

Natural language processing for dialects and low-resource language varieties

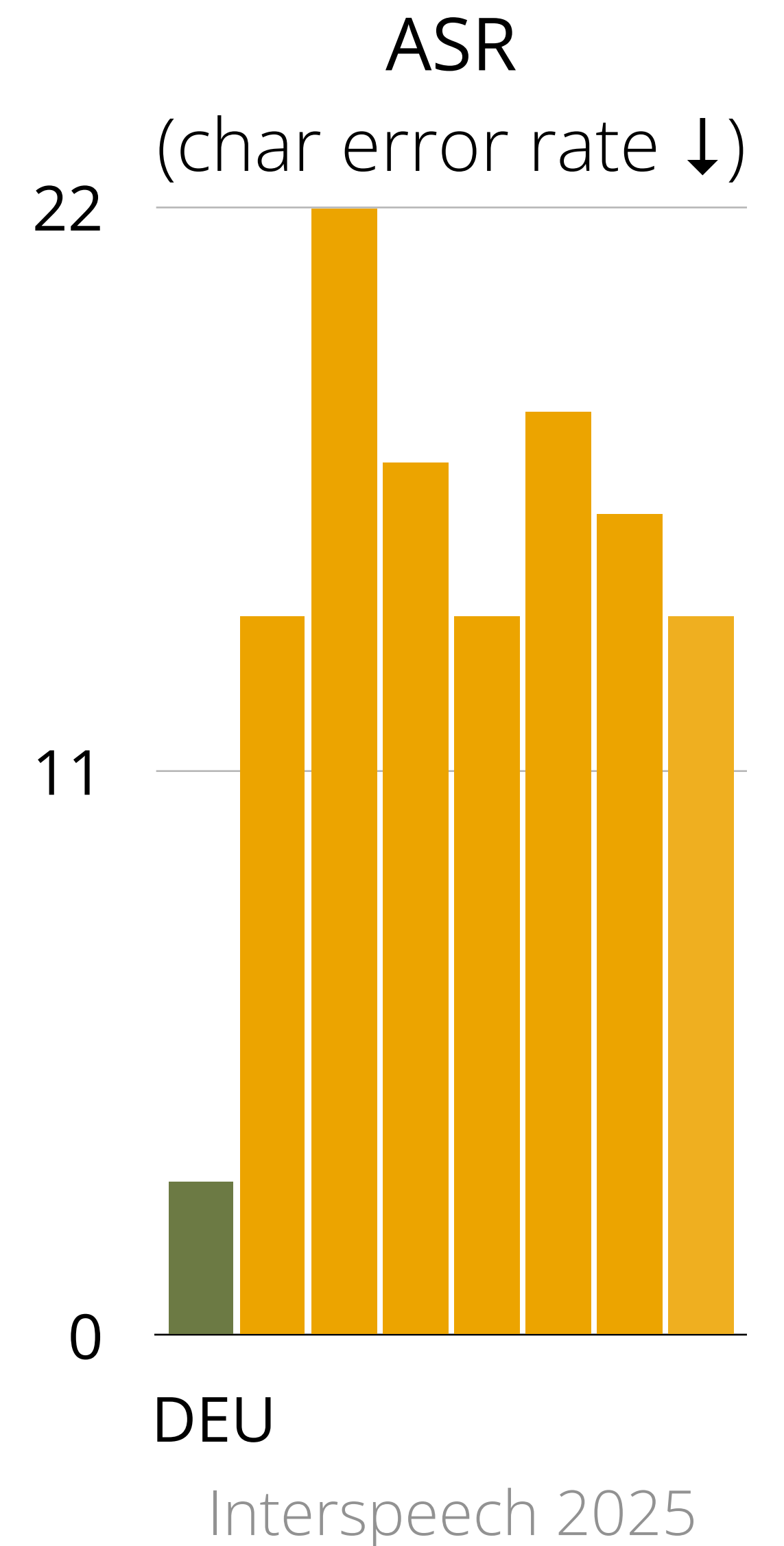
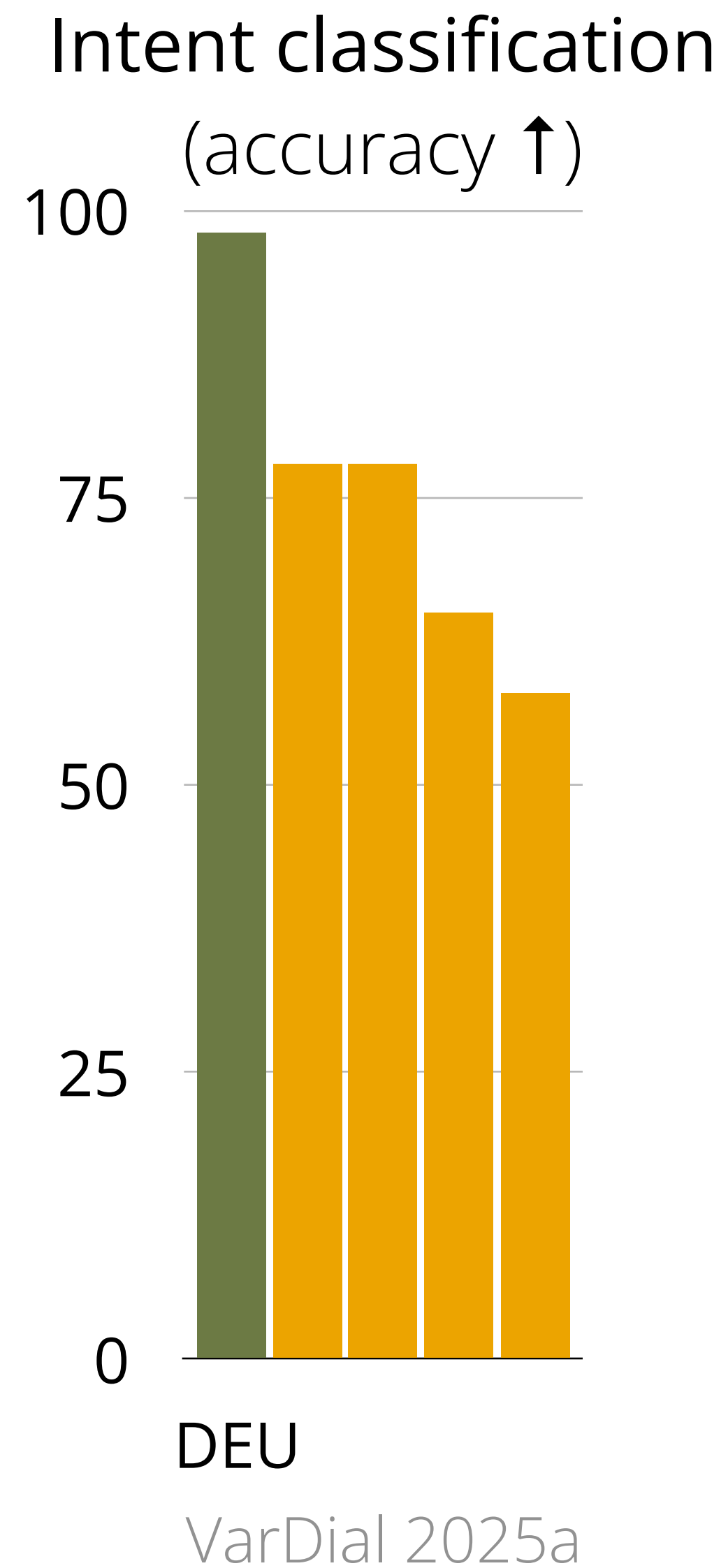
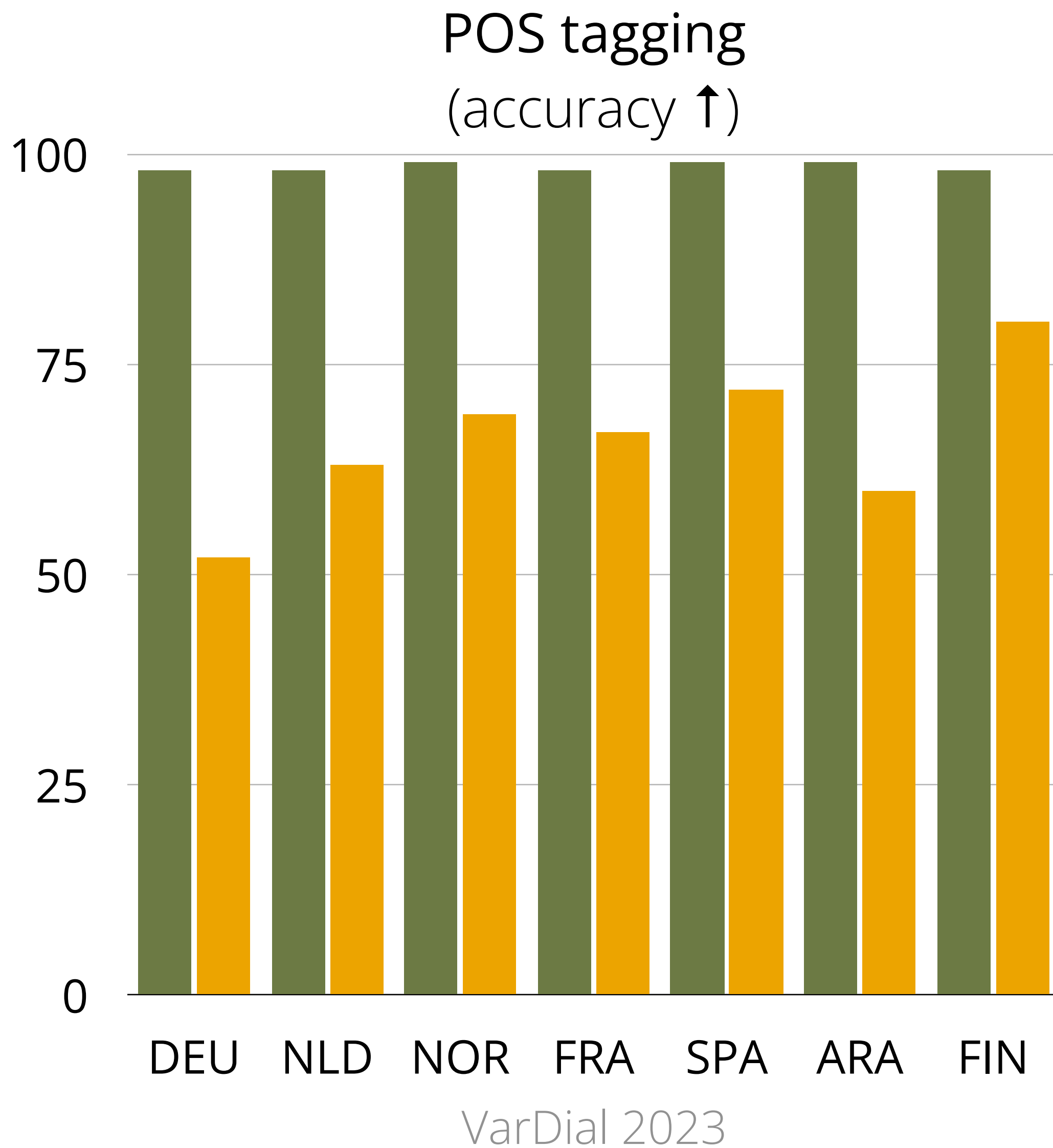
Verena Blaschke

LMU Munich, MCML | June 25, 2026

Ph.D. committee *Prof. Dr. Barbara Plank*
 Prof. Dr. Hinrich Schütze
 Prof. Dr. Dirk Hovy

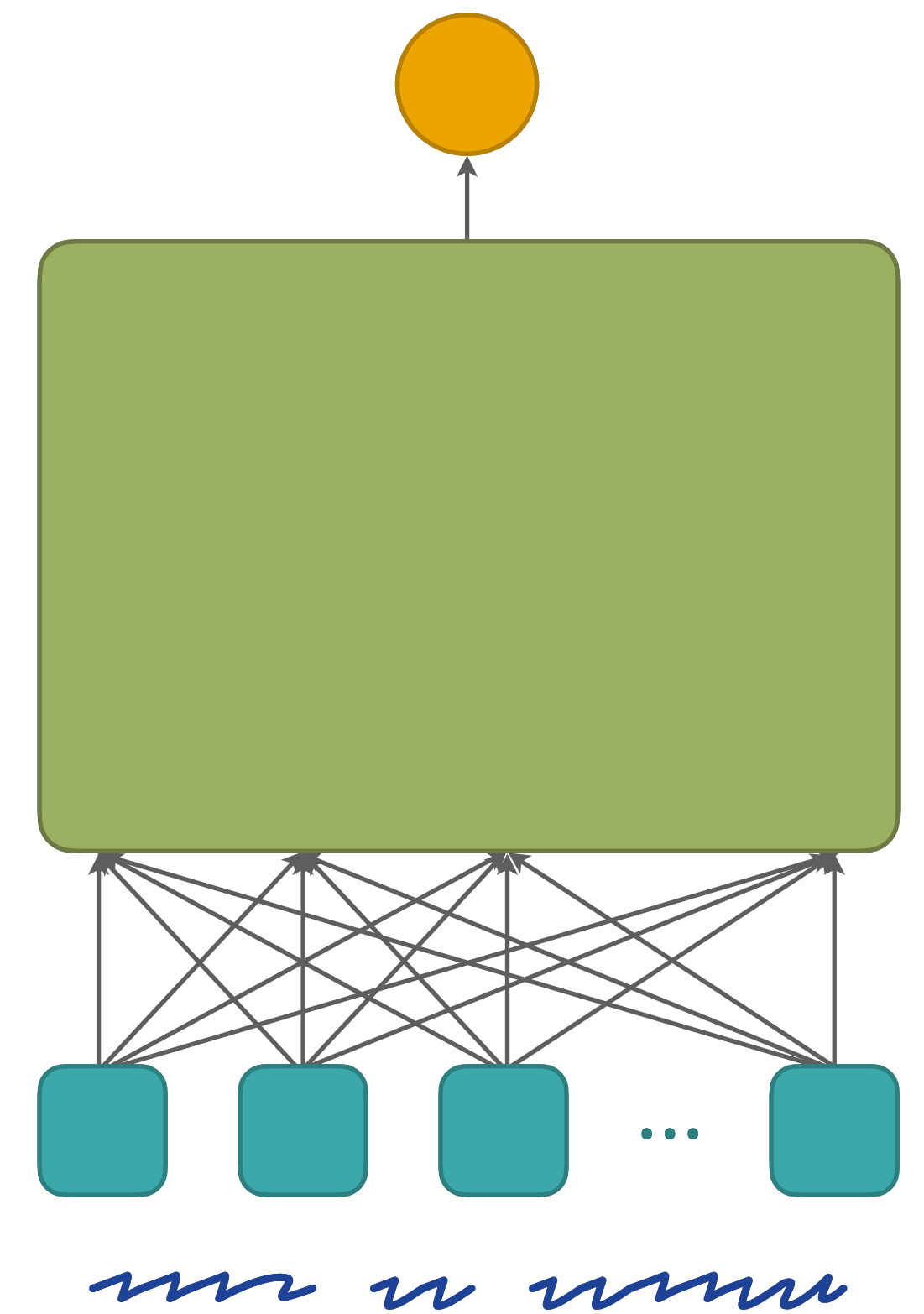
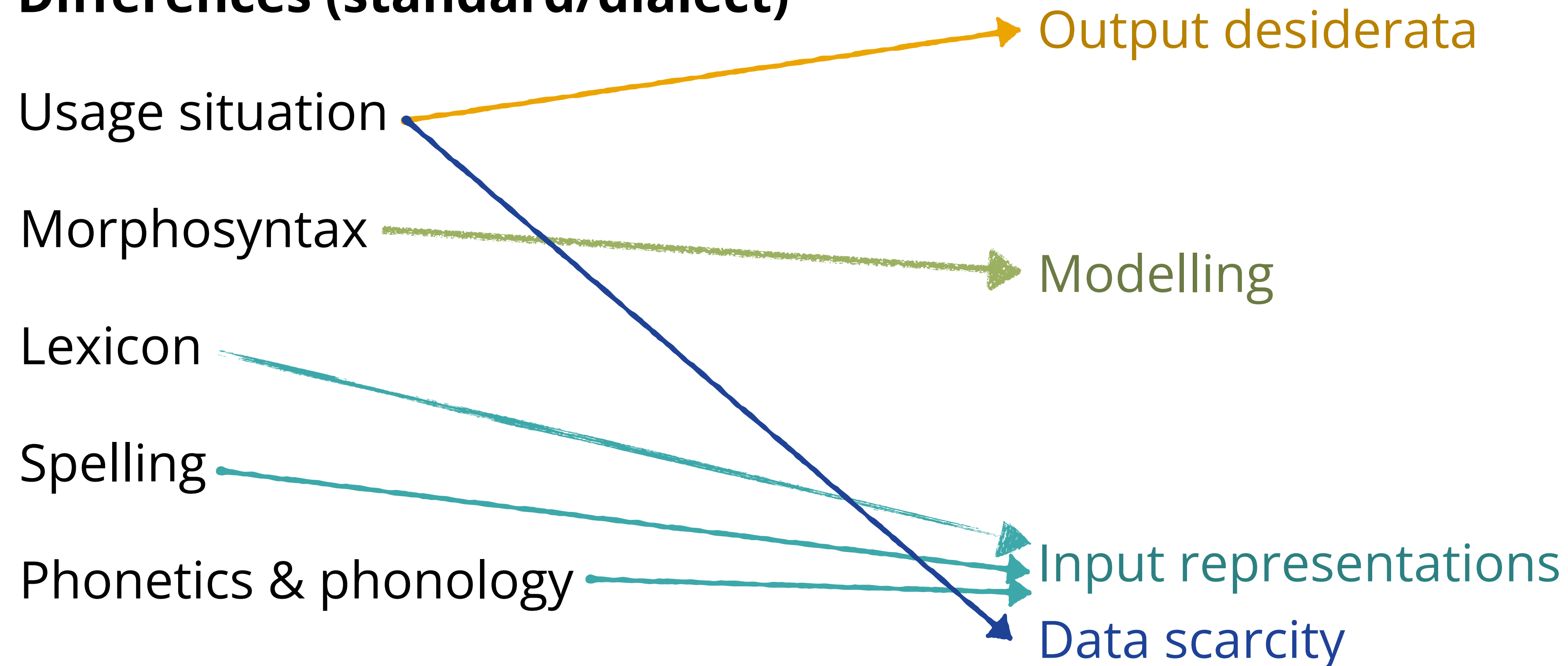


Performance on **standard languages** and related **non-standard varieties**



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

Differences (standard/dialect)



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

Verena Blaschke^{1,2}, Miriam Winkler¹, Constantin Förster³, Gabriele Wenger-Glemser³, Barbara

MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank

Verena Blaschke, Barbara Kovačić, S
Hinrich Schütze, Barbara Plank

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

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

Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages

Improving Dialectal Slot and Intent Detection with Auxiliary Tasks: A Multi-Dialectal Bavarian Case Study

Xaver Maria Krückl*, Verena Blaschke

Add Noise, Tasks, or Layers? MaiNLP at the VarDial 2025 Shared Task on Norwegian Dialectal Slot and Intent Detection

Analyzing the Effect of Linguistic Similarity on Cross-Lingual Transfer: Tasks and Experimental Setups Matter

Verena Blaschke^{1,2,*}, Masha Fedze

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

Verena Blaschke

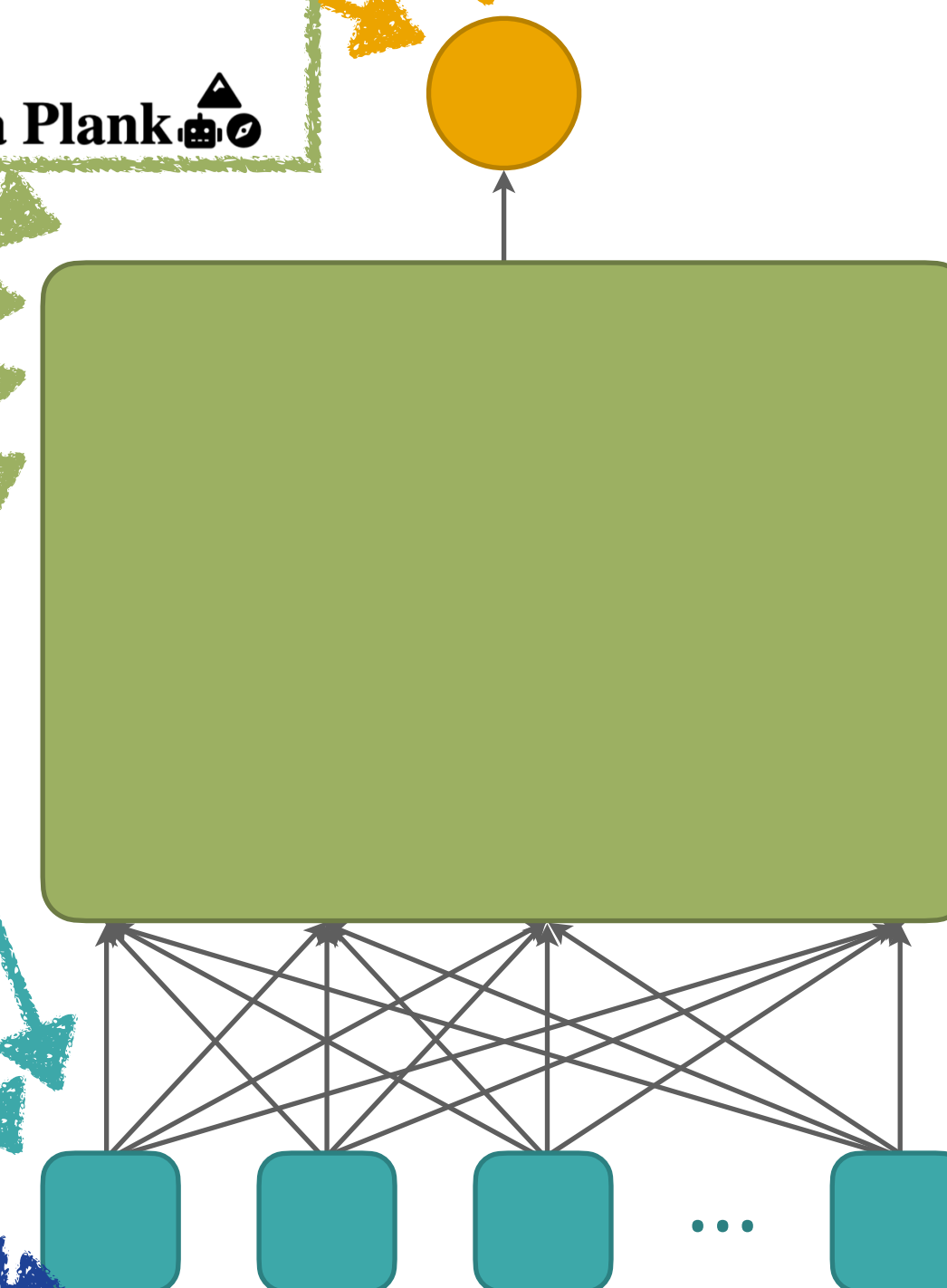
Hinrich Schütze

Barbara Plank

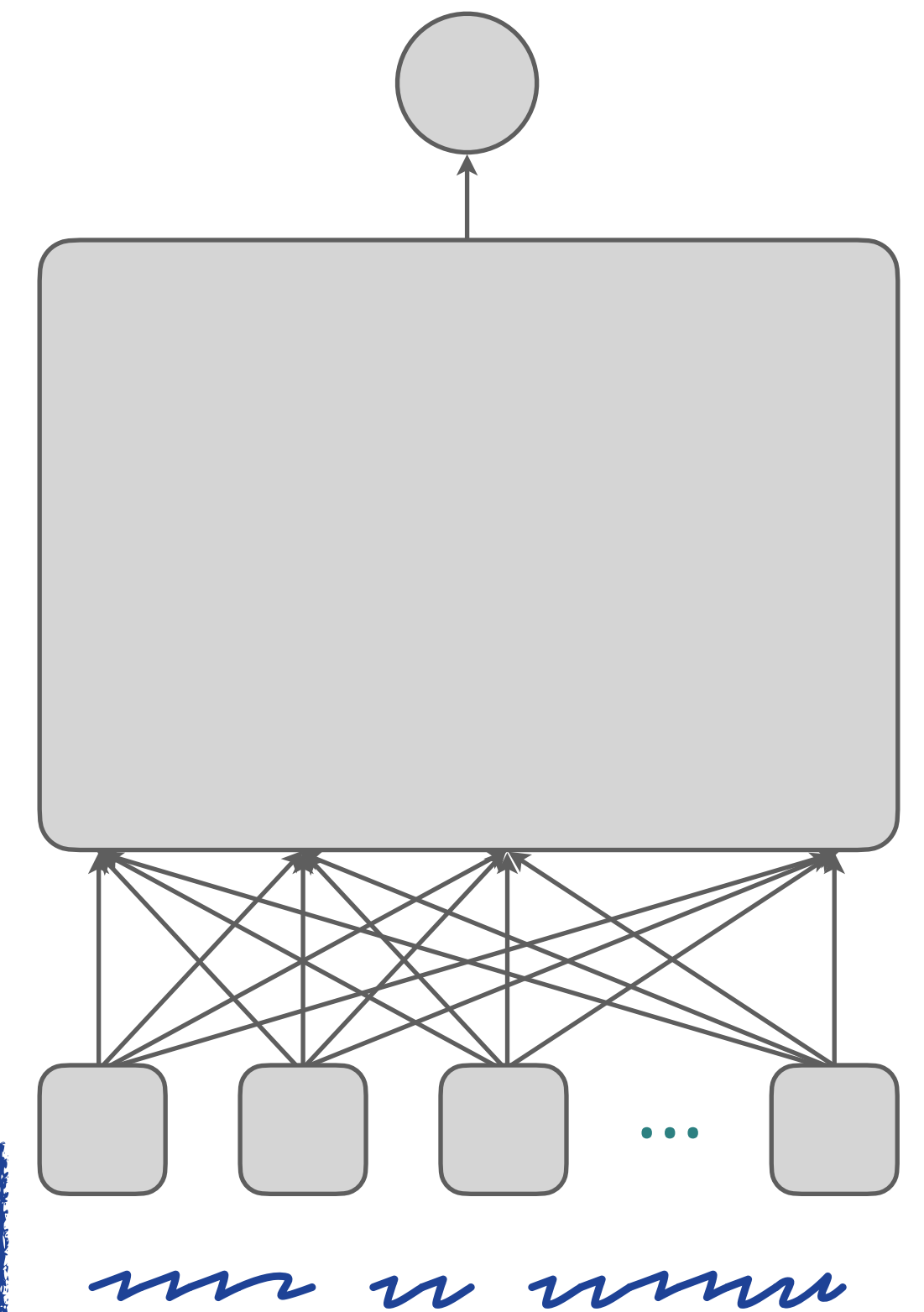
What Do Dialect Speakers Want?

A Survey of Attitudes Towards Language Technology for German Dialects

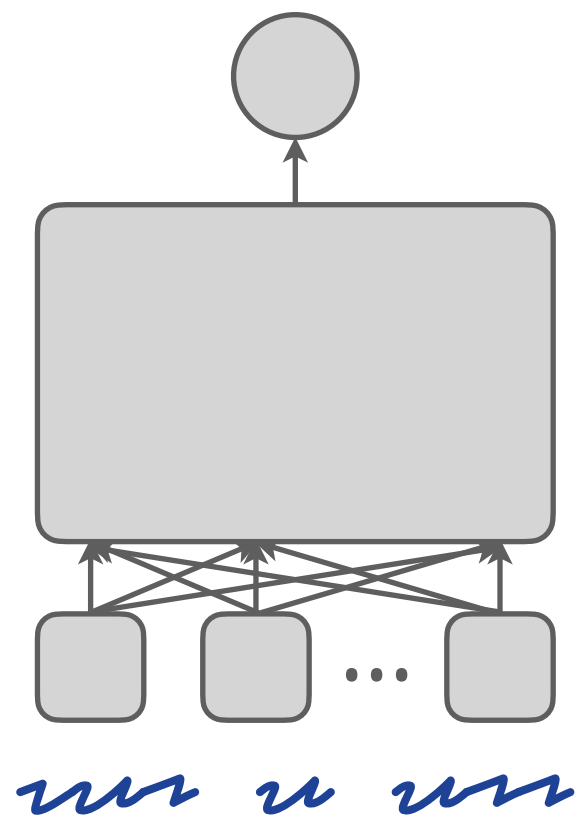
Verena Blaschke, Christoph Purschke, Hinrich Schütze, Barbara Plank



Data scarcity

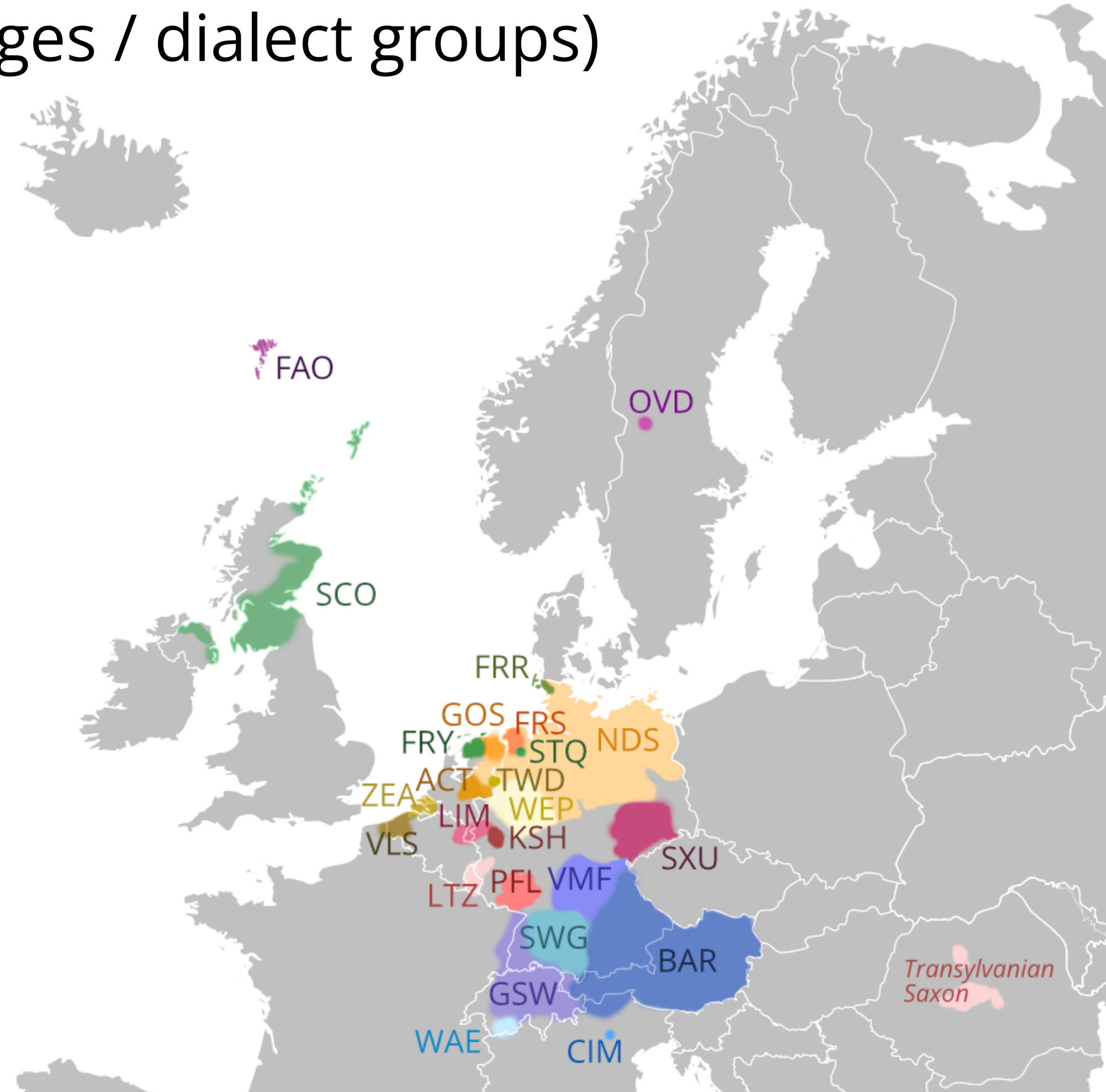


**A Survey of Corpora for
Germanic Low-Resource Languages and Dialects**

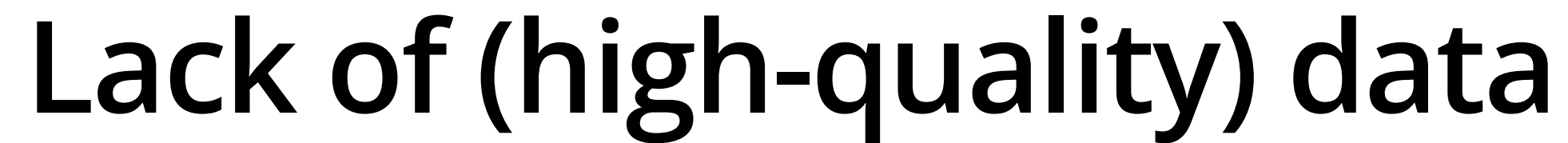


Lack of (high-quality) data

83 datasets
(~30 languages / dialect groups)



Language		Curated datasets				Uncurated text
		Annotations		Rep.		
North Germanic						
FAO	Faroese	D	M		  A	✓
NOR	(non-std.) Norwegian	D	M		   ¶	
OVD	Elfdalian	D			A ¶	
SWE	(non-std.) Swedish	D			 ¶	
DAN	(non-std.) Danish	D			unknown	✓
Anglo-Frisian						
SCO	Scots		M		 A ¶	✓
ENG	(non-std.) English	D	M		 ¶	
FRY	West Frisian	D	M		 A	✓
FRR	North Frisian					✓
STQ	Saterland Frisian					✓
Low German*						
NDS	Low Saxon	D	M	P	  A	✓
FRS	East Frisian Low Saxon					✓
GOS	Gronings			P		✓
TWD	Twents					✓
ACT	Achterhoeks					✓
WEP	Westphalian				 ¶	
Macro-Dutch						
VLS	West Flemish	D	M			✓
ZEA	Zeelandic					✓
Middle German						
LTZ	Luxembourgish				A	✓
KSH	Colognian					✓
LIM	Limburgish					✓
PFL	Palatine German					✓
PDC	Pennsylvania Dutch					✓
YID	Yiddish**		M		 	✓
SXU	Upper Saxon				 ¶	
Upper German						
DEU	(non-std.) German			P	  ¶	
VMF	East Franconian				 ¶	
BAR	Bavarian		M	S	P    ¶	✓
CIM	Cimbrian				 ¶	
SWG	Swabian					✓
GSW	Swiss Ger. & Alsatian	D	M	S	P    A ¶	✓
WAE	Walser	D		S	P 	✓



ges / dialect groups)

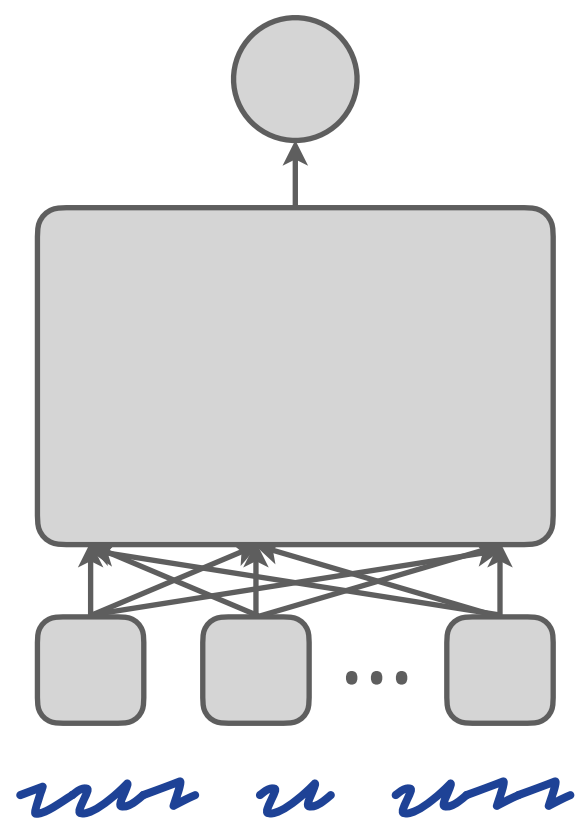
Sprache [Am Gwëntext werkeln]

"Beroda in technischn Frong
drod de mid grousa Mehrheit
des Gsetz üba de Rechtsstai
(CET)

“That's not Bavarian at all: [examples]”

— “I wouldn't agree that it's not Bavarian at all. But it needs to be fixed. Most of all, the genitive and preterite need to be replaced, and also some words. I'll help.”





New datasets for Bavarian

xSID de-muc (Munich Bavarian
slots/intents, VarDial 2025a)

Werds **de woch** **hoaß** wern?
 datetime weather/attribute
 → *weather/find*
Will it be hot this week?

xSID audio (spoken
German/Bavarian intents, ACL 2026)

 Wirds de Woch hoas?
 Wird es diese Woche heiß?
 → *weather/find*
Will it be hot this week?

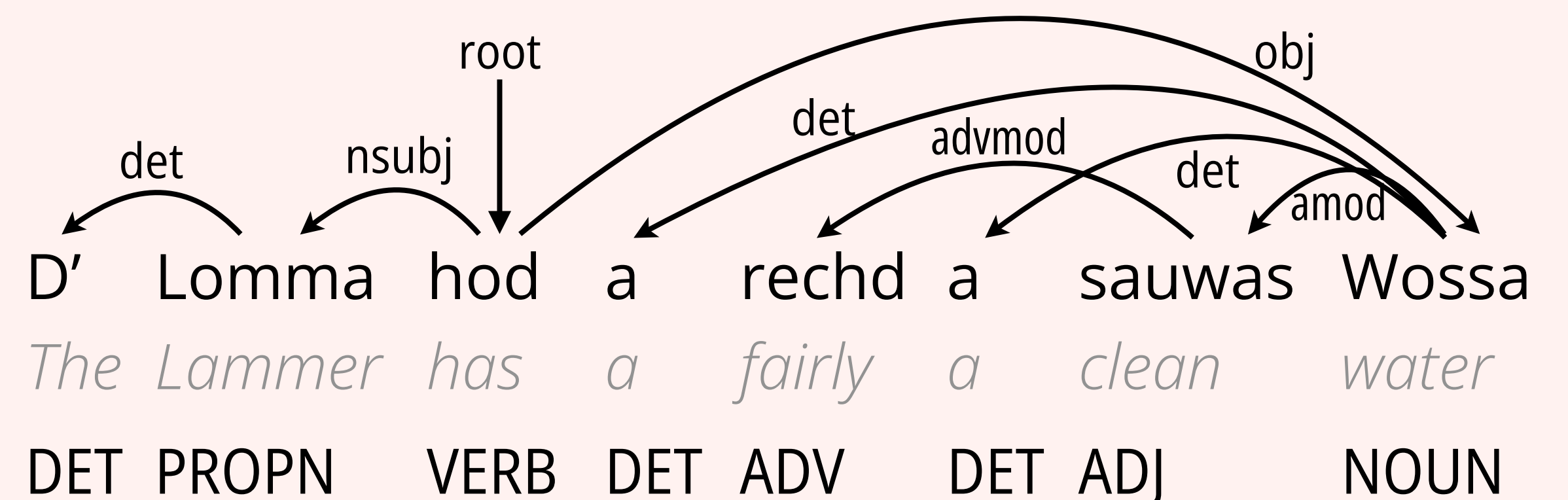
Betthupferl (Bavarian/Swabian/
Franconian ASR, Interspeech 2025)

No hodes Geldstüggle
am Eismo gem

Da hat sie das Geldstückchen
dem Eismann gegeben

She then gave the coin to the ice cream vendor

MaiBaam (1st Bavarian dependency
treebank, LREC-COLING 2024)



Input representations

**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

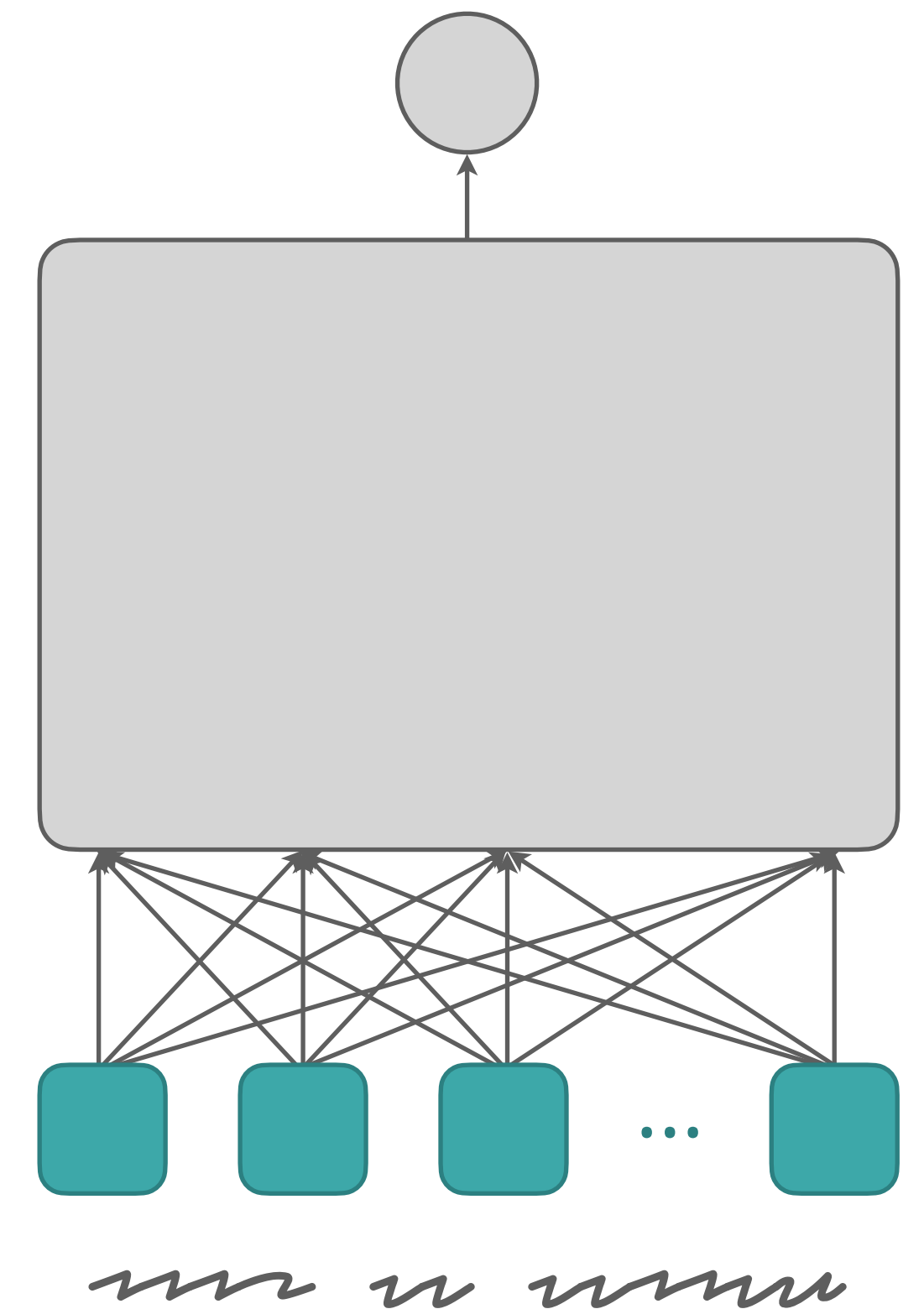
**Does Manipulating Tokenization Aid Cross-Lingual Transfer?
A Study on POS Tagging for Non-Standardized Languages**

Verena Blaschke

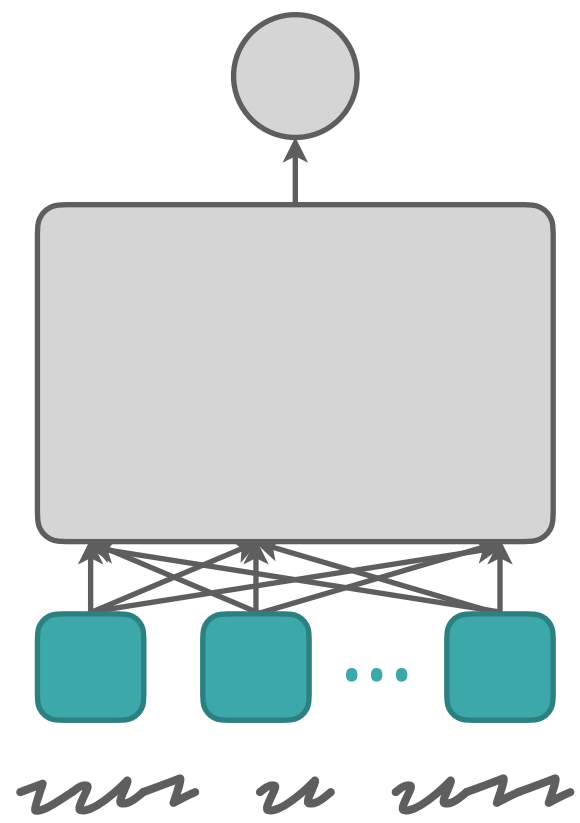
Hinrich Schütze

Barbara Plank

VarDial 2023



Spelling variation & subword tokenization



German

zweisprachig

bilingual

zwei sprach ig
two language y

Bavarian

zwaasprochig

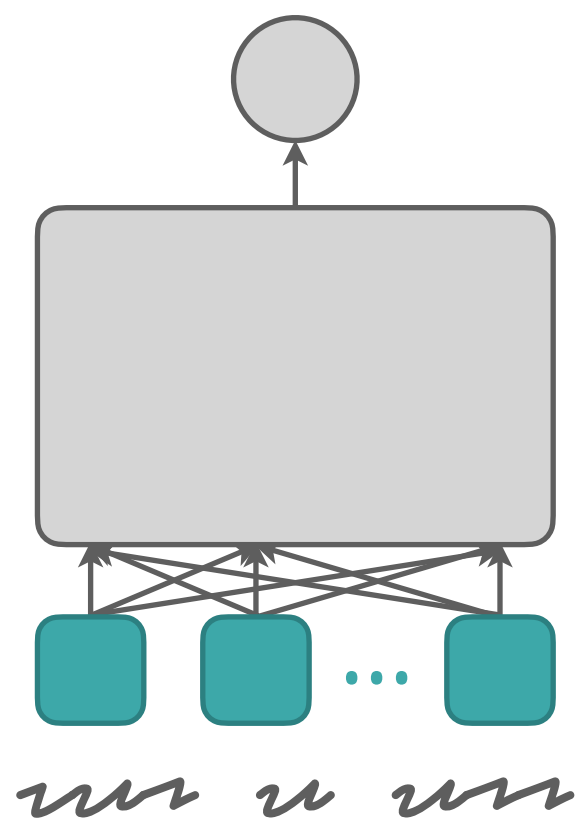
zwaspråchig

zwoasprochig

zwasprochig

...

z wa as pro chi g
z was pr å chi g
z wo as pro chi g
z was pr pro chi g



Spelling variation & subword tokenization

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

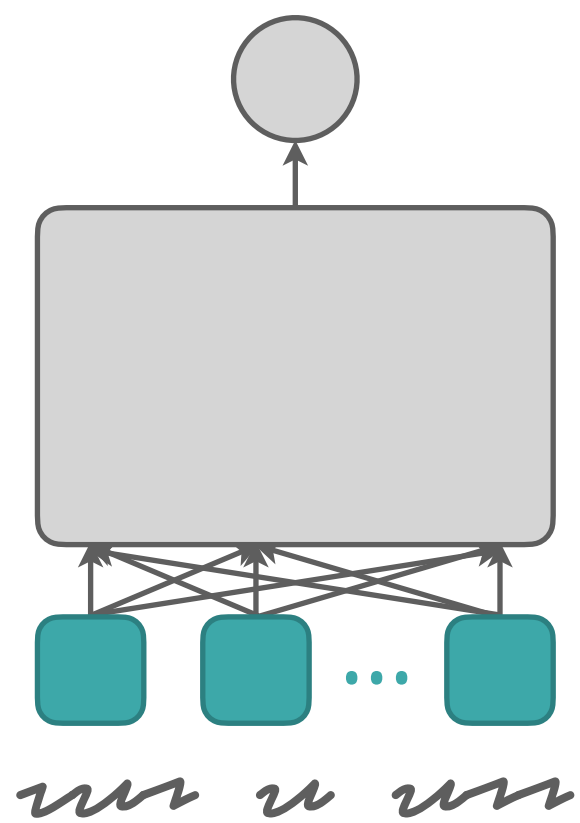
zwaspråchig

z was pr å chi g

...

Method originally proposed by Aepli & Sennrich (2022)

→ *How much of the fine-tuning data should be modified this way?*



Spelling variation & subword tokenization

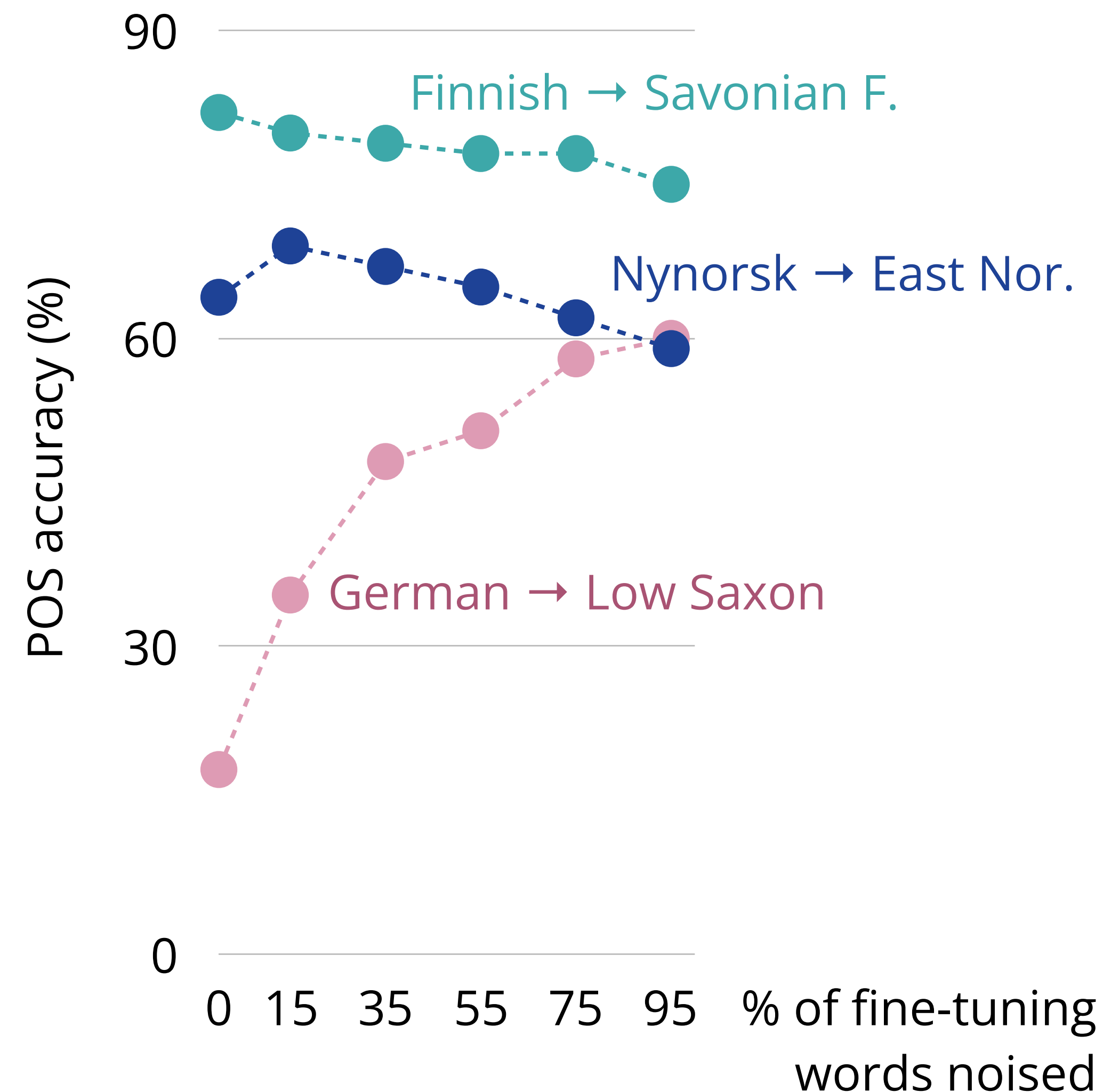
POS tagging experiments

- 7 dialect/variety groups
- 3 encoder models

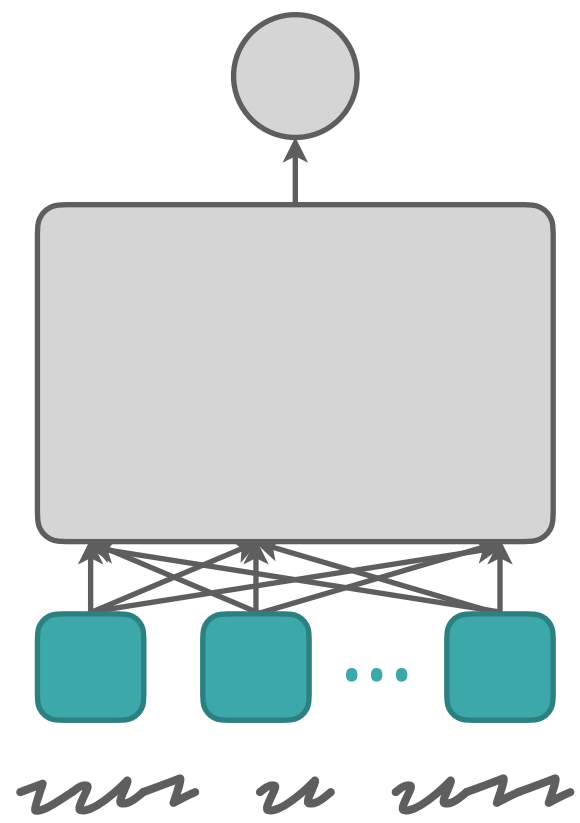
Noise injection increases token fertility

Similar token fertility in fine-tuning & test data
→ better transfer

Ideal noise rate depends on tokenizer & language



Written vs. spoken inputs



German

zweisprachig

zwei sprach ig

Bavarian

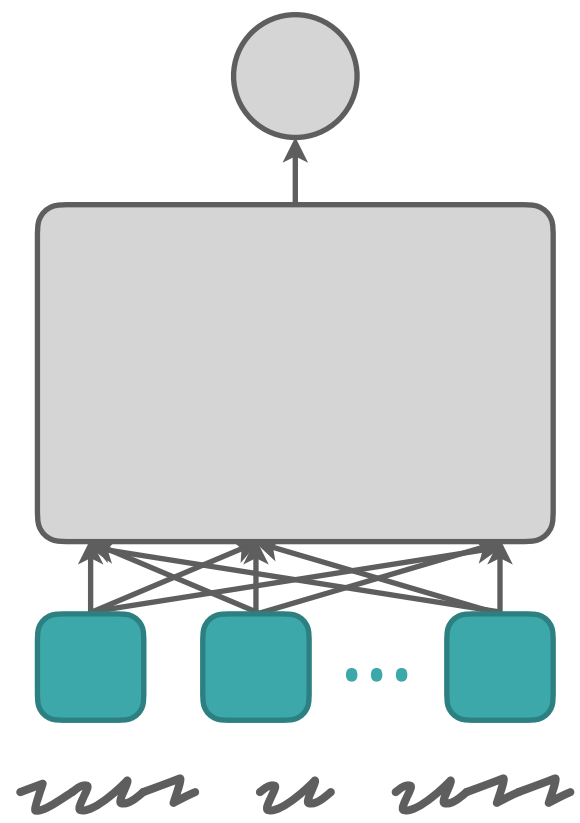
zwaasprochig

z wa as pro chi g



Speech:

- Main modality for dialects
- No need to deal with discrete tokens!



Written vs. spoken inputs

Sentence classification with parallel data

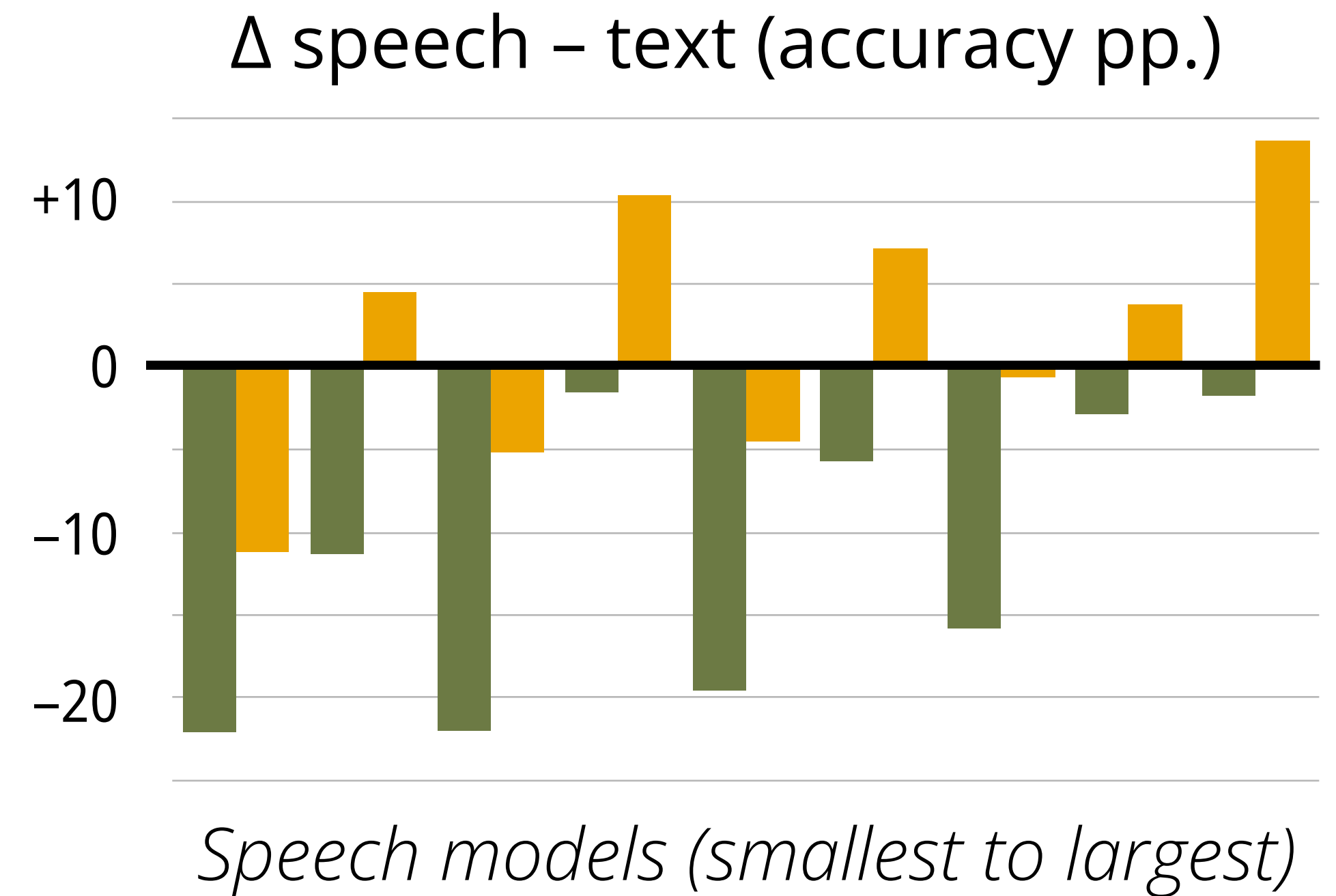
- Text vs. speech
- German vs. dialect

Dialect (often)

speech > text

Standard German

text > speech



→ Different trends across modalities for standard vs. non-standard varieties!

Modelling

MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank

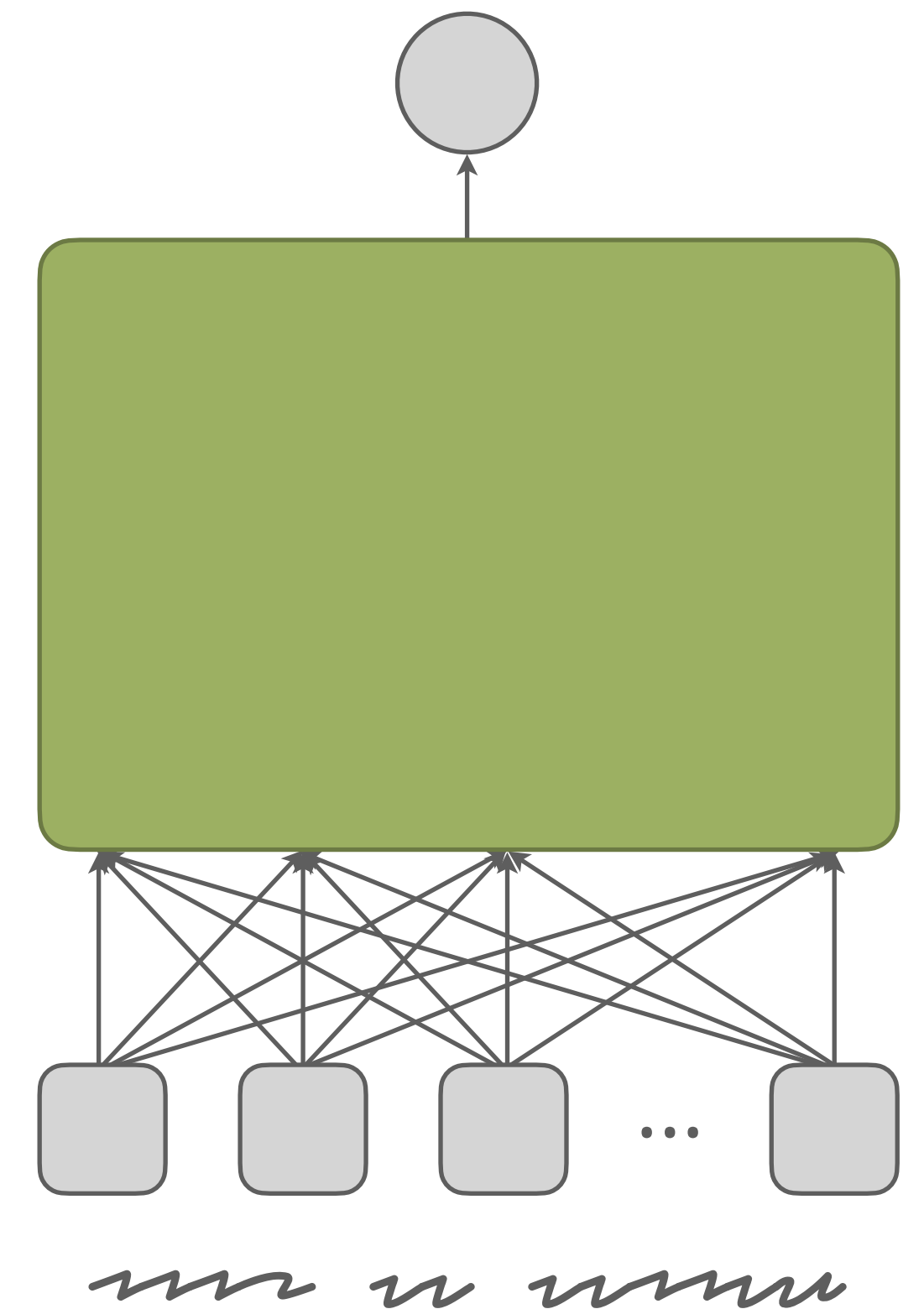
Verena Blaschke,[▲][🏠] Barbara Kovačić,[▲] Siyao Peng,[▲][🏠]
Hinrich Schütze,[▲][🏠] Barbara Plank[▲][🏠][🌐]

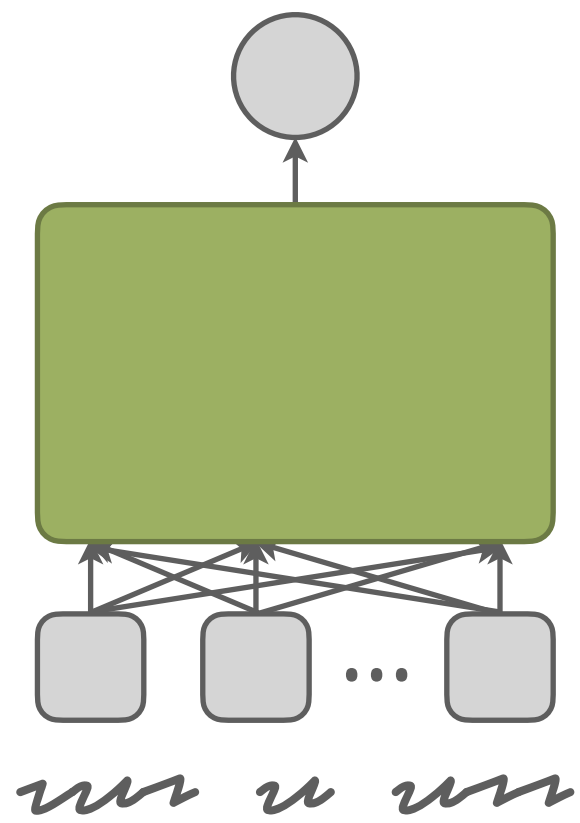
LREC-COLING 2024

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

Ekaterina Artemova [▲]^{*} Verena Blaschke[▲][🏠] Barbara Plank[▲][🏠][🌐]

EACL 2024

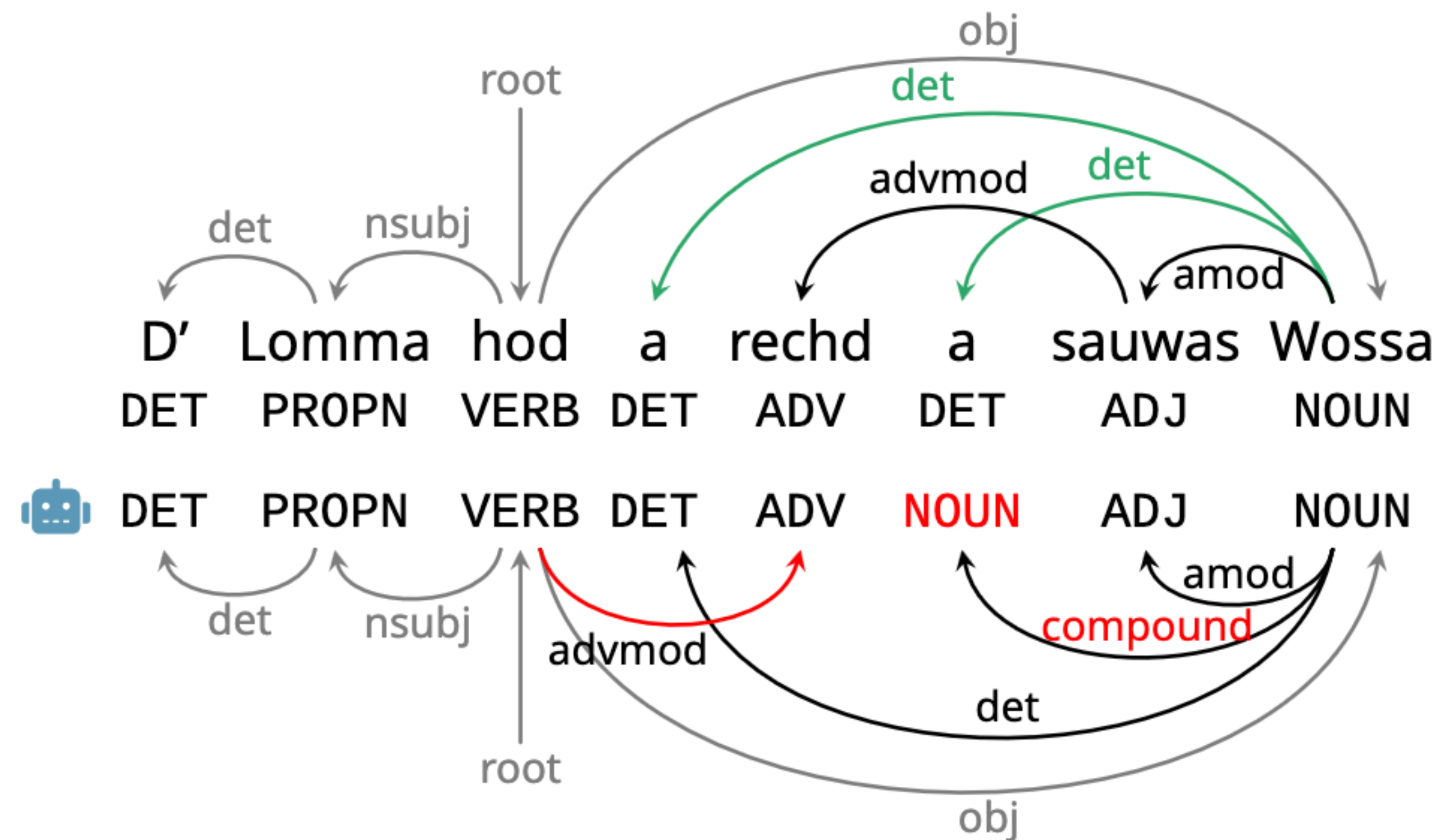




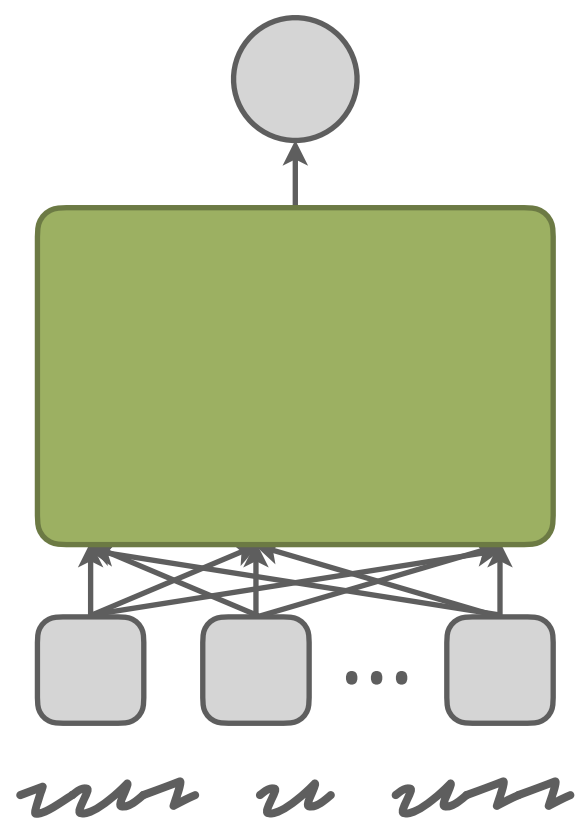
Lack of robustness to morphosyntactic variation

First Bavarian treebank

→ SOTA parsers struggle with Bavarian-specific constructions



*"The river Lammer
has fairly clean
water."*



Lack of robustness to morphosyntactic variation

Also a problem for applied tasks (like slot & intent detection)!

Test suite of morphosyntactic perturbations for German

B-music_item ✓ *B-artist* ✓ *I-artist* ✓
 Spiele das **Lied** von **Eddie** **Vinson** → *PlayMusic* ✓

Play the song by Eddie Vinson

B-music_item ✓ *B-artist* ✓ *B-contact* ✗
 Tue das **Lied** vom **Vinson** **Eddie** spielen → *SearchCreativeWork* ✗
Do the song by the Vinson Eddie play

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

Output desiderata

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

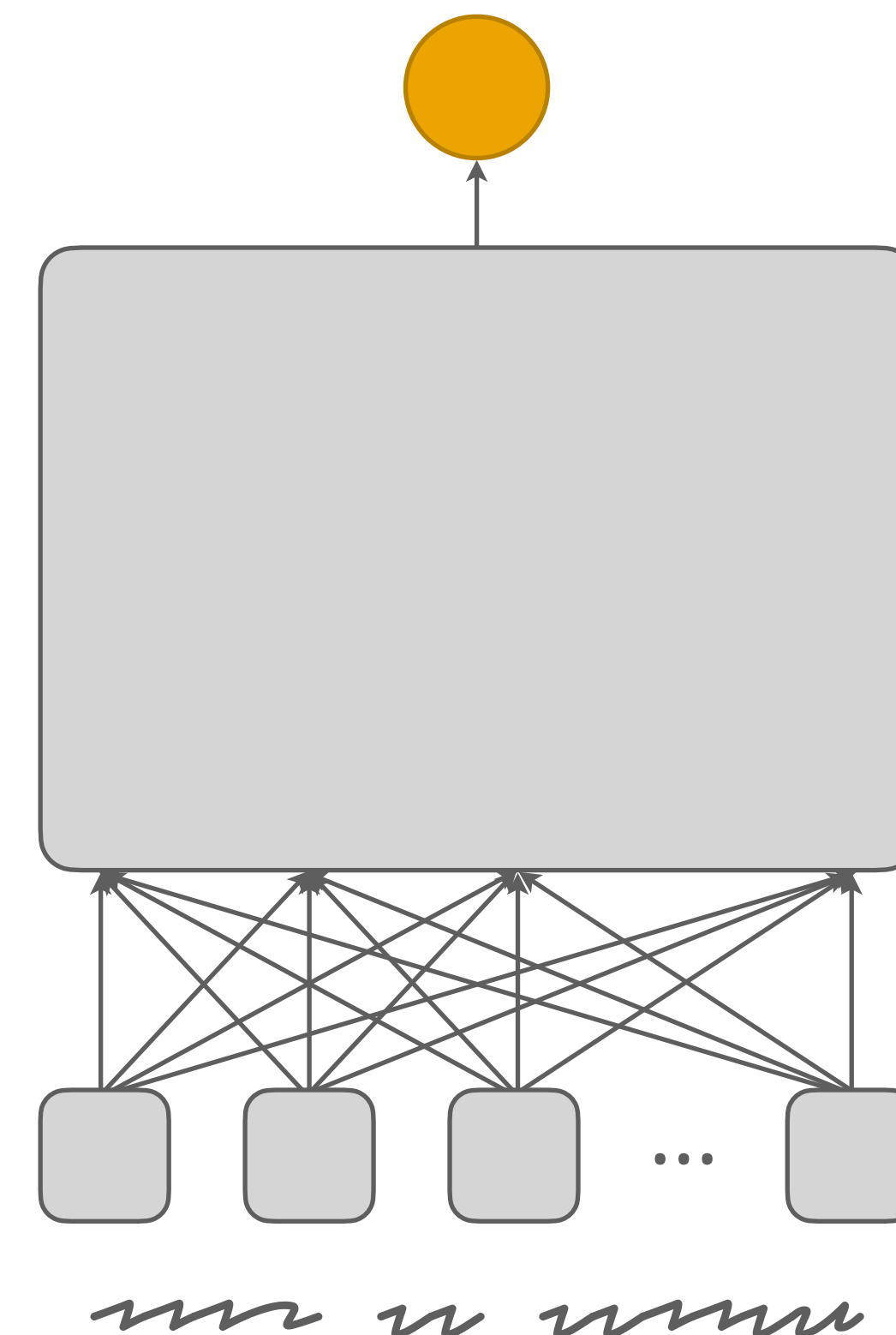
Verena Blaschke[▲] Christoph Purschke[●] Hinrich Schütze[▲] Barbara Plank[▲]

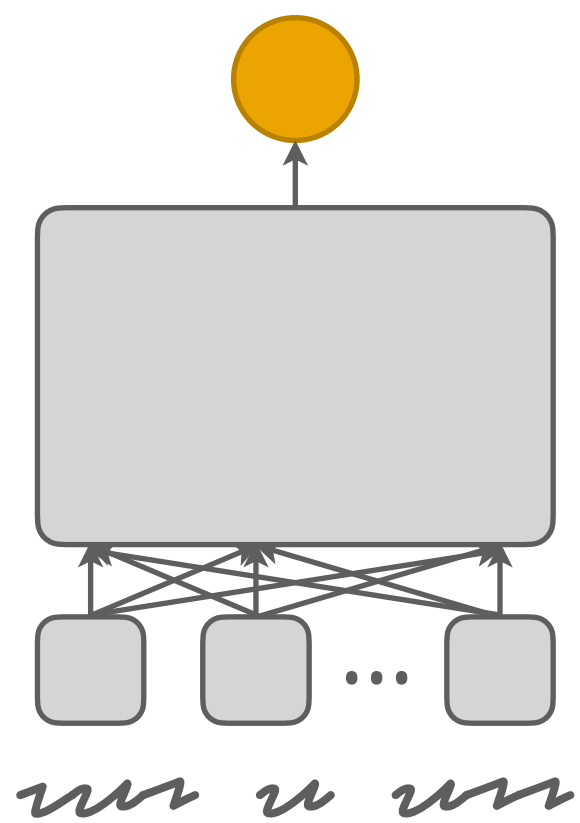
ACL 2024

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

Verena Blaschke^{1,2}, Miriam Winkler¹, Constantin Förster³, Gabriele Wenger-Glemser³, Barbara Plank^{1,2}

Interspeech 2025





Evaluation challenges

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, ...



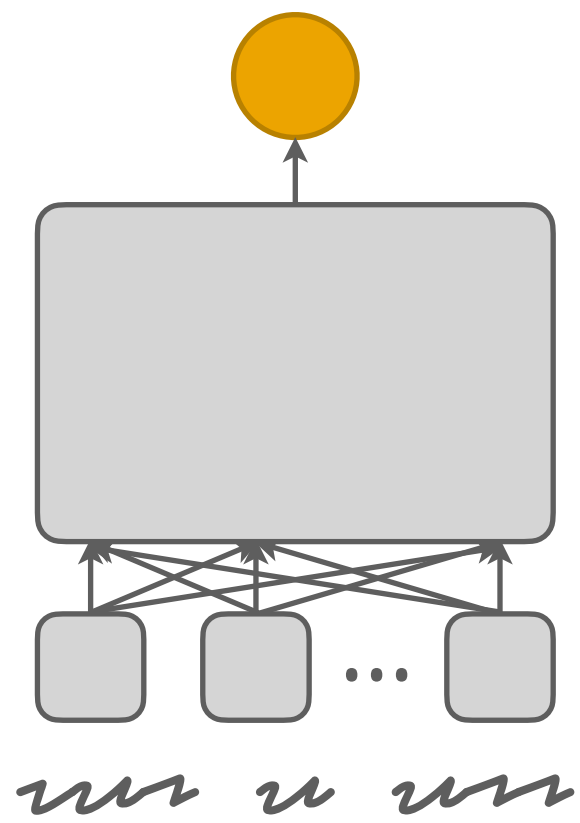
same as German reference

different, but acceptable

different & wrong

Best model (relative to German references):

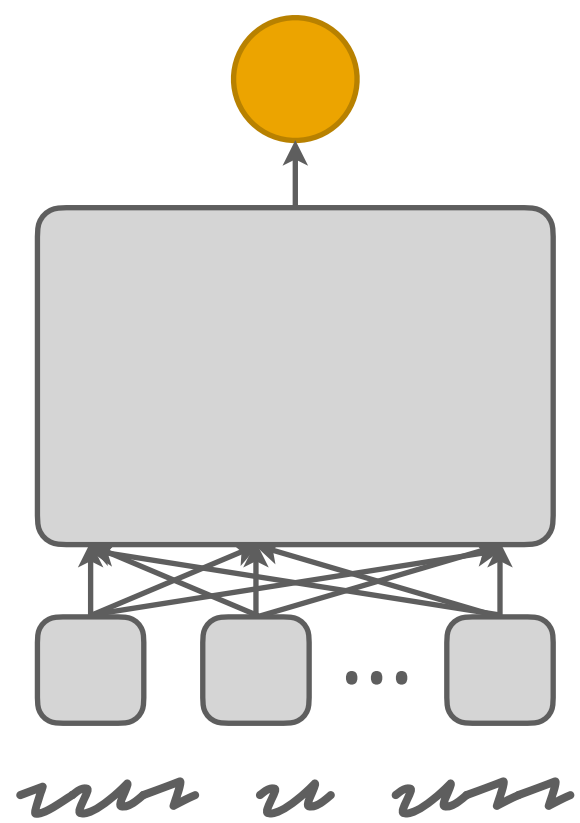
Correlations of automatic scores with human judgement: $0.48 \leq |\rho| \leq 0.63$



What tasks/outputs?

Dialect NLP for whom?

- Dialectologists → annotate/analyze data
- Machine learning researchers → testing generalization of models in sparse, heterogeneous data settings
- Language tech engineers → creating robust NLP systems (e.g., for subtitling, voice-based navigation systems, ...)
- Dialect speakers



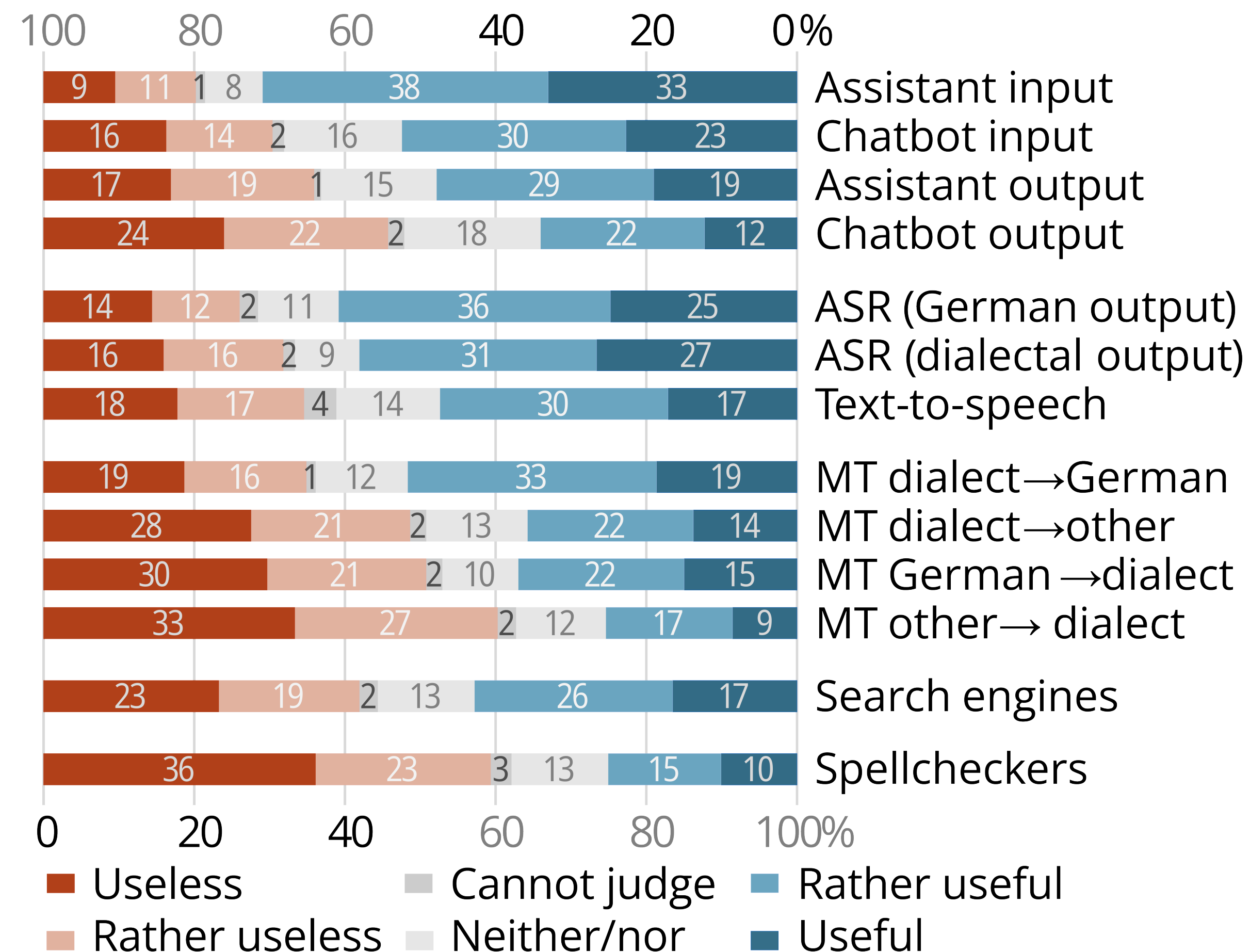
Opinions on hypothetical dialect technologies

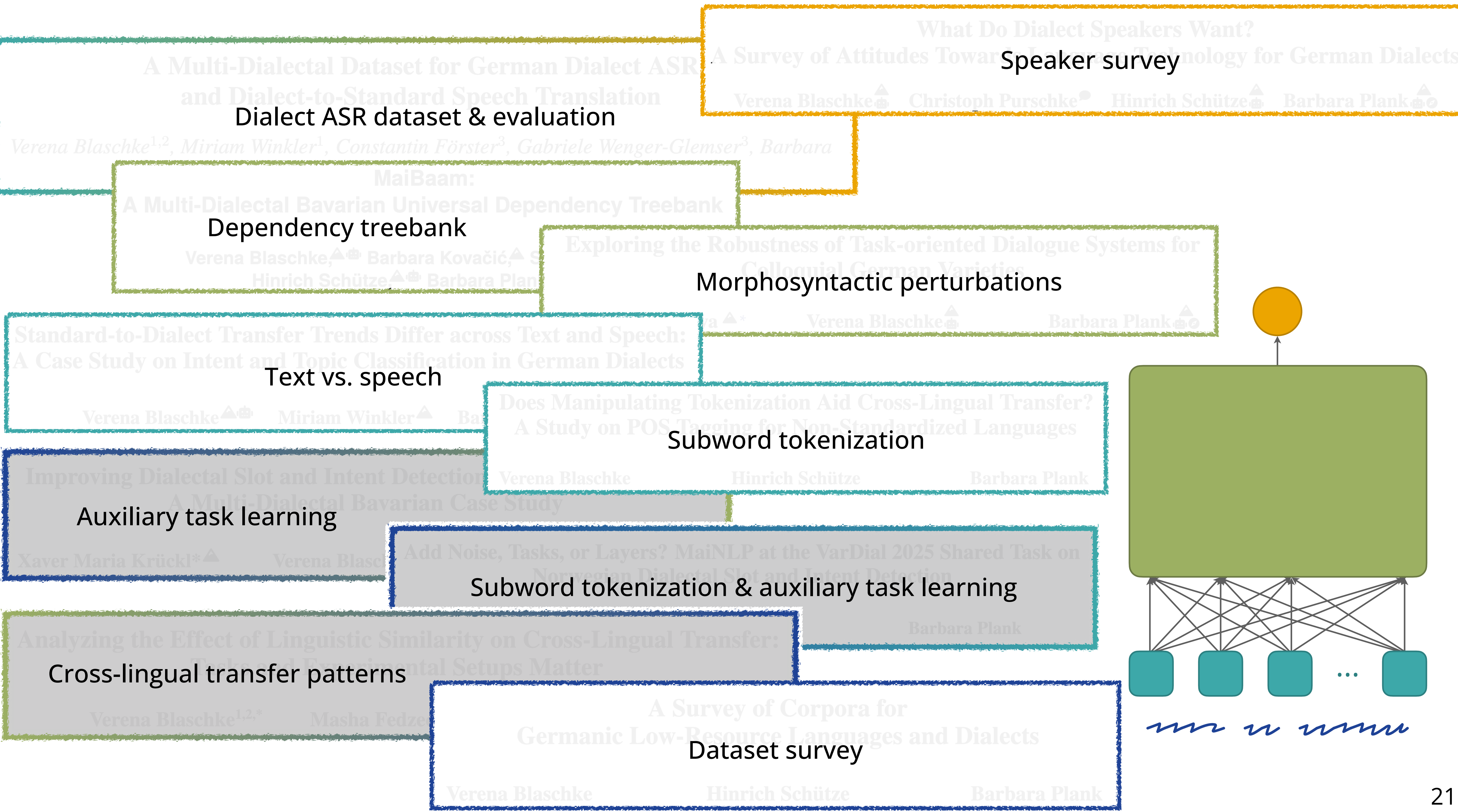
Survey of >300 speakers of German dialects

Trends:

- Handling dialectal input > generating dialectal output
- Speech > text

Language use → attitudes towards dialect technologies





Improving dialect NLP

- Better ways of selecting & modifying fine-tuning data

(VarDial 2023, ACL 2025, VarDial 2025a/b, ACL 2026)

- Targeting speaker interests (ACL 2024)

We need to pay attention to all NLP steps
& to dialectal variation on all linguistic levels
in order to create better dialect NLP systems!

Empirical analyses

- Transfer patterns across languages & tasks

(VarDial 2025a/b, ACL Findings 2025)

- Input representations (VarDial 2023, ACL 2026)
- Effect of variation (EACL 2024, LREC-COLING 2024, Interspeech 2025)

Resources

- Corpus overview (NoDaLiDa 2023)
- Multilingual dialect POS tagging benchmark (VarDial 2023)
- Datasets for German dialects (VarDial 2025a, LREC-COLING 2024, Interspeech 2025, ACL 2026)
- Morphosyntactic perturbation rules (EACL 2024)

