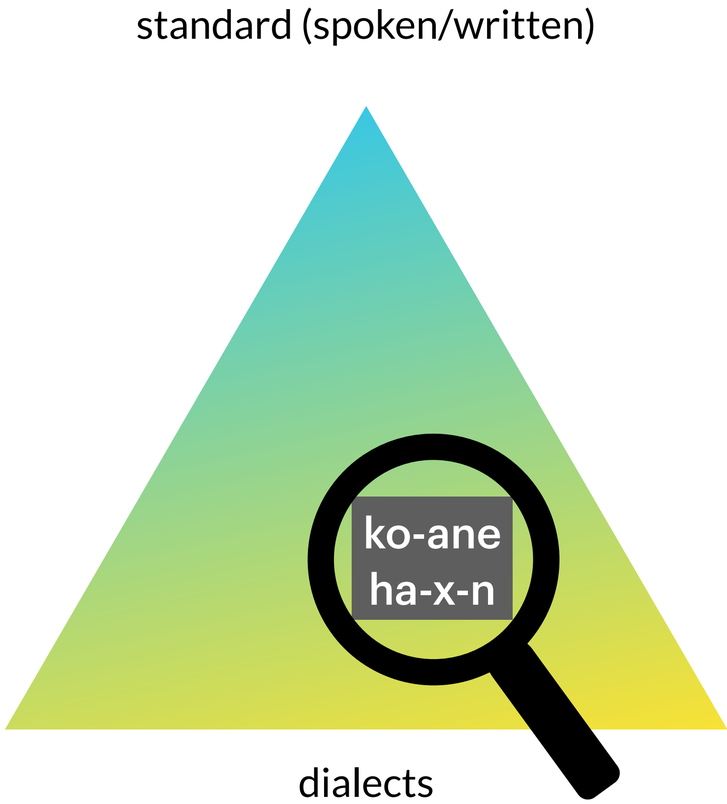
Many possible written representations

Written representations aren't necessarily comparable across datasets
(or always well-documented)!

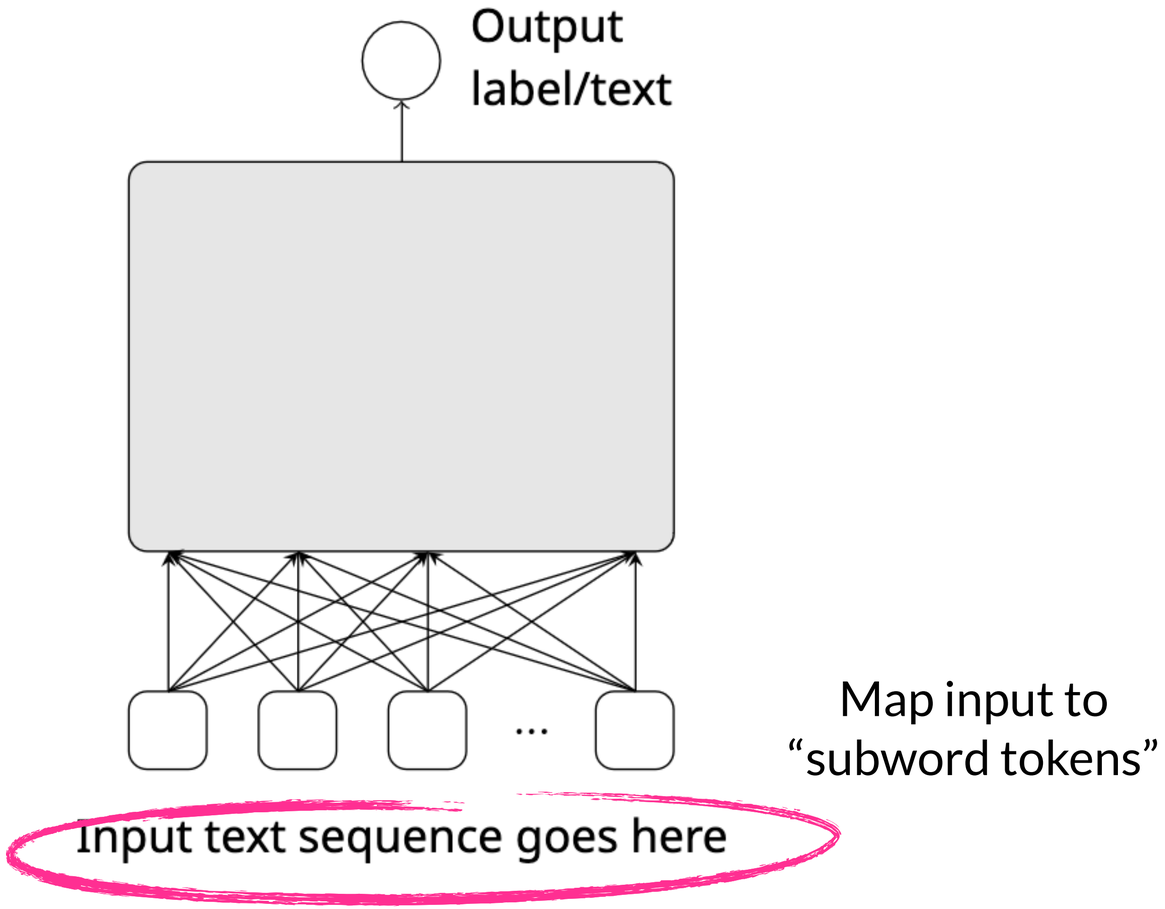
Ad-hoc spelling	I	mechad	wissn,	wia	hoas	dass	des	weard.
	I	mechat	wissen	wia	hoaß	das	des	weat.
	…							
	I	*want*	*know*	*how*	*hot*	*that*	*it*	*becomes*
Normalized	Ich	möchte	wissen,	wie	heiß	dass	das	wird.
	Ich	möchte	wissen,	wie	heiß		es	wird.
Phonemic transcriptions	iː	mɛçɐt	vɪsn	vɪɐ	hɔas	das	des	vɛɐt.
	I	mächad	wissn,	wia	hoàs	dass	des	weard.

What makes dialects challenging?

- Linguistic differences
- Data challenges
- Representation & modelling challenges
- Evaluation obstacles



What's the input representation?



Tokenization and non-standardness

- State-of-the-art LMs are based on subwords (not characters). This representation is sensitive to slight surface variations: minor changes will lead to different segmentations & representations:

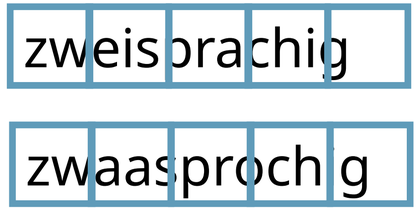
German	zweisprachig *bilingual*	zwei sprach ig *two language y*
Bavarian	zwaasprochig	z wa as pro chi g
	zwaspråchig	z was pr å chi g
	zwoasprochig	z wo as pro chi g
	zwasprochig	z was pr pro chi g
	…	

Tokenization and non-standardness — approaches

German	zweisprachig	`zwei` `sprach` `ig`
Bavarian	zwaasprochig	`z` `wa` `as` `pro` `chi` `g`
	zwaspråchig	`z` `was` `pr` `å` `chi` `g`
	…	

Possible approaches

- Normalize input (translate into standard) — but might lose valuable information
- Also change the tokenization of the training data *(Aepli & Sennrich, 2022; Blaschke et al., 2023)*
- Use models without tokens, e.g., apply computer vision methods to screenshots of text
  (Rust et al., 2023; Muñoz-Ortiz et al., 2025)



Written vs. spoken inputs

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

Another token-free approach: directly work with audio data if the task allows!

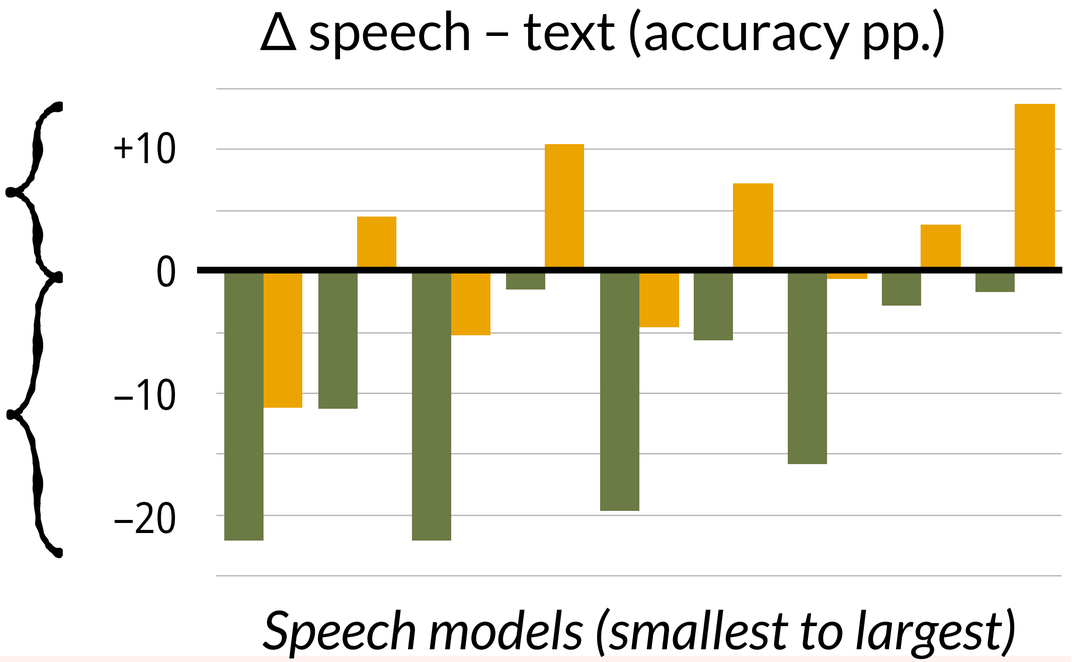
- Continuous input representations (no discrete tokens)

Sentence classification with parallel data

- Text vs. speech
- German vs. dialect

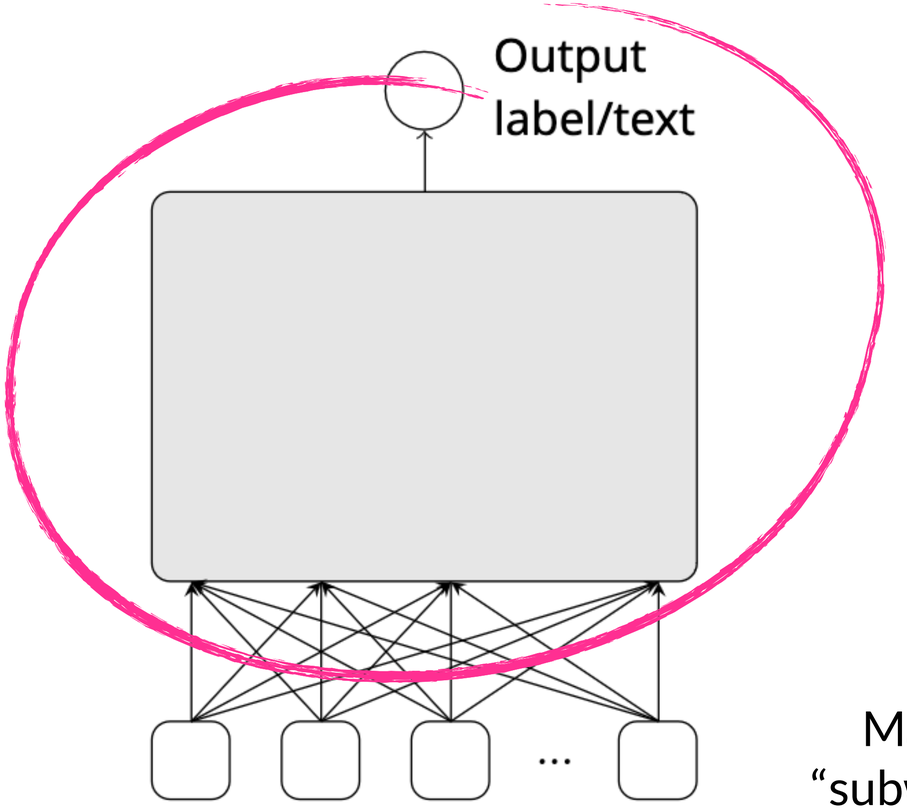
Dialect (often) speech > text

Standard German text > speech



→ Different trends across modalities for standard (German) vs. non-standard (Bavarian) varieties!

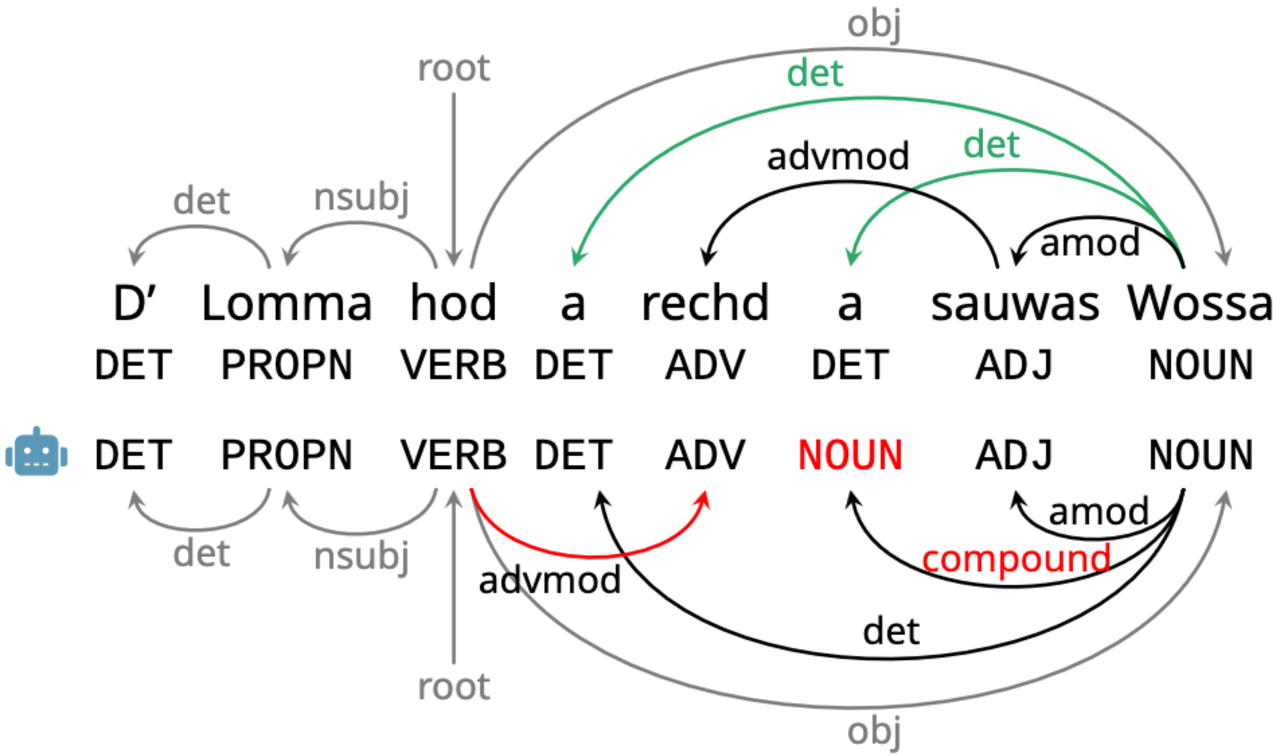
Modelling challenges



Map input to
“subword tokens”

Lack of robustness to morphosyntactic variation

First Bavarian treebank
→ State-of-the-art parsers struggle with Bavarian-specific constructions



“The river Lammer has fairly clean water.”

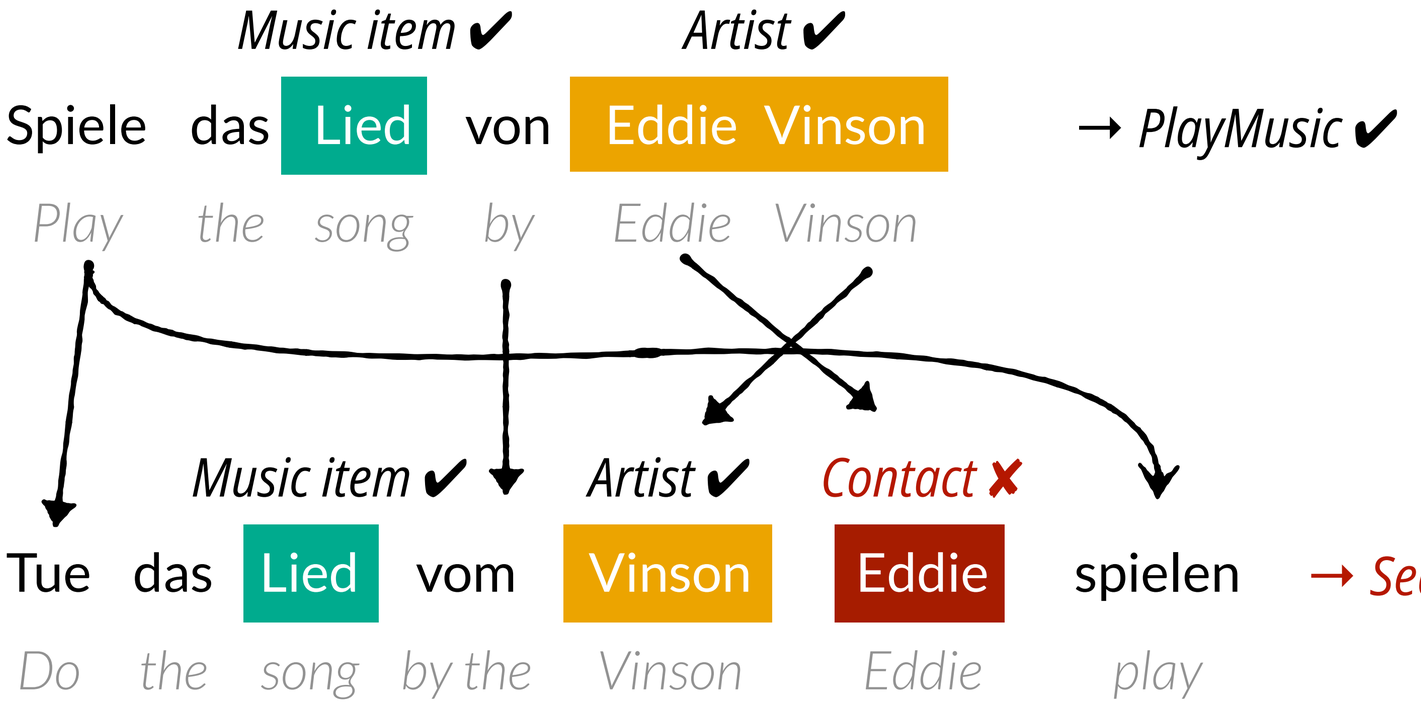
MaiBaam:
A Multi-Dialectal Bavarian Universal Dependency Treebank

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, Barbara Plank

Lack of robustness to morphosyntactic variation

Also a problem for applied tasks, like identifying user intents and relevant slots

Test suite of morphosyntactic perturbations for German



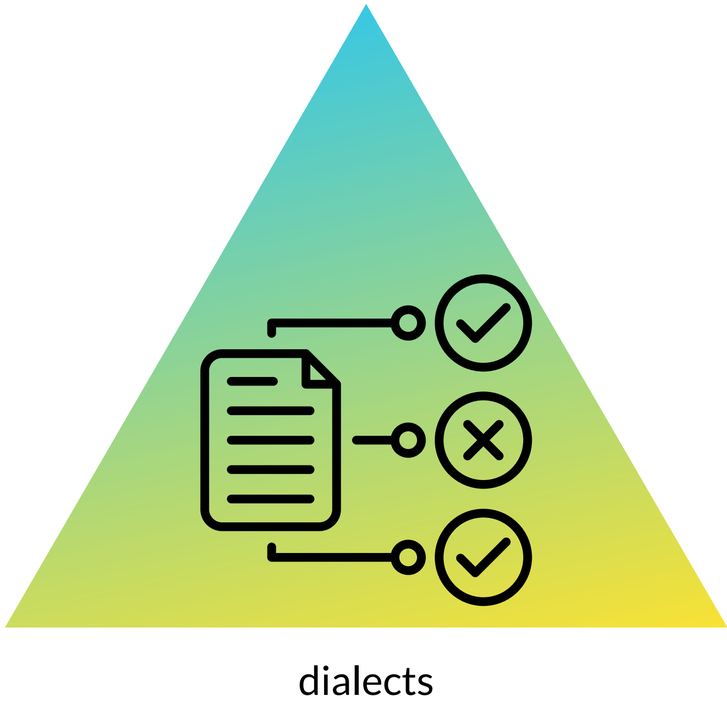
- Impacts both slot & intent detection
- Changing the word order especially affects slot filling

Exploring the Robustness of Task-oriented Dialogue Systems for Colloquial German Varieties

Ekaterina Artemova * Verena Blaschke Barbara Plank

What makes dialects challenging?

- Linguistic differences
- Data challenges
- Representation & modelling challenges
- Evaluation obstacles



Evaluation challenges

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]

ASR challenges: linguistic differences (pronunciation, lexis, morphosyntax)

But also: How to evaluate the output without a single gold standard?

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

same as German reference — different, but acceptable — different & wrong

Best model (relative to German references):
Correlations of automatic scores with human judgement: $0.48 \leq |\rho| \leq 0.63$

Evaluation challenges

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]

ASR challenges: linguistic differences (pronunciation, lexis, morphosyntax)

But also: How to evaluate the output without a single gold standard?

In order to build better dialect transcription systems, we need to
- improve our automatic metrics!
- know what output style (level of normalization) is the most relevant for, e.g, dialectologists!

same as German reference

different, but acceptable

different & wrong

Best model (relative to German references):
Correlations of automatic scores with human judgement: $0.48 \leq |\rho| \leq 0.63$

Conclusion

- By grounding our discussion on Bavarian, we illustrate broader challenges in dialect technology and propose pathways towards inclusive, variation-aware models
- Collaborations between linguists & NLPers to improve dialect NLP & make NLP more useful for dialectologists
  - Accessible/open, well-documented data
  - Better evaluation of NLP systems, also with dialectologists' use-cases in mind

Slack for
variation-aware NLP