Beyond the Standard: Dialectal Variation at the Heart of NLP

Barbara Plank
LMU Munich

May 16, 2026

DialRes workshop @ LREC 2026
Palma, Mallorca

European Research Council
Munich Center for Machine Learning

The Beauty of Working with Language

There's heavy rain

It's raining heavily

It's raining cats and dogs

It's raining a whole ton

Heavy precipitation in this area

Starkregen in dieser Region

Es regnet sehr stark

Es schüttet aus Kübeln

Es schüttet aus Eimern

Was für ein Wolkenbruch

7000+

Corpus

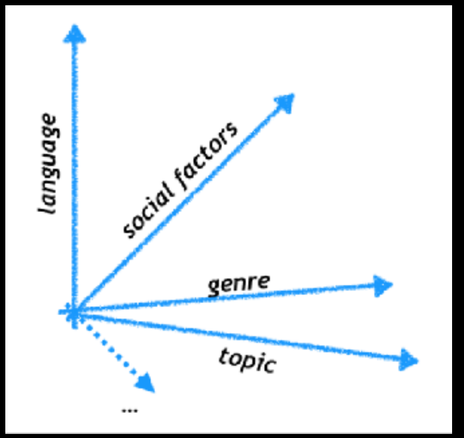
Sample

Variety of a language

Grieve et al. (2025): a population of texts by one or more external factors (e.g. register)

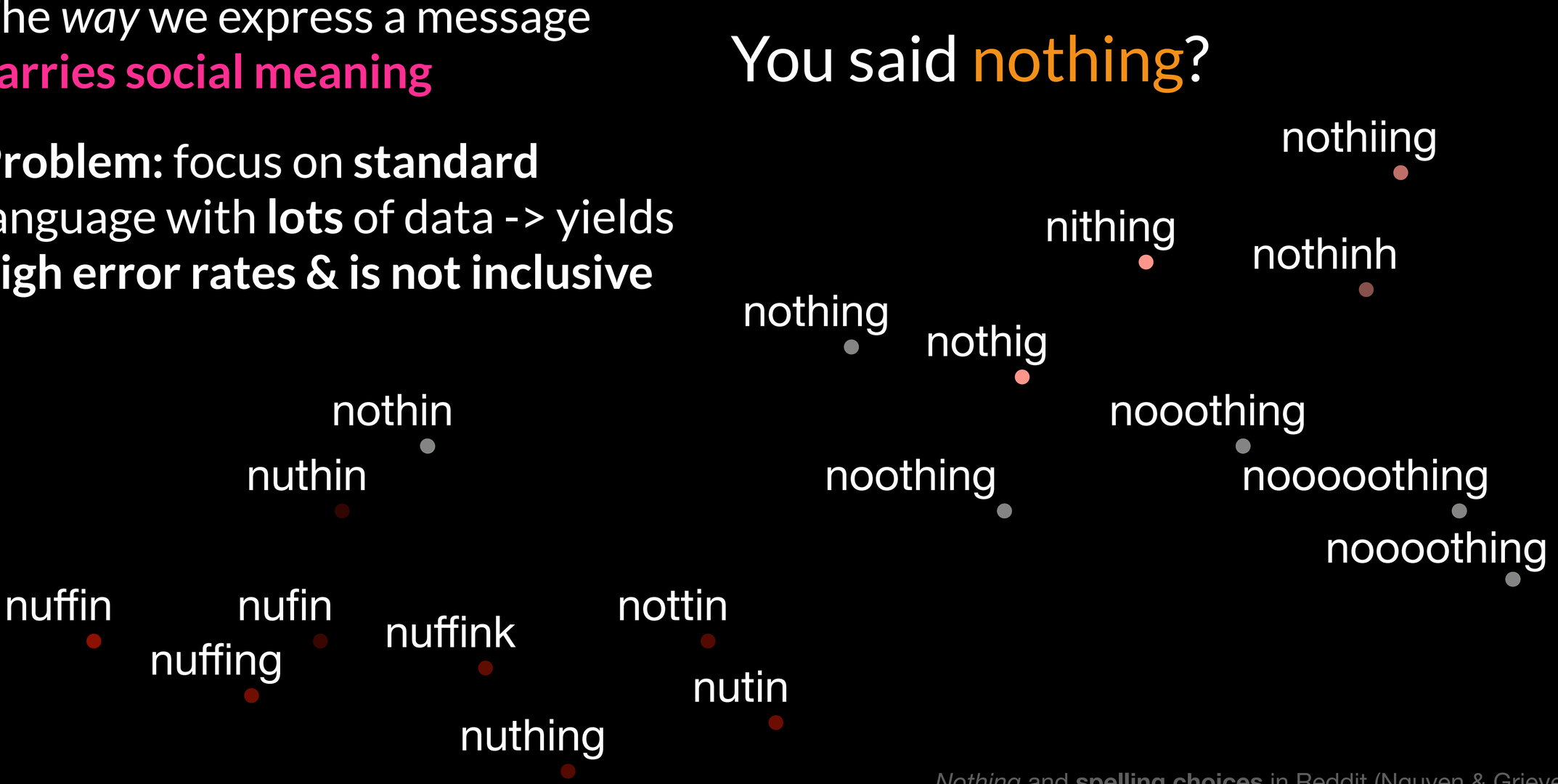
Variety Space

Plank (2016)



Language is full of variation

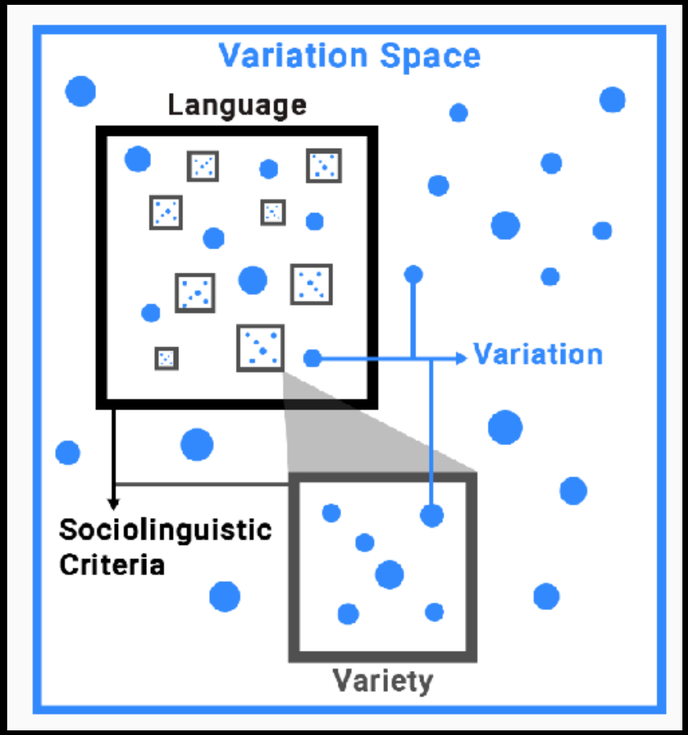
- The *way* we express a message **carries social meaning**
- **Problem:** focus on **standard** language with **lots** of data -> yields **high error rates & is not inclusive**



Nothing and **spelling choices** in Reddit (Nguyen & Grieve, 2020)

Language is full of variation

- The ways in which people speak and write varies in multiple dimensions: regionally, socially, historically, between generations, stylistically, etc. (Berruto, 2010).
- In NLP, variation is typically seen as **noise**
- **Sociolinguistics** to embrace importance of variation throughout research setup:
  - The sum of all different ways of speaking and writing defines what we call the “variation space”, i.e., the totality of all existing linguistic variants

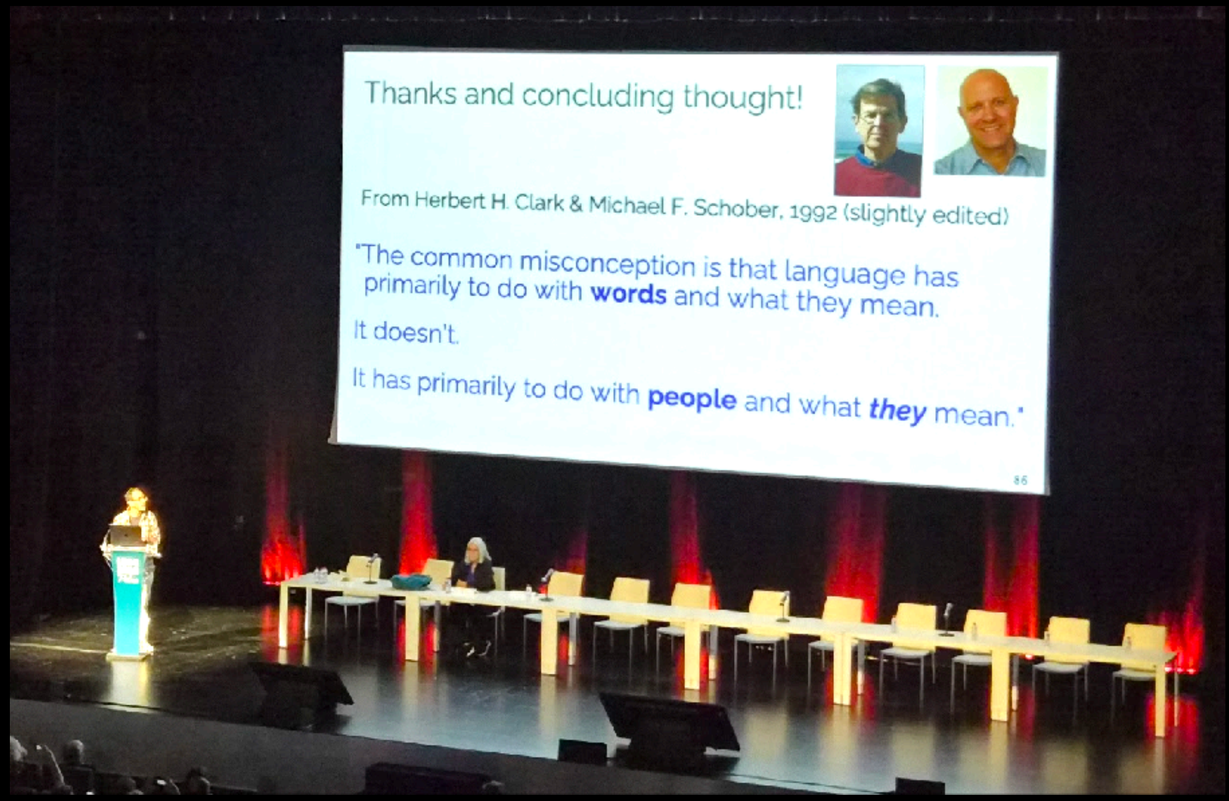


Variation is the Norm!

Language is for and by people

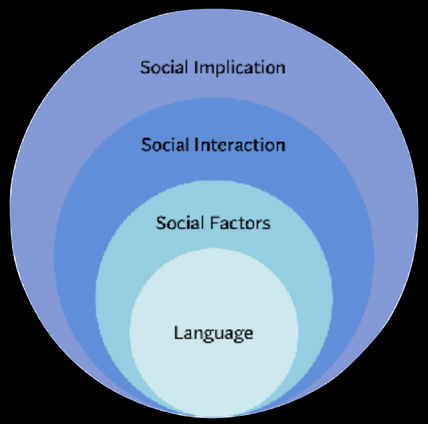
"The common misconception [is] that language use has primarily to do with words and what they mean. It doesn't. It has primarily to do with people and what they mean.

From Herbert H. Clark & F. Schober, 1992.



Dan Jurafsky's keynote at LREC 2026

Socially aware NLP



The Call for Socially Aware Language Technologies

Diyi Yang Stanford University diyiy@stanford.edu
Dirk Hovy Bocconi University mail@dirkhovy.com
David Jurgens University of Michigan jurgens@umich.edu
Barbara Plank LMU Munich bplank@cis.lmu.de

↑ **Language Variation**

Performance ↓

typology

domain

genre

topic

register

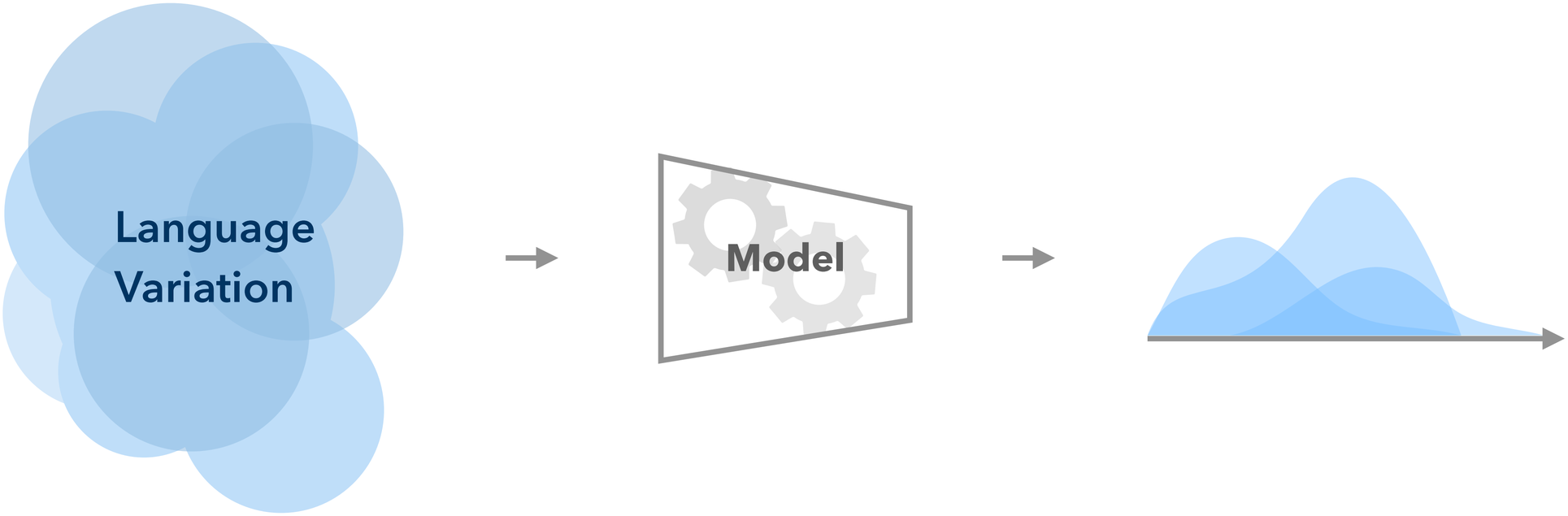
dialect

social context

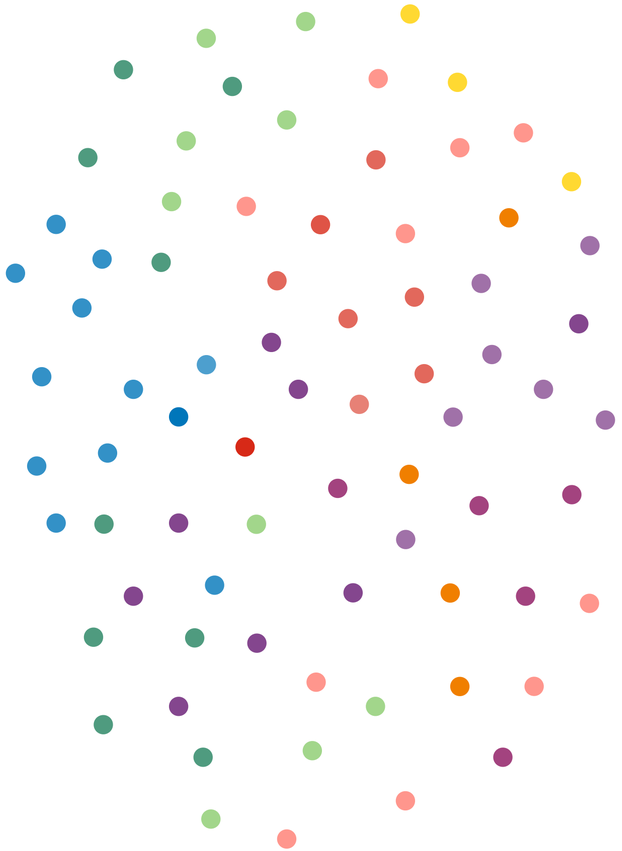
Sekine (1997); Gildea (2001); Plank (2011); Ramesh Kashyap et al. (2021)
Biber (1988); Karlgren and Cutting (1994); Biber (1995); Lee (2001)

Slide by Max-Müller-Eberstein.

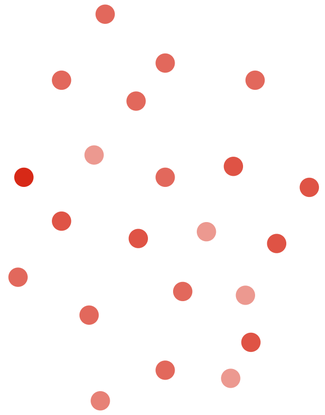
How do we make sure everyone is understood?



Source



Model

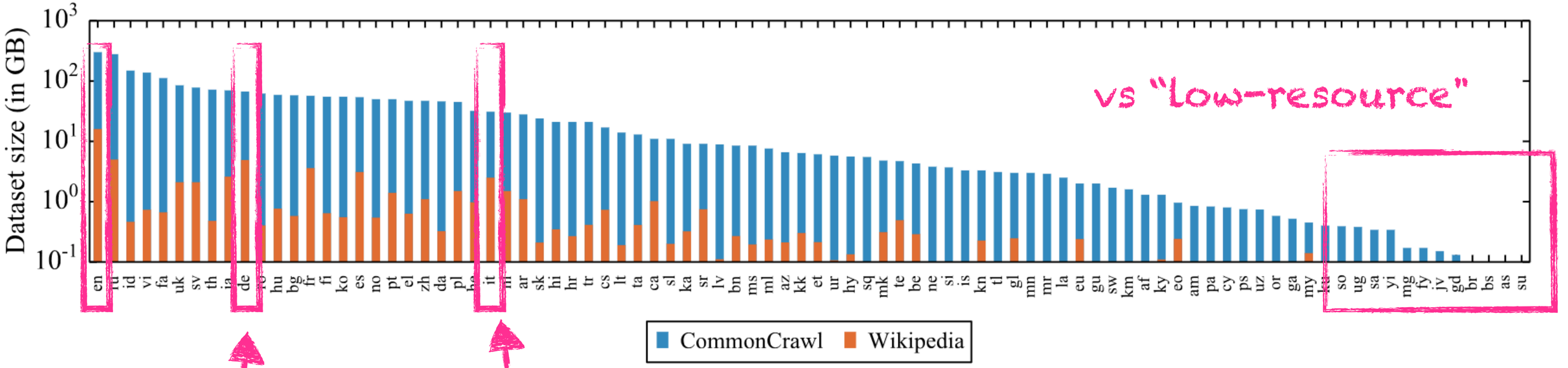


Target

Transfer

The "monolithic" view on languages

"high-resource" languages (EN, DE, IT ...)



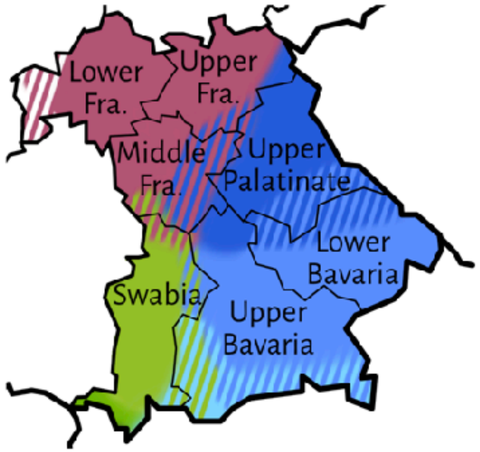
vs "low-resource"

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 XLM-R.

Examples: German or Italian

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

Language Variation is the Everywhere❤️



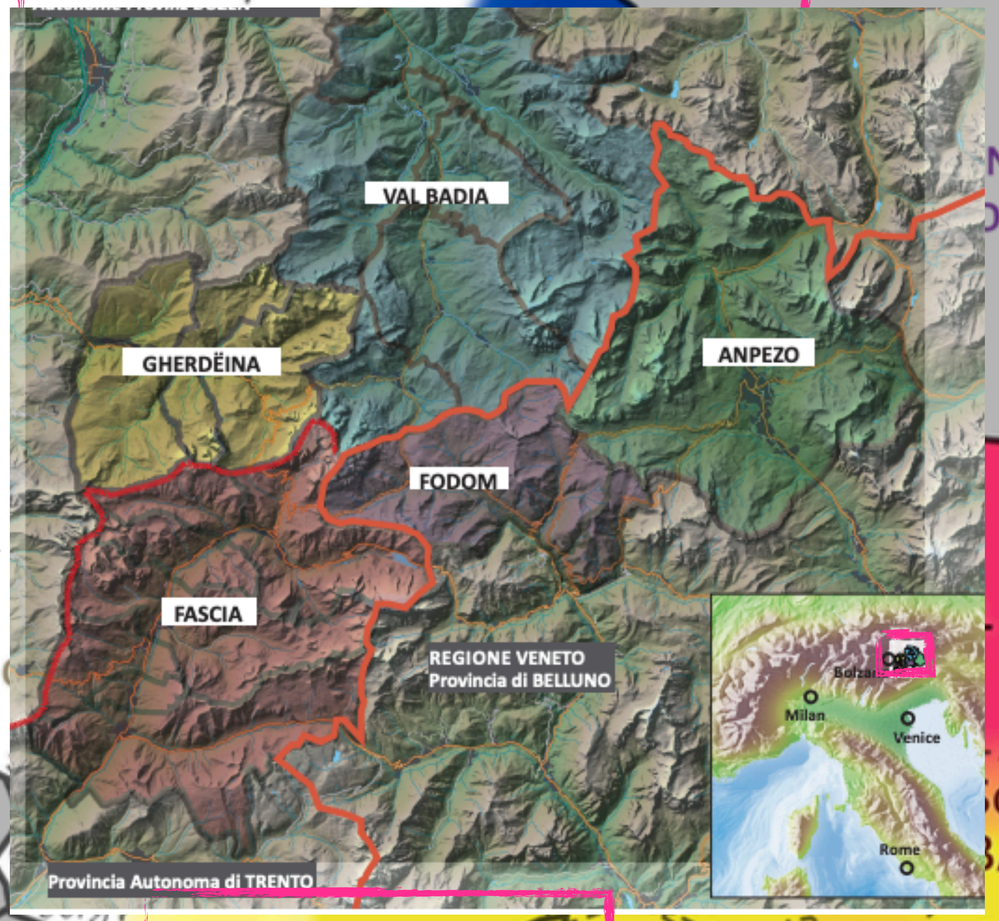
Franconian
East F.

Bavarian
North B.
Central B.
South B.

Alemannic
Swabian

Transition area between dialects

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.

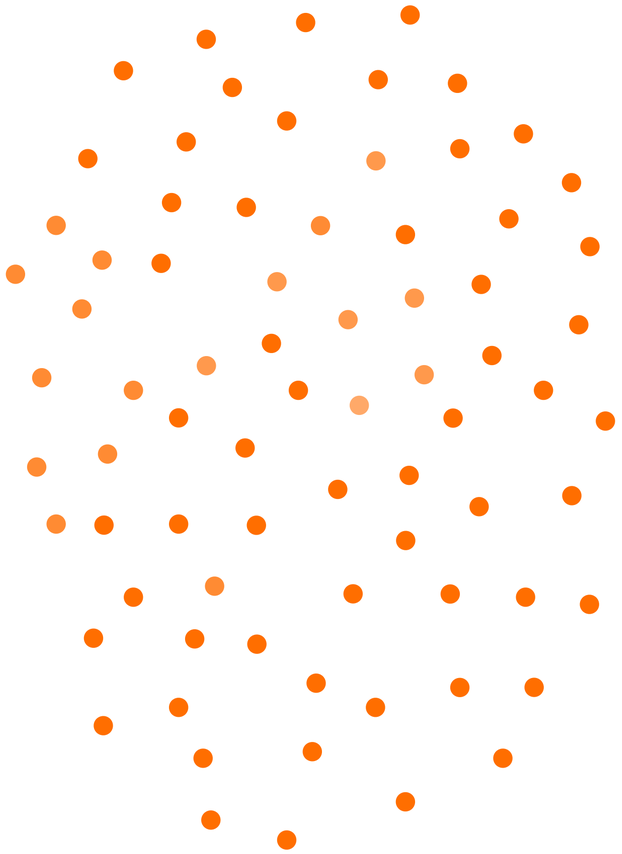
Dialect: "Sprache der Heimat" [Language & Home] (Heidegger/Horan)

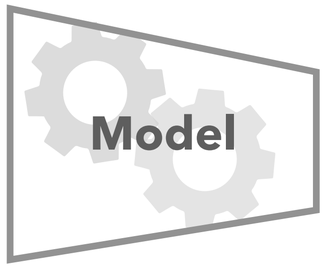
“Heidegger [..] articulated the fundamental yet intangible **relationship** between **dialect as the mother tongue** and **a sense of ‘home’** (Heidegger, 1983). [..]

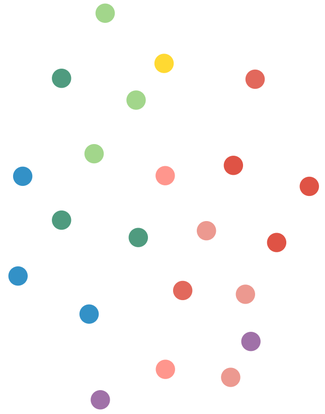
This connection is **particularly strong with dialect** [..]”

Beyond standard: dialectal transfer

German
standard







South Tyrolean
Tyrolean
Austrian
Bavarian
Franconian
Swabian
Alemannic
Low Saxon

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

Outline

Motivation: Beyond “standard" language

Part I - The Problem: Dialects & language variation

- Why are dialects challenging for NLP?
- What resources exist (for German dialects)?

Part II - The How and Why: Dialect transfer & User needs

- Which transfer strategies exist across data, models, and representations?
- What do dialect speakers actually want?

Conclusion and Outlook

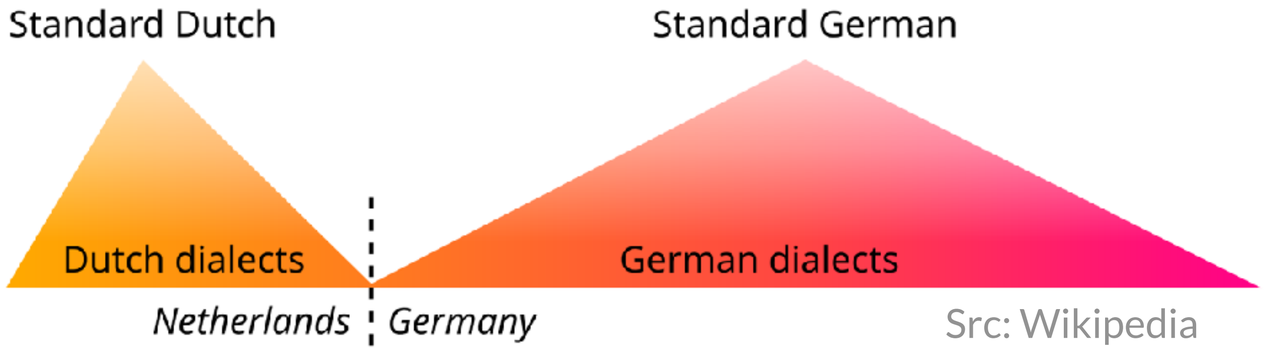
Part 1: Dialects and Language Variation

What makes dialects challenging

What do we mean by dialect?

- Language varieties (in this talk):
  - Non-standardized
  - Closely related to a standard language

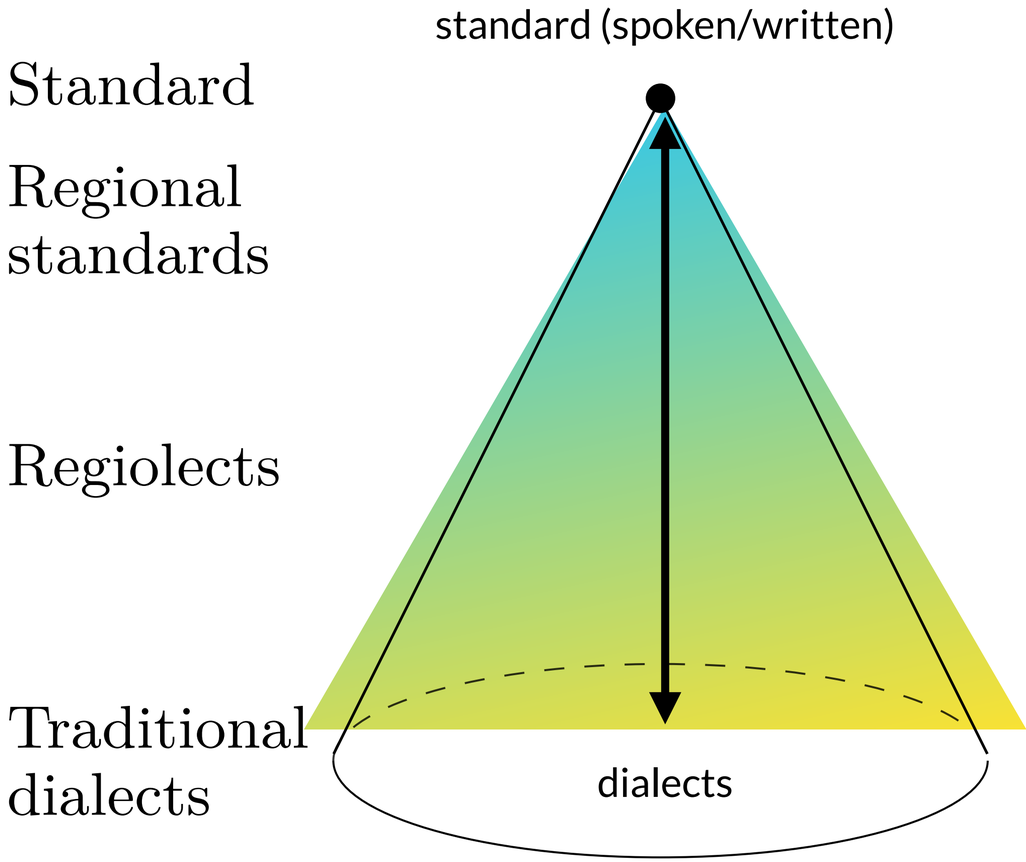
- Often: continuum standard – dialect
- Often: subdialects



Slide parts adapted from Verena Blaschke

What makes dialects challenging?

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



The cone model of dialect-standard variation based on Auer.
3d cone illustration by Verena Blaschke.

Linguistic differences

Differences from the standard language

- Pronunciation ($\rightarrow$ spelling)
- Lexicon

[German]	Sie	haben	keine	**Beine**	
[Bavarian]	Se	hom	koane	**Haxn**	ned
	They	*have*	*no*	*legs*	*not*

“They [=fish] have no legs”

Slide by Verena Blaschke

Linguistic differences

Differences from the standard language

- Pronunciation ($\rightarrow$ spelling)
- Lexicon
- Grammar: morphology, syntax

[German]	Sie	haben	**keine**	Beine	
[Bavarian]	Se	hom	**koane**	Haxn	**ned**
	They	*have*	**no**	*legs*	**not**

“They [=fish] have no legs”

Slide by Verena Blaschke

Linguistic differences

Differences from the standard language

- Pronunciation ($\rightarrow$ spelling)
- Lexicon
- Grammar: morphology, syntax
- Usage context
    - Dialect speakers typically also write (+ speak?) the standard

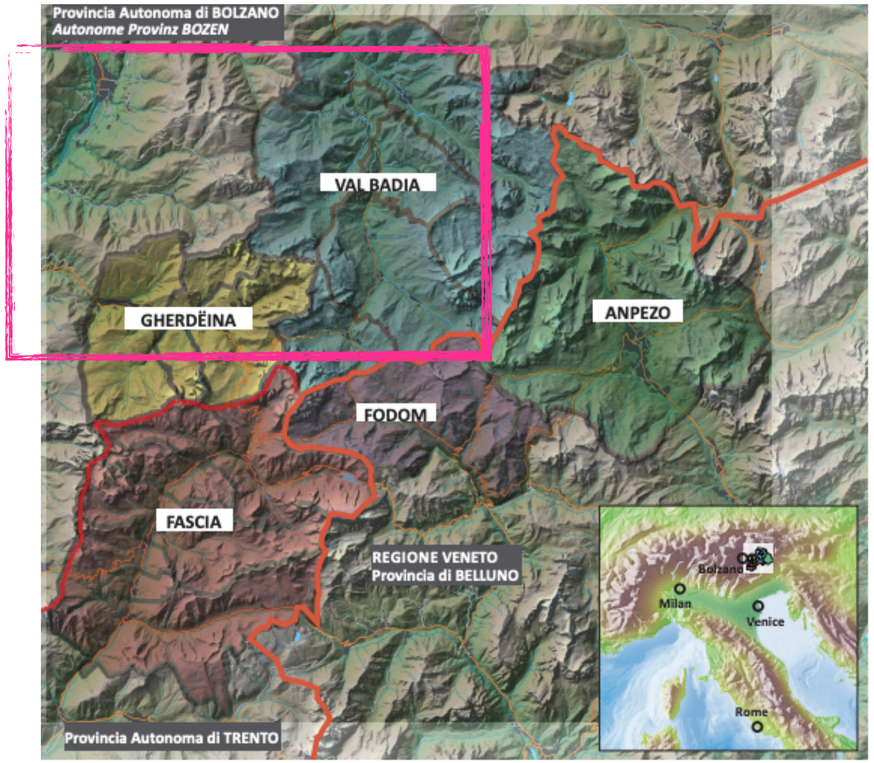
[German]	Sie	haben	keine	Beine	
[Bavarian]	**Se**	**hom**	koane	Haxn	**ned**
	They	*have*	*no*	*legs*	*not*
	De	**ham**	koane	Haxn	_
	Dei	**hobm**	koane	Haxn	_

"They [=fish] have no legs"

Linguistic differences

English	The walls and roofs of ice caves can collapse and cracks can get closed.
Italian	Le pareti e il soffitto delle caverne di ghiaccio sono soggetti a crolli e le crepe possono richiudersi.
Val Badia	I parëis y le sössot di andri da dlacia pó tomé ite y les sfësses pó se stlüje pro.
Gherdëina	I parëies y i plafons dla ciavernes tla dlacia possa tumé ite y la sfëntes possa se stlù.

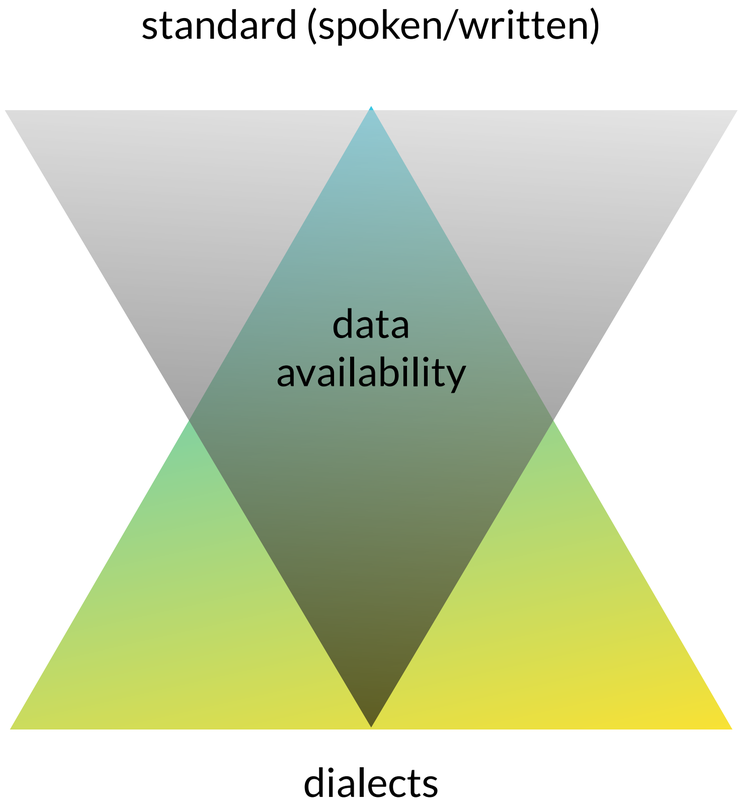
Table 1: Example translations from the FLORES+ dev split.



Src: Frontull et al. 2025

What makes dialects challenging?

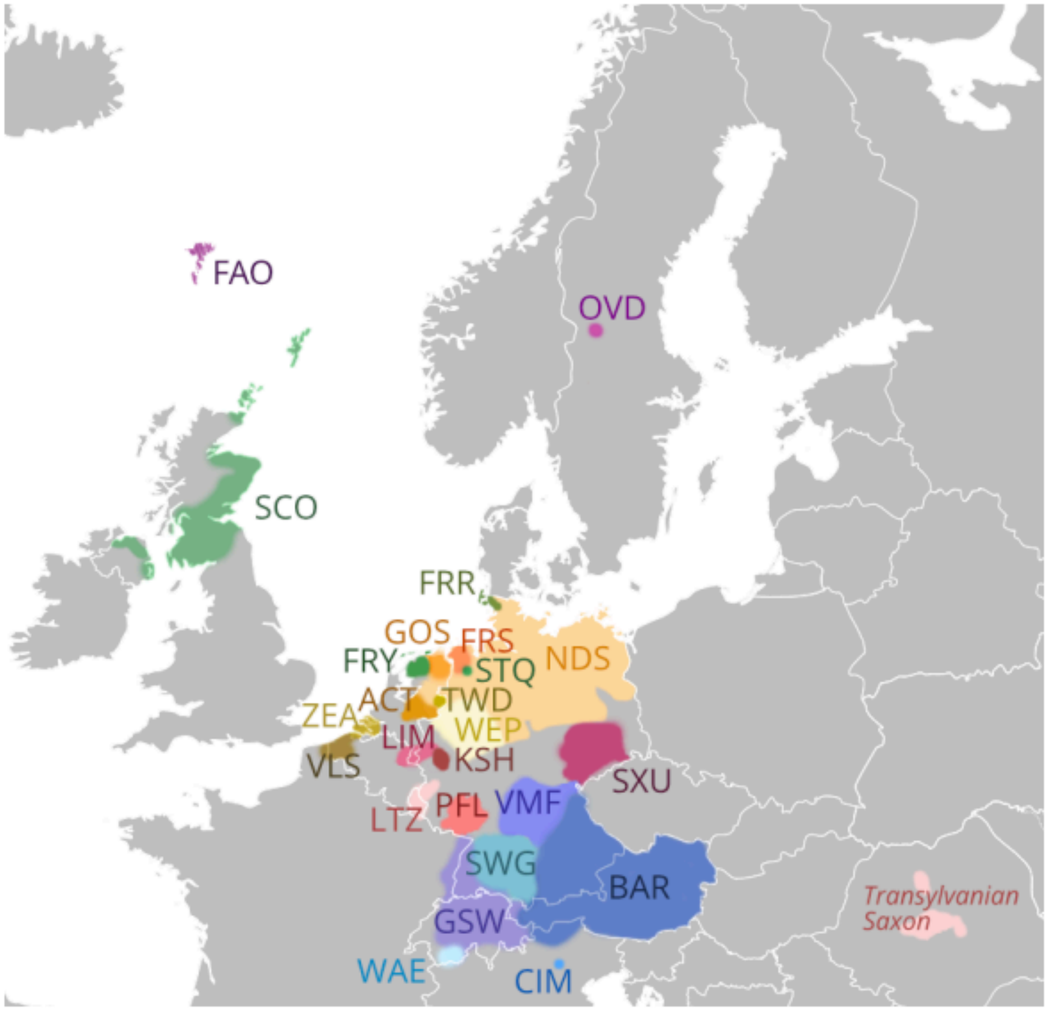
- Linguistic differences
- Data challenges
- Representation challenges
- Evaluation obstacles



Data challenges for Dialect corpora

- Limited availability
- Quality & written representations

- Example for Germanic dialects:
  - 100+ datasets for 35 Germanic dialects
    + small languages



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

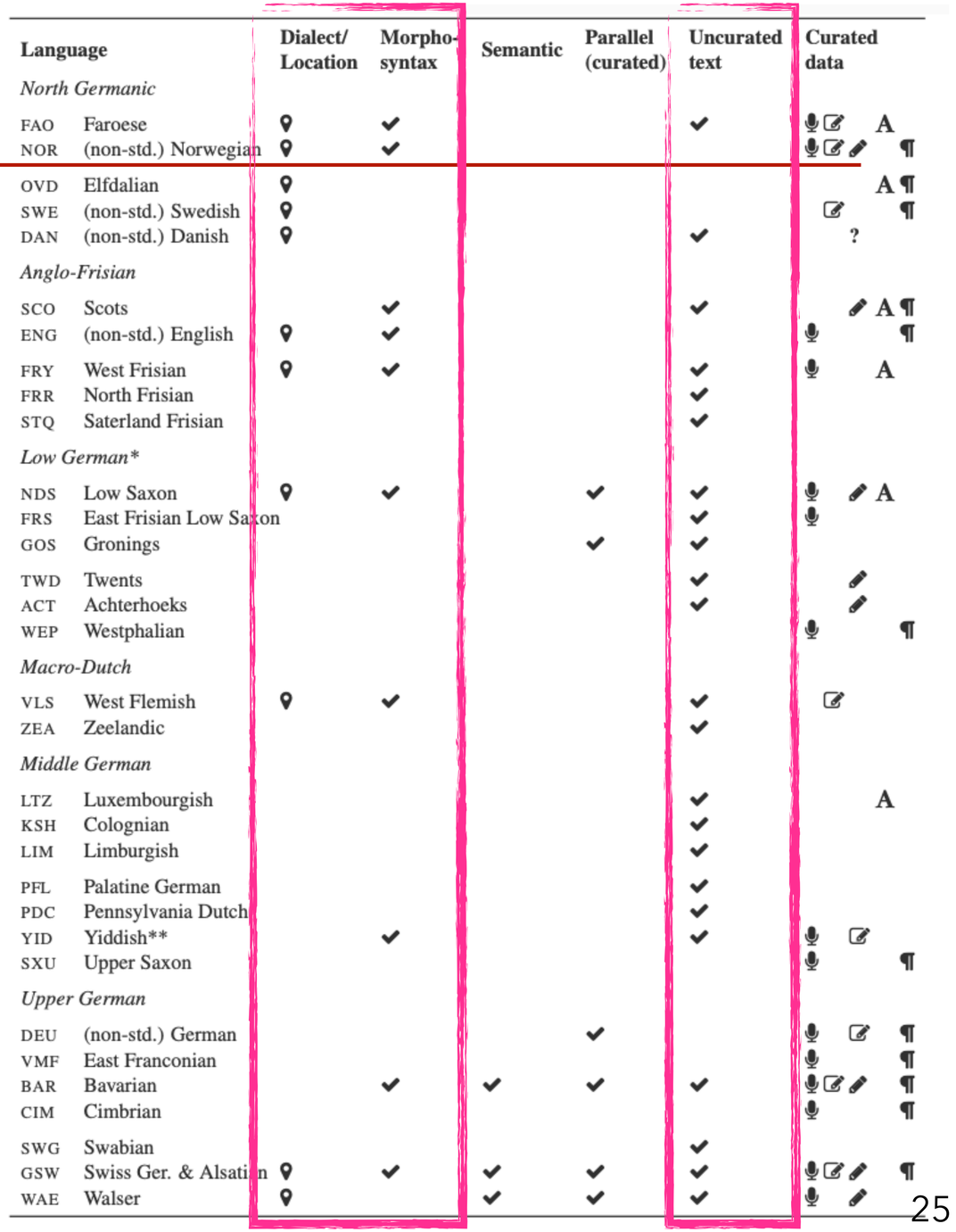
Verena Blaschke Hinrich Schütze Barbara Plank

https://aclanthology.org/2023.nodalida-1.41.pdf

Repository: github.com/mainlp/germanic-lrl-corpora

Annotations

- What, if any, annotations did we find?
  - Morphosyntax (POS, Universal dependencies, phrase structures)
  - Gelocation and dialect group
  - Rarer: translations, paraphrases, sentiment, slot and intent detection



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

Blaschke et al. 2023

Wikipedias for Dialects

Wikipedia & Language		Articles (01/2023)	Manual edits (2001–2022)	Manual edits (2022)	Monthly editors (2022)
nds	NDS (Germany)* (📍)	84 k	44 %	99 %	30
lb	LTZ	61 k	43 %	85 %	56
fy	FRY	50 k	60 %	99 %	54
sco	SCO	39 k	53 %	63 %	70
als	GSW + SWG + WAE (📍)	30 k	69 %	100 %	58
bar	BAR (📍)	27 k	68 %	63 %	39
frr	FRR (📍)	17 k	79 %	85 %	16
yi	YID	15 k	49 %	97 %	35
li	LIM	14 k	42 %	75 %	21
fo	FAO	14 k	41 %	99 %	29
vls	VLS (📍)	8 k	45 %	79 %	16
nds-nl	NDS (Netherlands)* (📍)	8 k	40 %	68 %	14
zea	ZEA	6 k	47 %	98 %	10
stq	STQ	4 k	38 %	81 %	8
ksh	KSH + other Ripuarian (📍)	3 k	32 %	99 %	6
pfl	PFL + oth. Rhen. Franc., Hessian (📍)	3 k	65 %	72 %	6
pdc	PDC	2 k	27 %	92 %	6
en	ENG	6608 k	90 %	92 %	102 574
de	DEU	2765 k	91 %	93 %	16 141
nl	NLD	2114 k	68 %	66 %	3521
da	DAN	289 k	63 %	64 %	711
is	ISL	56 k	54 %	79 %	118

Blaschke et al. 2023

Quality & Written Representations

- Mostly unannotated
  - .. and sometimes uncurated
    - quality issues especially for low-resource (Kreutzer+, TACL 2022; Abadji+, LREC 2022)
    - Scots Wikipedia issue
    - Example: "West Flemish" QED corpus
- In various (often undocumented) written representations (e.g., phonetic/phonemic transcription, pronunciation spelling, LRL orthography, standard orthography)

```
<w id="33.28">07,</w>
<w id="33.29">624&amp;</w>
<w id="33.30">lt;</w>
<w id="33.31">br</w>
<w id="33.32">/</w>
<w id="33.33">&amp;</w>
<w id="33.34">gt;</w>
<w id="33.35">Καλά</w>
<w id="33.36">,</w>
```

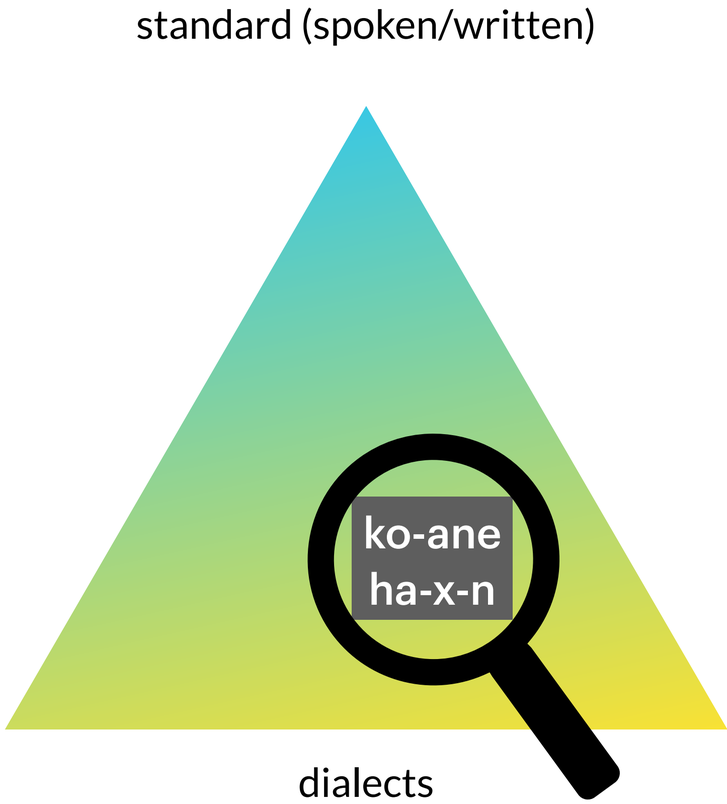
From the Swiss German ArchiMob corpus (Scherrer et al., 2019b):
6a ¶ können sie ihre jugendzeit beschreiben
6b chönd sii iri jugendziit beschriibe
"Can you describe your youth?"

From the BISAME corpus (STIH, 2020):
7a Niema hat salamols gweßt as die Werter vum franzeescha kumma.
"Nobody knew then that these words came from French."

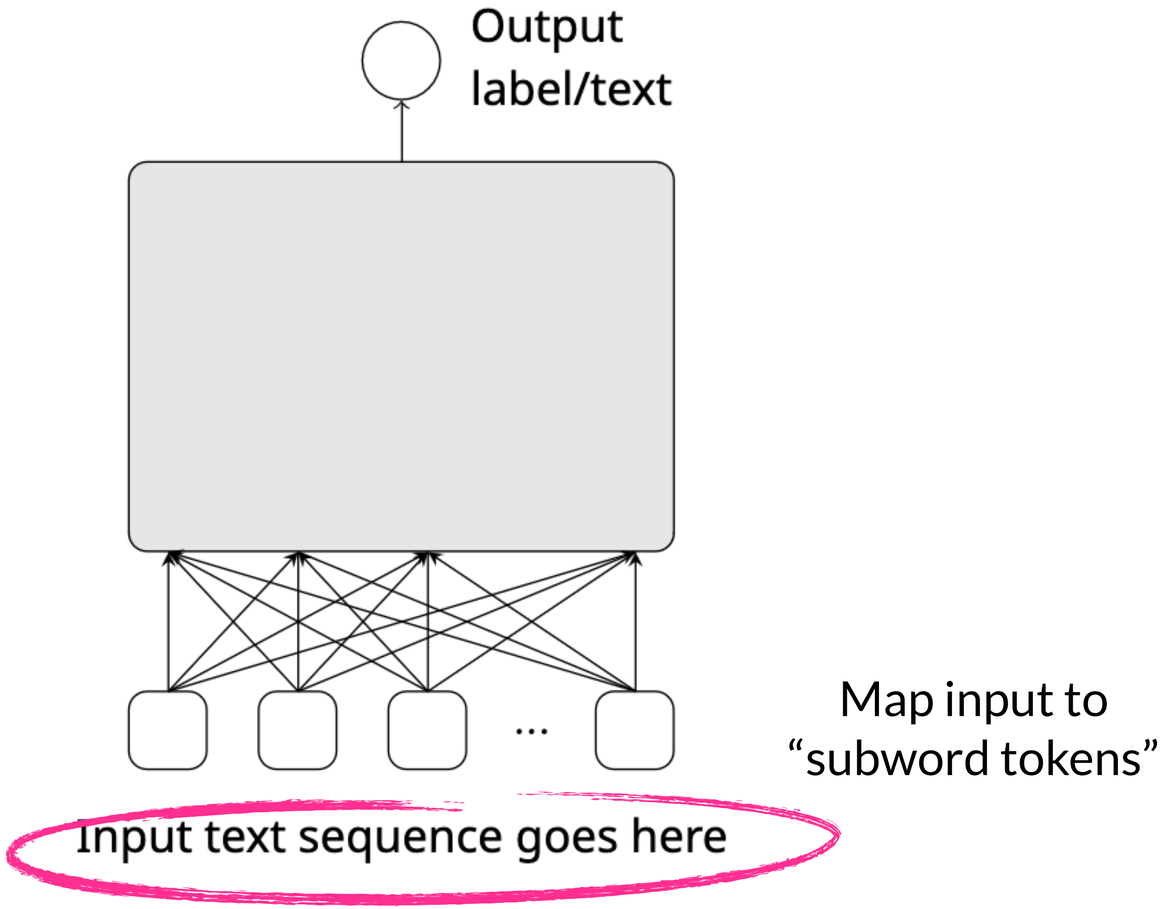
Blaschke et al. 2023

What makes dialects challenging?

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

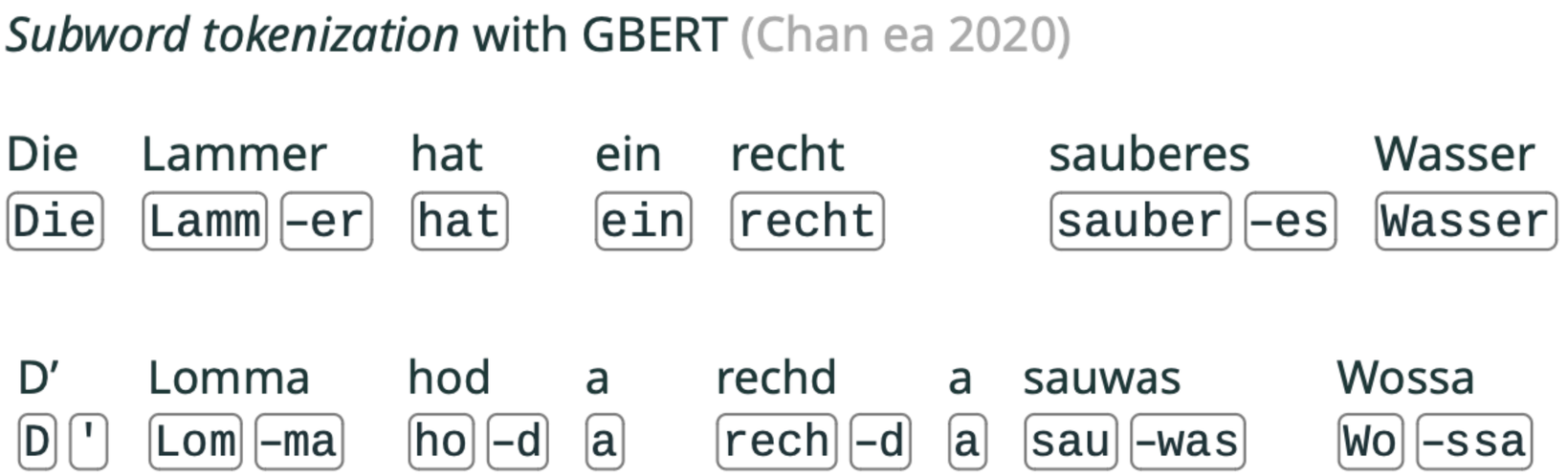


What's the input representation?



Tokenization and non-standardness

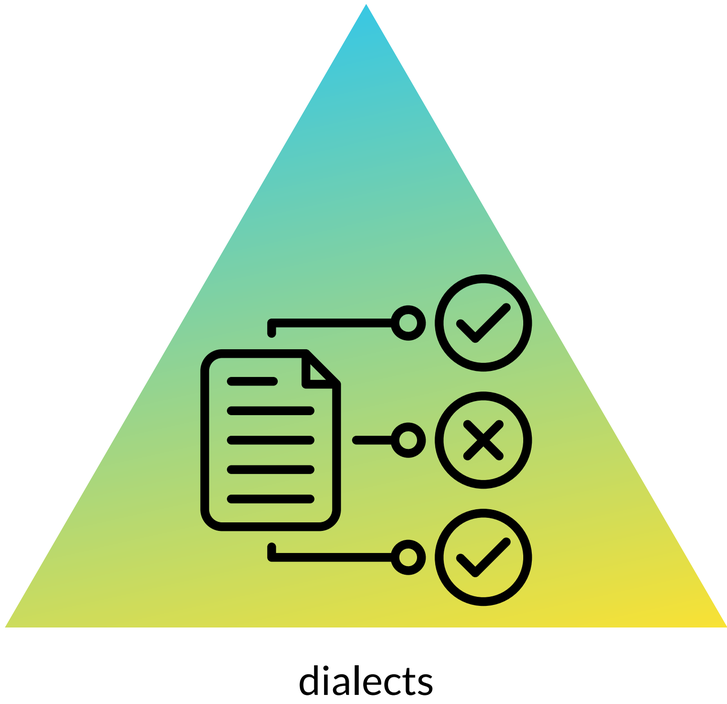
- State-of-the-art LMs are based on subwords (not characters). This representation is sensitive to slight surface variations: minor changes will lead to different segmentations & representations
  - Example:



Sentence via `bar.wikipedia.org/wiki/Låmma`
Chan ea "German's next language model" 2020

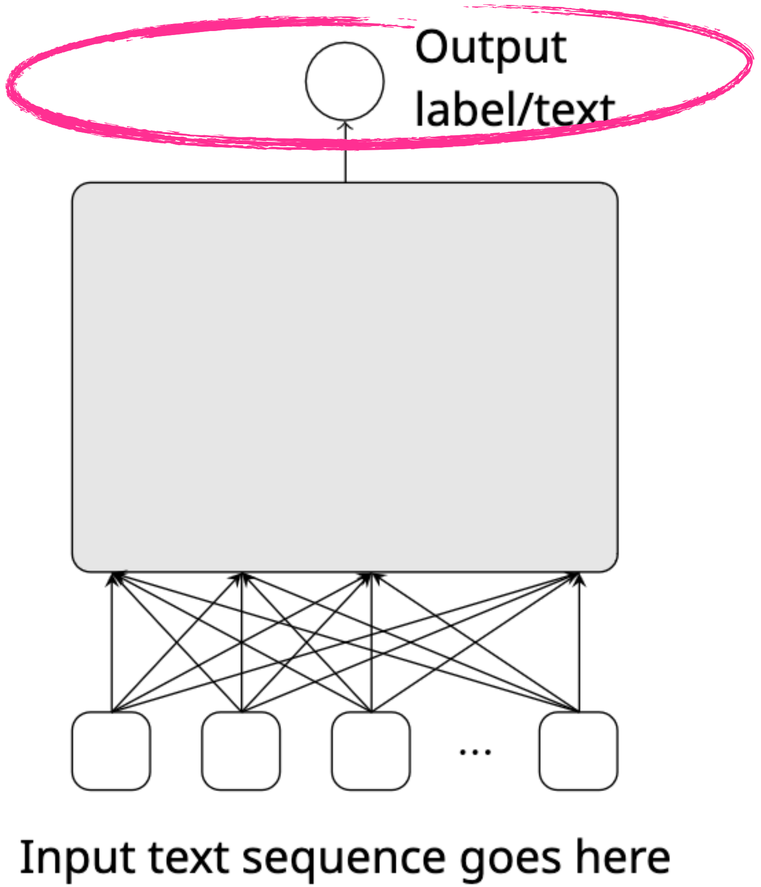
What makes dialects challenging?

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



How to evaluate systems for dialects?

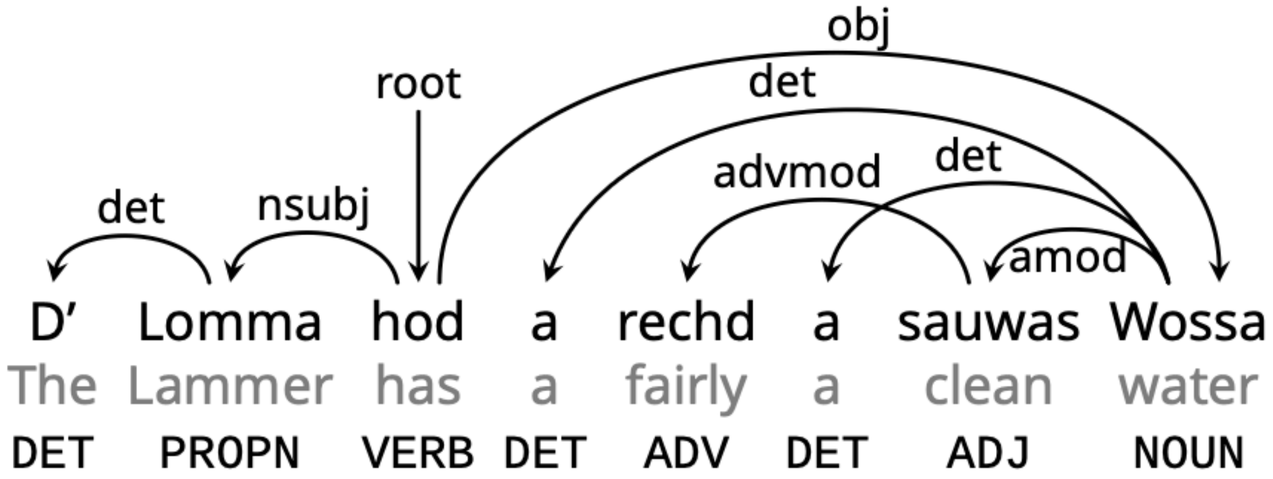
- If we have “gold data” (translations/transcriptions/annotations) we can measure the **performance gap (dialect gap)**



Example: MaiBaam Treebank for Bavarian

Universal Dependencies (de Marneffe ea 2021)

- Cross-linguistic comparability
- Multi-dialectal



Blaschke ea "MaiBaam: A multi-dialectal Bavarian Universal Dependency tree-bank" LREC-COLING 2024
de Marneffe ea "Universal Dependencies" *Computational Linguistics* (2021)

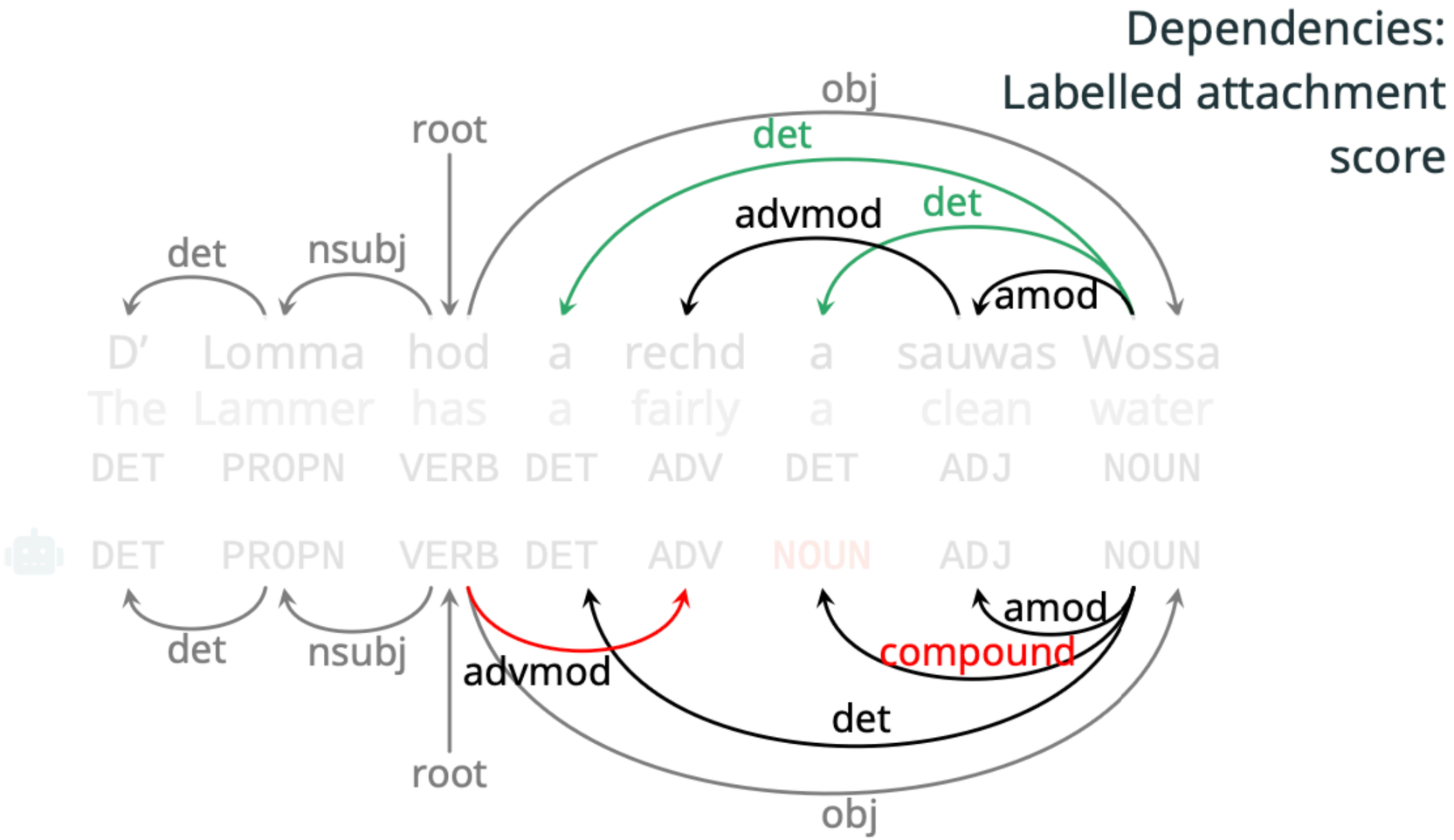
Quantifying the dialect gap (Standard German vs Bavarian)

Train on German data (there is no Bavarian training data!),
test on German vs. Bavarian

Model	Test lang	Acc (%)	LAS (%)	Input representation
Stanza	DEU	95.9	83.7	
GBERT	DEU	96.8	83.1	
UDPipe	DEU	96.5	84.9	
Stanza	BAR	40.9	23.1	Full words
GBERT	BAR	57.4	30.1	Subword tokens
UDPipe	BAR	80.5	67.3	Subword tok. + characters

Blaschke et al. 2024

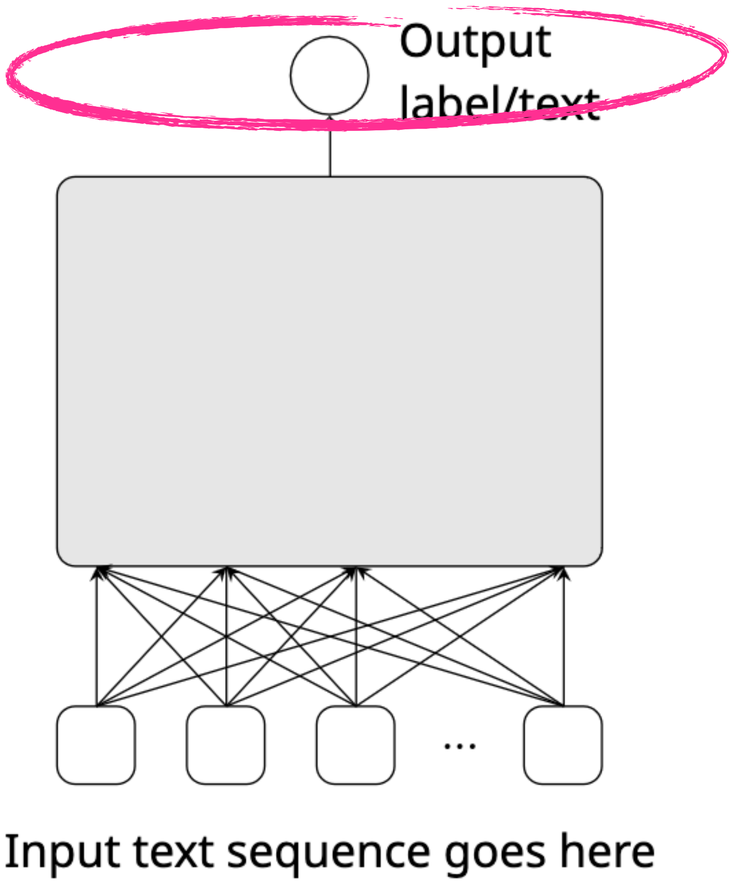
Brittle towards uncommon structures



Blaschke et al. 2024

Evaluation obstacles

- In many case, we do not have any gold data yet
  - **Solution:** Create resources
    - Selected examples of resources we contributed recently



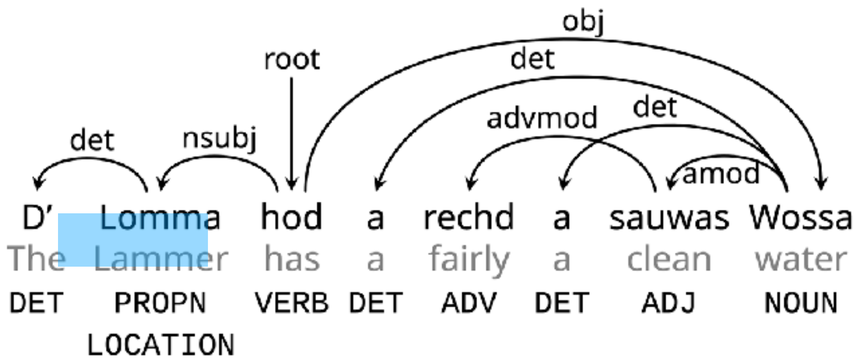
Multi-Dialectal German NLP benchmarks

- **NaBaSID**: Bavarian Slot and Intent Detection:
  - Translations and Natural Data (Winkler et al., 2024; Krückl et al., 2025; Winkler et al., 2026)

Bavarian	Wos is grod für a weda in äding ?
	(en) *What is the weather like in* Altötting *right now* ?

https://aclanthology.org/2024.lrec-main.1297/ https://aclanthology.org/2025.vardial-1.10.pdf

- **MaiBaam**: Universal dependencies (Syntax):
  - Brittleness toward non-common structures



MaiBaam:
A Multi-Dialectal Bavarian Universal Dependency Treebank

https://aclanthology.org/2024.lrec-main.953

- **BarNER**: Named entities:
  - Largest publicly available manually annotated NER dataset, 2 genres

Sebastian, Basti, Wastl?!
Recognizing Named Entities in Bavarian Dialectal Data

https://aclanthology.org/2024.lrec-main.1262/

- **Betthupferl**: Multi-dialectal ASR

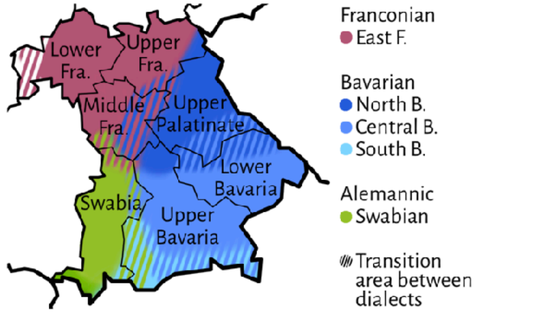
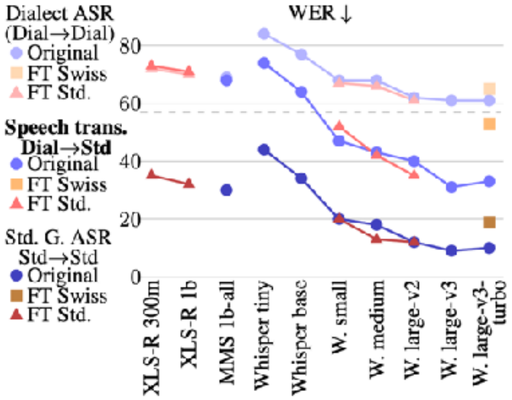


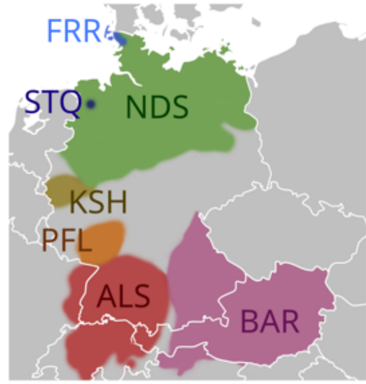
Figure 1: *The dialects and administrative regions included in Betthupferl (dialect division after [5]).*



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

https://www.isca-archive.org/interspeech_2025/blaschke25_interspeech.pdf

Multi-Dialectal German NLP benchmarks



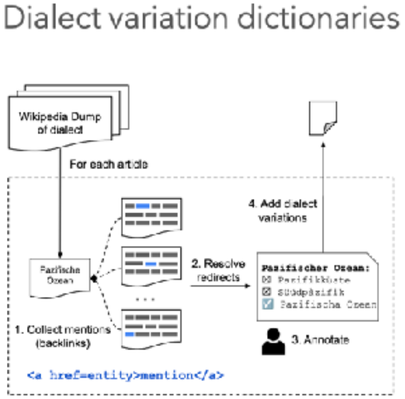
- **BLI**: Bilingual Lexicon Induction
  - Mine bilingual dictionaries (Artemova & Plank, 2023)

Bavarian	German
Eihgmoant	Eingemeindet
Sidlichn	Südlichen
Augschburg	Augsburg

https://aclanthology.org/2023.nodalida-1.39/

- **Dialect Variation Dictionaries**:
  - Mine Wikipedia for spelling variants

German	Bavarian	Translation?
zweisprachig	*zwaasprochig*	yes
(“bilingual”)	*zwaspråchig*	yes
	zwoasprachign	infl.
	dreisprochige (“trilingual”)	no
	...	...
dazwischen	*dozwischn*	yes
(“in between”)	*dawischn* (“to catch”)	no
	daktischen (“tactical”)	no
	...	...



Make Every Letter Count: Building Dialect Variation Dictionaries from Monolingual Corpora

Robert Litschko[1,2] Verena Blaschke[1,2] Diana Burkhardt[1] Barbara Plank[1,2] Diego Frassinelli[1]

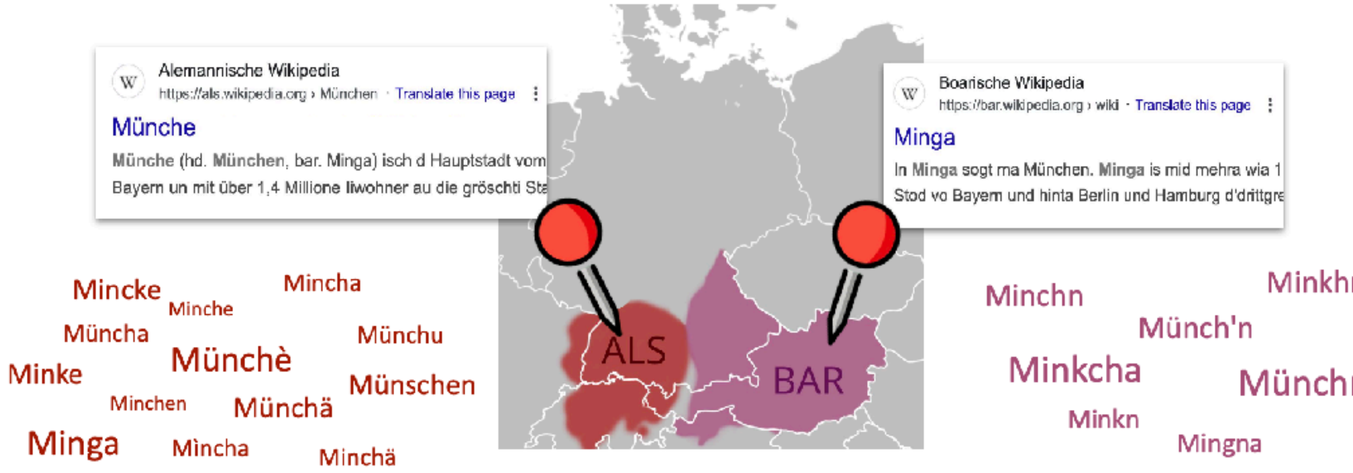
https://aclanthology.org/2025.findings-emnlp.762.pdf

- **CDIR**: Cross-dialect IR
  - How to access culture-specific information that can be found in dialect Wikis

Cross-Dialect Information Retrieval: Information Access in Low-Resource and High-Variance Languages

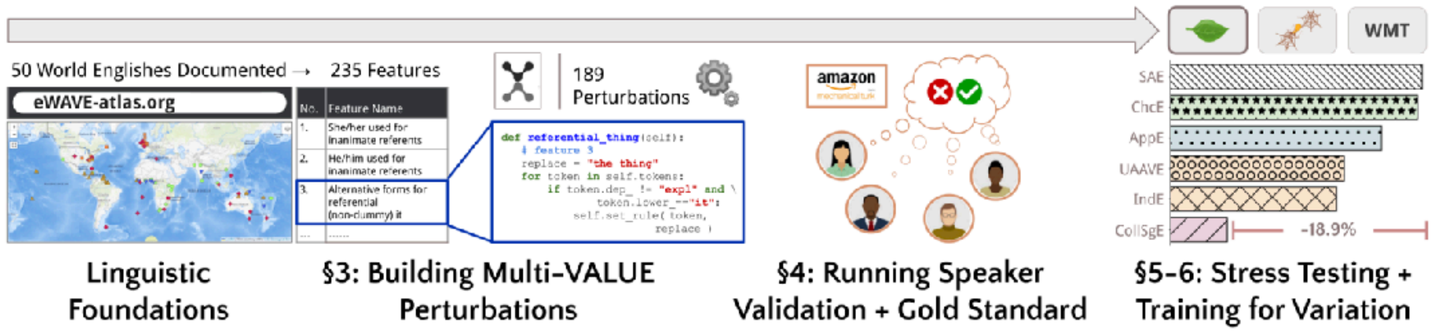
Robert Litschko[1,2] Oliver Kraus[1] Verena Blaschke[1,2] Barbara Plank[1,2]

https://aclanthology.org/2025.coling-main.678.pdf



How to evaluate systems for dialects without benchmark?

- No “gold data”
- **Idea**: linguistically-informed perturbations for robustness testing (Ziems et al., 2023 for English)

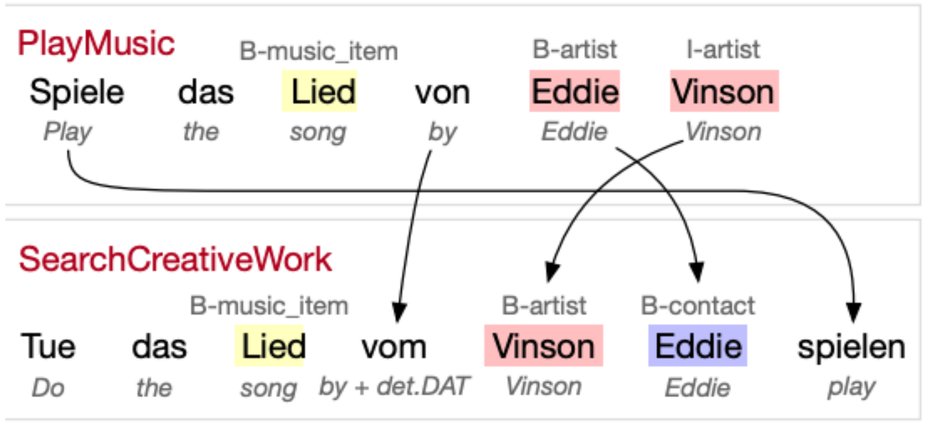


Multi-VALUE: A Framework for Cross-Dialectal English NLP

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, Diyi Yang

https://aclanthology.org/2023.acl-long.44.pdf

- Robustness Testing for German Varieties by collecting knowledge of dialect syntax
  -> to craft perturbations:



Category	Perturbation	Example: Before → After
Noun Phrase		
	`possession_von`	des Baums → von dem Baum
	von construction instead of genitive	*the*.GEN *tree's* → *of the*.DAT *tree*
	`possession_pron`	Kafkas Werke → Kafka seine Werke
	Dative with poss. pron. instead of genitive	*Kafka's works* → *Kafka*.DAT *his works*
	`article_name`	Franz Kafka → der Franz Kafka
	Article before personal names	*Franz Kafka* → *the Franz Kafka*

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

Ekaterina Artemova, Verena Blaschke, Barbara Plank

https://aclanthology.org/2024.eacl-long.28.pdf

How to evaluate systems for dialect generation?

- **Challenge**: spelling variations - high amount of variation - e.g. **Machine Translation** for language varieties without standard orthography (e.g. Aepli et al. 2023):

GSW	... ufere	Webs ii te	aa glueg e t	w ä rd e .
GSW	... ufere	Webs i te	ah gluegt	w e rd ä .
de	... auf einer	Webseite	angeschaut	werden.
en		... *viewed on a website.*		

[German]	Sie	haben
[Bavarian]	**Se**	**hom**
	They	*have*
	De	**ham**
	Dei	**hobm**

- **Idea:** contrast sets (Sun et al., 2023). Similar at surface level, semantically different. If segments paired with C are get scored higher than A,B, then the metric is said is not dialect-robust.

A: S e chs Mitarbeiter s i wäg e Verletzige behandlet worde.
B: S ä chs Mitarbeiter s y wäg Verletzige behandlet worde.

Six members of staff have been treated for injuries.

C: Sechs Mitarbeiter si wäge Verletzige **beschtraft** worde.

*Six members of staff **were punished because of** injuries.*

A Benchmark for Evaluating Machine Translation Metrics on Dialects Without Standard Orthography

Noëmi Aepli[1] **Chantal Amrhein**[1,2] **Florian Schottmann**[2,3] **Rico Sennrich**[1,4]

https://aclanthology.org/2023.wmt-1.99.pdf

How to evaluate systems for dialect to standard (e.g. ASR)?

- **Example: Automatic speech recognition/translation:** human judgements only moderately correlate with automatic metrics relative to Standard German references ($0.48 \leq |\rho| \leq 0.59$) (Blaschke et al., 2025)
  - **Error analysis** revealing valid alternatives not captured by metrics
  - We need better **metrics** that capture dialectal nuances in generation tasks

"Everybody, immediately spread out and search for Mathilda's coin, or I'll show you what's what!"

Std	Sofort	alle	ausswärmen	und		Mathildas		Geldstück	suchen,	sonst	zeige	ich	euch,	wo's		langgeht.				
	At once	*all*	*spread out*	*and*		*Mathilda.*GEN		*coin*	*search*	*else*	*show*	*I*	*you*	*where it*		*runs along.*				
Dial	Sofort	alle	ausschwärma	und	da	Mathilda	ihr	Geldstückle	sung,	sonst	zach	ich	eich,	wo	da	Bartl	an	Most	hoid.	
					the	*Mathilda*	*her*							*where*	*the*	*Barthel*	*the*	*cider*	*fetches.*	
Hypo	Sofort	alle	Ausschwärmer	und	der	Mathilda	ihr	Geldstück	lesung.	Sonst	zeig	ich	euch,	wo	der	Badl	den	Most	holt.	
	✓	✓	✗ *swarmers*	✓	✓	✓	✓	✓	✗ *reading*	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	

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

https://www.isca-archive.org/interspeech_2025/blaschke25_interspeech.pdf

Outline

Motivation: Beyond “standard" language

Part I - The Problem: Dialects & language variation

- Why are dialects challenging for NLP?
- What resources exist (for German dialects)?

Dialects often amplify issues we see in cross-lingual NLP

Part II - The How and Why: Dialect transfer & User needs

- Which transfer strategies exist across data, models, and representations?
- What do dialect speakers actually want?

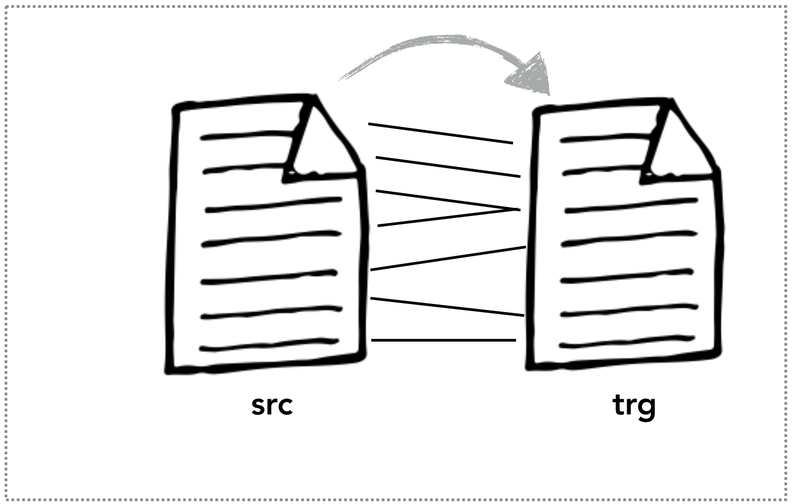
Conclusion and Outlook

Part 2: Dialect Transfer & User Needs

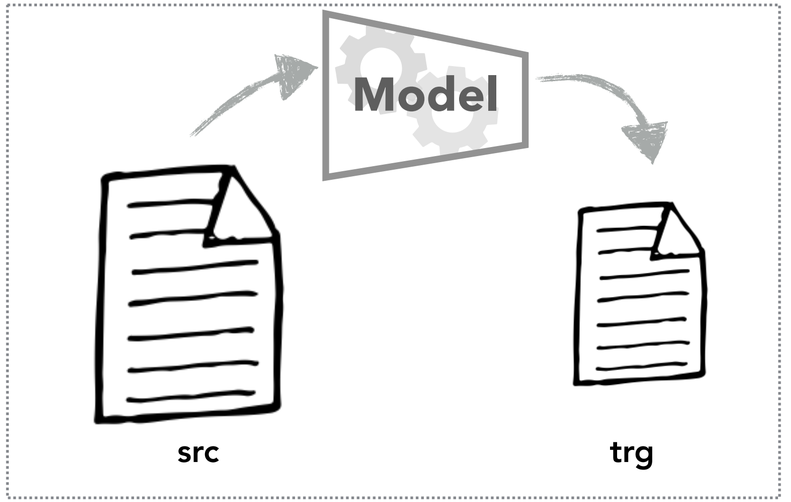
Approaches and what users want

Transfer to tackle the lack of data - Two broad approaches:

Data adaptation:

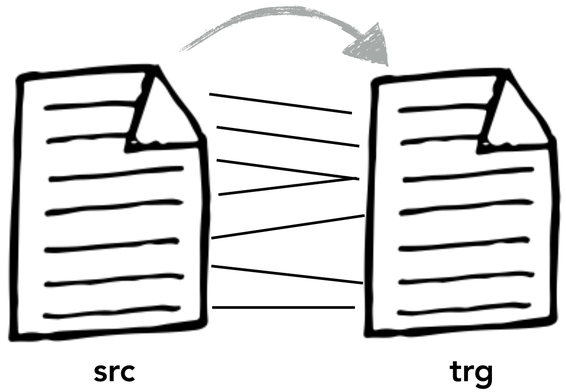


Model adaptation:

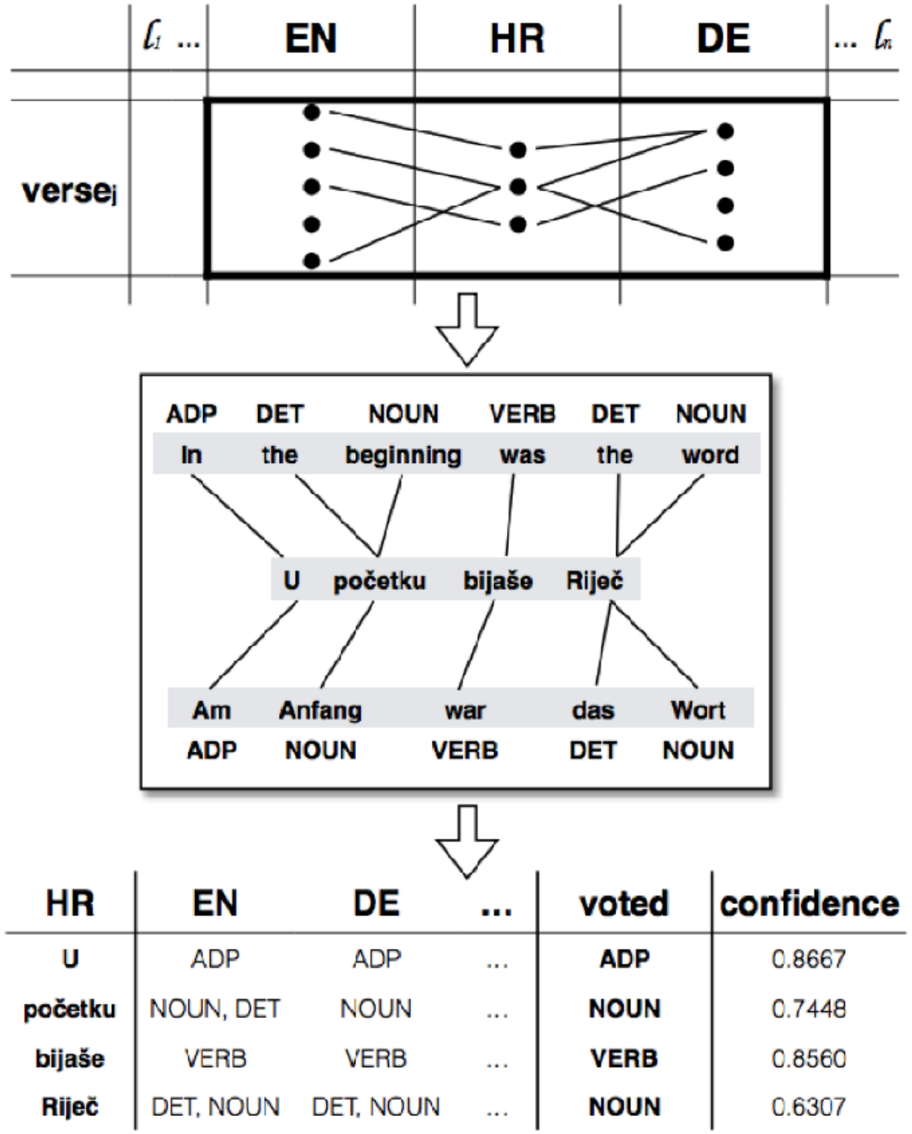


- **Data adaptation:** creates new training data
- **Model adaptation:** adjusts the model (or its inputs)

Data adaptation

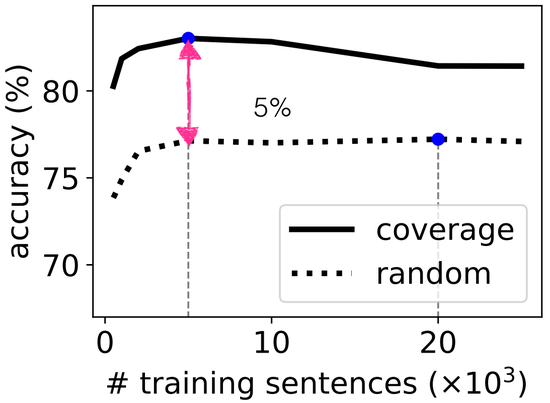


Multi-source annotation projection



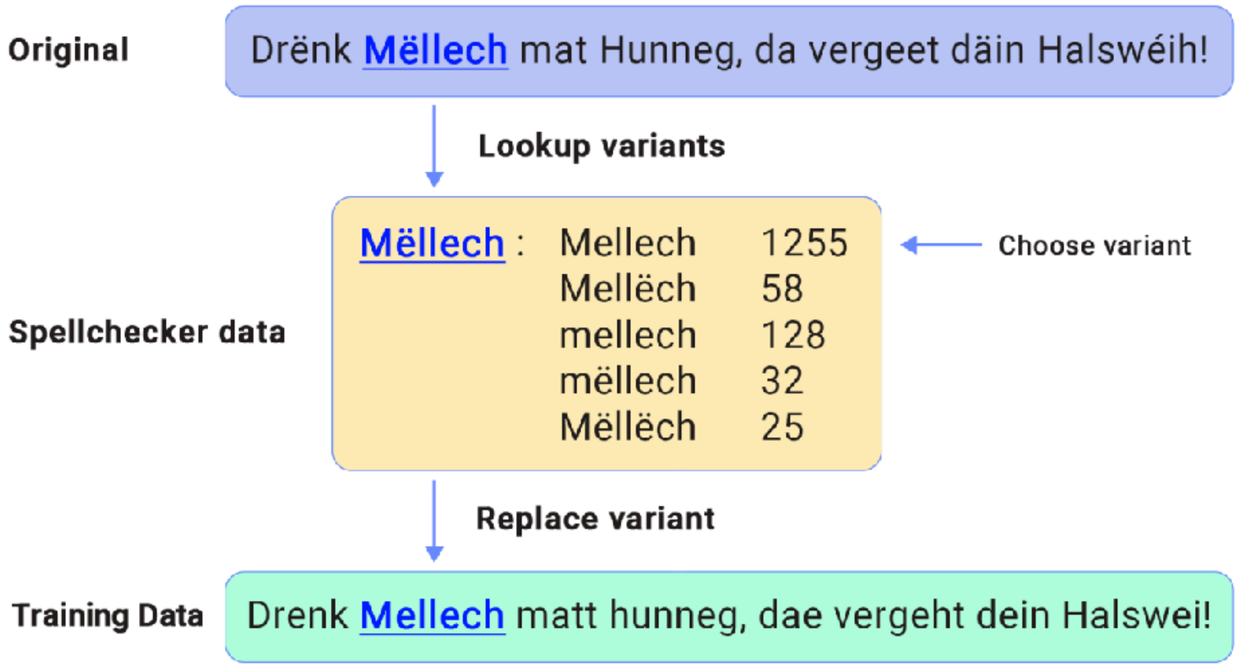
HR	EN	DE	...	voted	confidence
U	ADP	ADP	...	**ADP**	0.8667
početku	NOUN, DET	NOUN	...	**NOUN**	0.7448
bijaše	VERB	VERB	...	**VERB**	0.8560
Riječ	DET, NOUN	DET, NOUN	...	**NOUN**	0.6307

- Example: **Project** POS tags from multiple high-resource (21) source languages through parallel data to low-resource language (LRL)
  (Plank & Agic, 2018 EMNLP)
- **Select** instances by word-alignment coverage
- **Train** on projected data obtained via annotation projection
- **Findings**: less but higher-quality data is useful



Perturb

- Example: Linguistically-informed training data creation (e.g. for Luxembourgish text normalisation by Lutgen et al., 2025):
  - Use real-life spellchecker online data for data perturbation. Train normaliser on data augmented with perturbation motivated from real-world usage data



Neural Text Normalization for Luxembourgish Using Real-Life Variation Data

Anne-Marie Lutgen[1], Alistair Plum[1], Christoph Purschke[1], Barbara Plank[2,3],

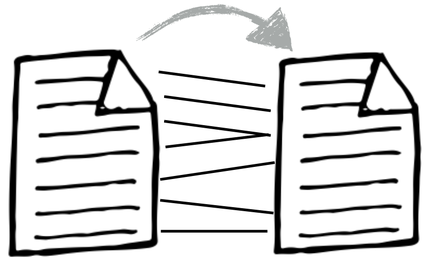
https://aclanthology.org/2025.vardial-1.9.pdf

Overview of approaches

Herzlich
Herz##lich
Herzlich w

Adaptation of representations

Data adaptation

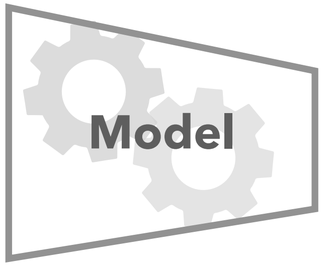


Annotation projection

Translation

(Synthetic) data augmentation

Model adaptation



FT / Aux.tasks / Adaptive / Contrastive

Prompting

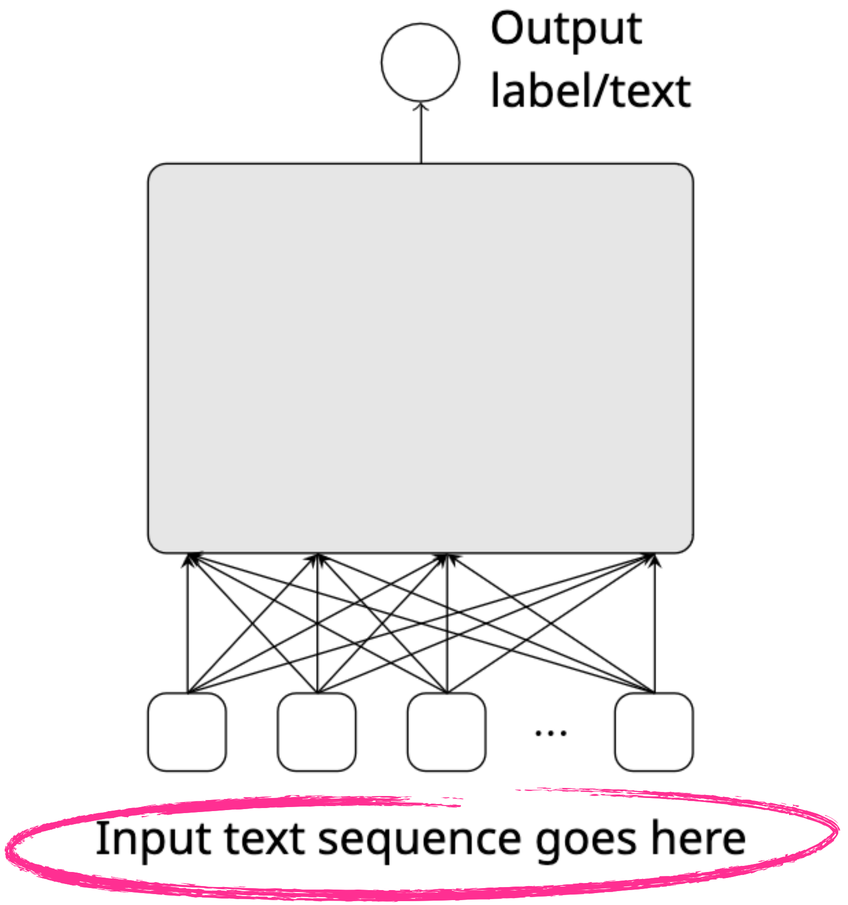
Model editing / steering

Combinations not covered: e.g. TADA (task-agnostic dialect adapters for English by Held et al., 2023) combines data synthesis with contrastive learning

Transfer strategies across data, models, and **representations**

Better input representations?

- Idea: **Change input** to make it closer reflect variation:
  - noise injection
  - perturbations
  - code-switching with dictionaries
- Other modalities:
  - visual representations
  - audio directly



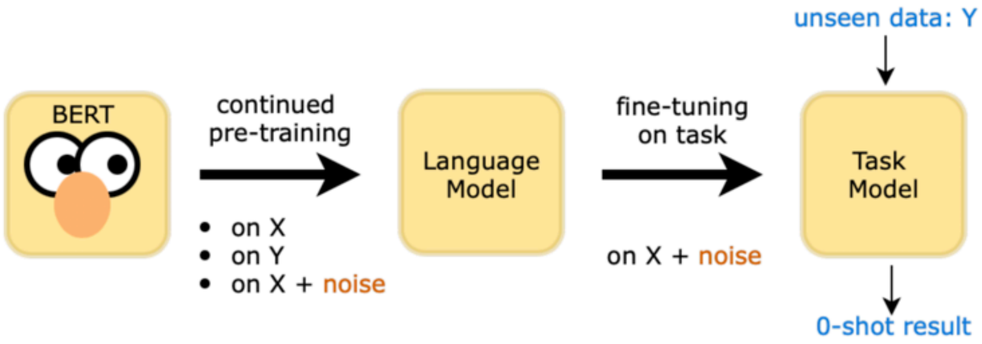
Better input representations? Option 1: Character “noise"

- Does noise injection help cross-lingual transfer?
  - Example: Inject 15% of fine-tuning words with noise (Aepli & Sennerich, 2022)

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
D(e	Lammer	hat	ein	recht		saubenes	Wasser
D (e	Lamm -er	hat	ein	recht		sau -ben -es	Wasser

Processing Swiss German - A dialect w/o standard orthography

- 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
- Aepli & Sennerich (2022) propose **noise injection** for Swiss German (POS and topic classification)
  - X=German, Y=Swiss German
- Continued pre-training on Y combined with noise injection on X worked best:



Noise	DE-BERT	DE-BERT + GSW	DE-BERT + DE	DE-BERT + DE + Noise
✗	50.66	**72.1**	52.08	53.88
✔	72.77	**82.11**	71.13	70.45

Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise

Noëmi Aepli[1] and Rico Sennrich[1,2]

- How does it work on other non-standard language varieties?

Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages

Verena Blaschke
Center for Information and Language Processing (CIS), LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
blaschke@cis.lmu.de

Hinrich Schütze
inquiries@cislmu.org

Barbara Plank
bplank@cis.lmu.de

(Blaschke et al., 2023 VarDial@EACL)

Results

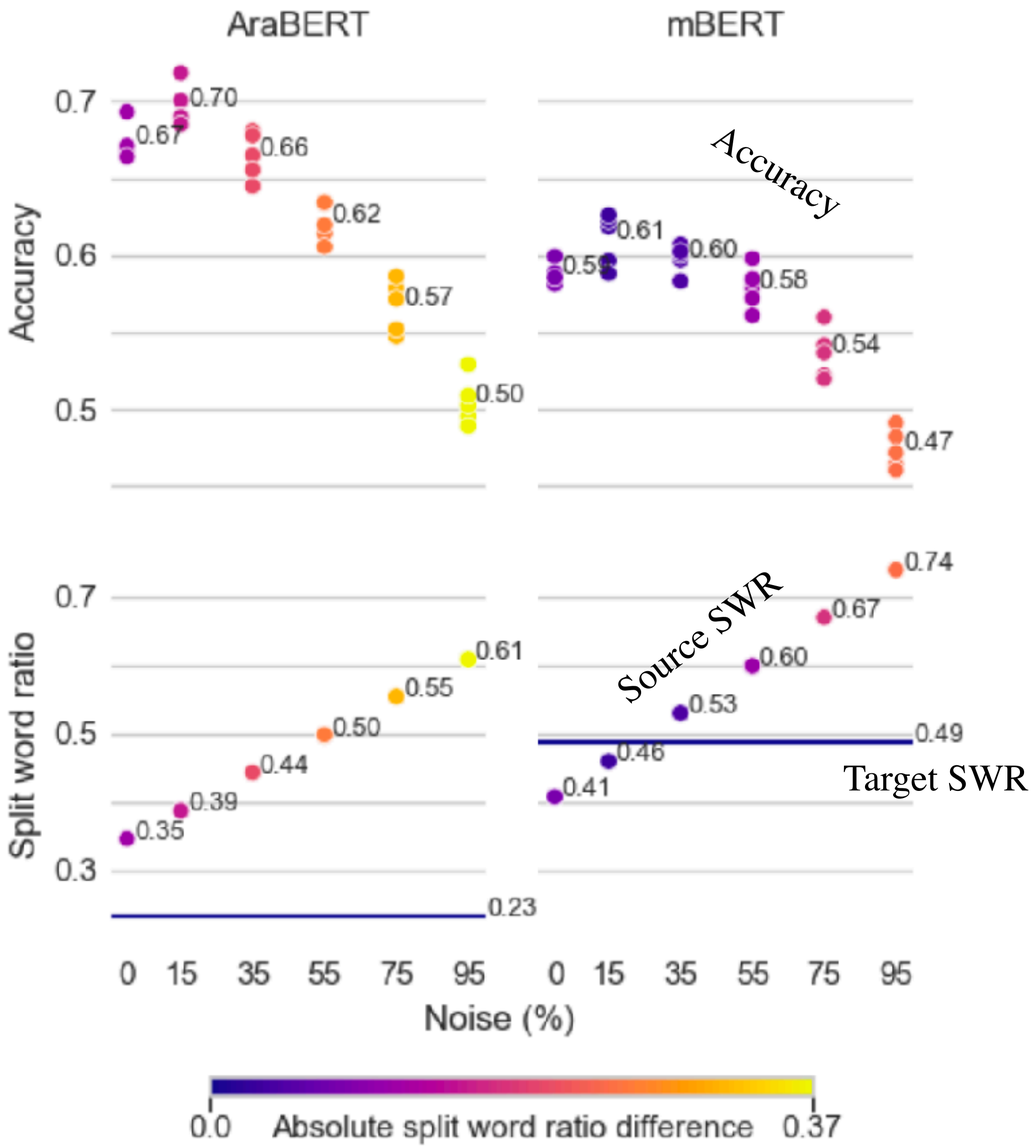
- Huge drops from standard languages (dark) to non-standards (yellowish)
- Pre-trained LM choice matters
- **Higher noise rates >15%** often help

Source	Target	Monolingual PLM						mBERT						XLM-R					
	Noise:	0	15	35	55	75	95	0	15	35	55	75	95	0	15	35	55	75	95
German	Alsatian G.	44	71	76	77	78	77	58	76	78	78	77	76	46	71	76	78	77	77
German	Swiss German	55	78	80	80	79	78	62	78	78	79	78	77	56	77	79	79	79	78
German	*German*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*	*98*
German	Low Saxon*	18	35	48	51	58	60	36	61	66	68	67	67	26	44	58	71	71	71
Dutch	Low Saxon*	52	62	63	64	64	63	73	75	75	75	73	72	63	71	73	73	73	72
Dutch	*Dutch*	*98*	*97*	*97*	*95*	*93*	*83*	*97*	*97*	*97*	*96*	*95*	*92*	*98*	*98*	*97*	*96*	*96*	*94*
Bokmål	East N.	35	60	67	65	62	60	57	60	58	57	56	54	66	63	63	62	61	59
Bokmål	North N.	36	63	69	67	65	62	61	61	61	60	60	58	70	66	66	65	64	62
Bokmål	West N.	33	59	66	63	61	59	58	57	56	55	54	53	67	62	61	60	59	57
Nynorsk	East N.	64	69	67	65	62	59	59	59	56	56	55	53	67	66	64	62	60	57
Nynorsk	North N.	67	72	69	68	65	63	62	61	59	60	59	57	71	68	67	66	64	62
Nynorsk	West N.	65	69	66	64	63	60	58	58	56	56	56	54	68	64	63	61	60	58
Bokmål	*Bokmål*	*99*	*98*	*98*	*97*	*96*	*91*	*98*	*98*	*97*	*97*	*96*	*92*	*99*	*98*	*98*	*98*	*97*	*93*
Nynorsk	*Nynorsk*	*98*	*98*	*97*	*97*	*95*	*90*	*97*	*97*	*96*	*96*	*94*	*90*	*98*	*97*	*97*	*96*	*95*	*92*
French	Picard	48	52	52	52	51	48	68	73	74	73	73	72	67	74	76	76	75	75
French	*French*	*89*	*88*	*86*	*83*	*78*	*66*	*98*	*98*	*97*	*97*	*96*	*93*	*98*	*98*	*98*	*98*	*97*	*94*
French	Occitan*	41	44	45	45	45	44	86	87	86	85	85	83	77	81	83	83	82	82
Spanish	Occitan*	62	69	70	69	69	69	83	84	83	82	81	79	72	79	78	79	78	77
Spanish	*Spanish*	*99*	*99*	*97*	*97*	*96*	*89*	*99*	*99*	*98*	*96*	*96*	*91*	*99*	*99*	*98*	*98*	*97*	*93*
MSA	Egyptian A.	67	70	66	62	57	50	59	61	60	58	54	47	64	66	65	62	57	50
MSA	Gulf Arabic	66	69	65	61	56	49	65	65	62	60	55	49	66	66	65	61	57	49
MSA	Levantine A.	64	65	62	58	53	47	56	57	55	53	50	45	59	61	60	57	53	46
MSA	Maghrebi A.	51	54	53	50	46	42	50	51	49	48	46	42	51	53	52	50	47	42
MSA	*MSA*	*94*	*93*	*89*	*83*	*78*	*67*	*96*	*95*	*91*	*85*	*79*	*69*	*96*	*95*	*91*	*86*	*80*	*70*
Finnish	Ostroboth. F.	81	80	79	77	78	75	78	78	76	74	73	70	81	85	86	86	86	84
Finnish	SE Finnish	81	79	77	75	76	73	75	75	73	70	69	66	81	84	84	84	84	82
Finnish	SW Finnish	75	73	72	71	71	70	68	68	67	64	63	61	76	80	80	81	81	79
Finnish	SW trans. area	79	78	77	76	76	74	72	72	70	68	67	65	79	84	84	85	84	83
Finnish	Savonian F.	82	80	78	76	76	73	77	79	76	73	72	69	81	84	85	85	85	83
Finnish	Tavastian F.	81	80	79	78	78	75	76	77	76	73	72	69	81	85	86	86	86	84
Finnish	*Finnish*	*98*	*98*	*98*	*97*	*96*	*94*	*96*	*96*	*96*	*95*	*94*	*93*	*98*	*97*	*97*	*97*	*96*	*94*

Table 1: **Accuracy scores (in %) by language combination, language model and noise level.** Scores are averaged over five initializations. Target languages marked with an asterisk* appear in the training data for mBERT. Rows *in*

How much noise?

- **Split word ratio** correlates best with accuracy: the smaller the difference (similar split ratios in src and target) the higher the tagging accuracy

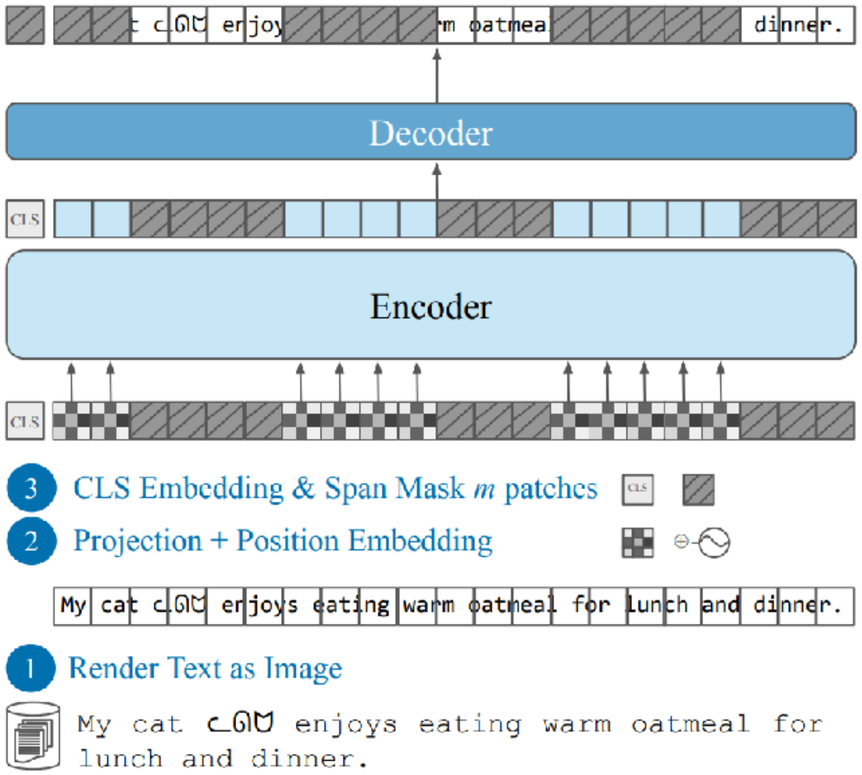


Better input representations? Option 2: Visual representations

- Language modeling with pixels (Rust et al., ICLR 2023): render text as image
- Found to be robust to orthographic attacks:

Attack	Sentence
NONE	Penguins are designed to be streamlined
CONFUSABLE	Pe𝛑guins are 𝑑esig𝛑𝑒𝑑 to be 𝑠tre𝛂mlin𝑒d
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

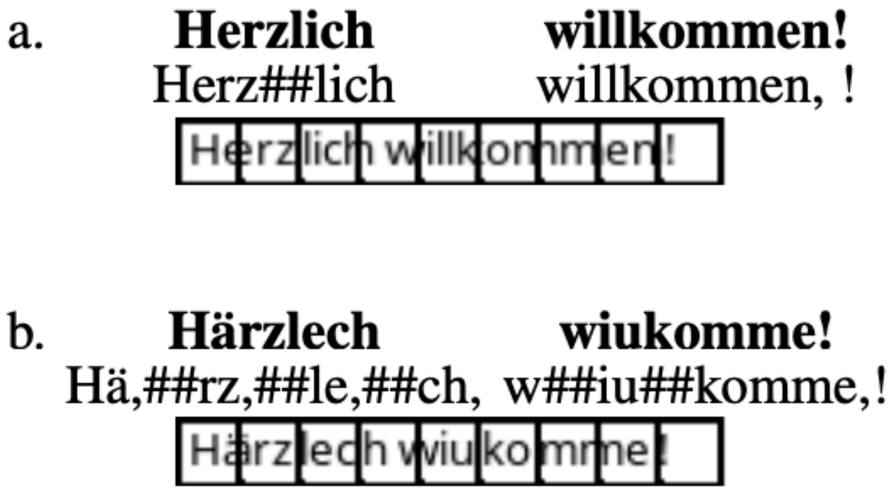
- How well does PIXEL work on non-standard languages?



(a) PIXEL pretraining

PIXEL for non-standard language varieties

- Example: We explore the potential of PIXEL-based models for transfer learning from standard to non-standard language varieties (Muñoz-Ortiz et al., 2025) using German as a case study
  - We trained a German PIXEL model and compare it to a token-based model trained on the same data
  - Tasks: POS, parsing, intent, topic detection
  - Standard vs dialect variants



Evaluating Pixel Language Models on Non-Standardized Languages

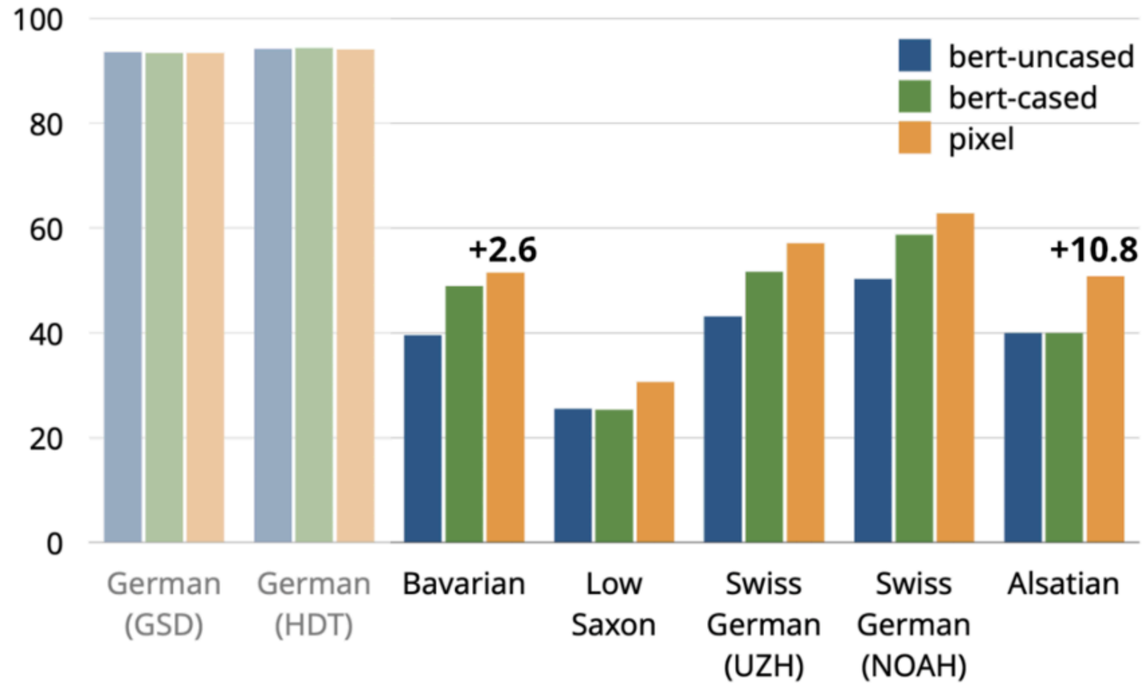
Alberto Muñoz-Ortiz **Verena Blaschke** **Barbara Plank**

https://aclanthology.org/2025.coling-main.427.pdf

German Pixel Results

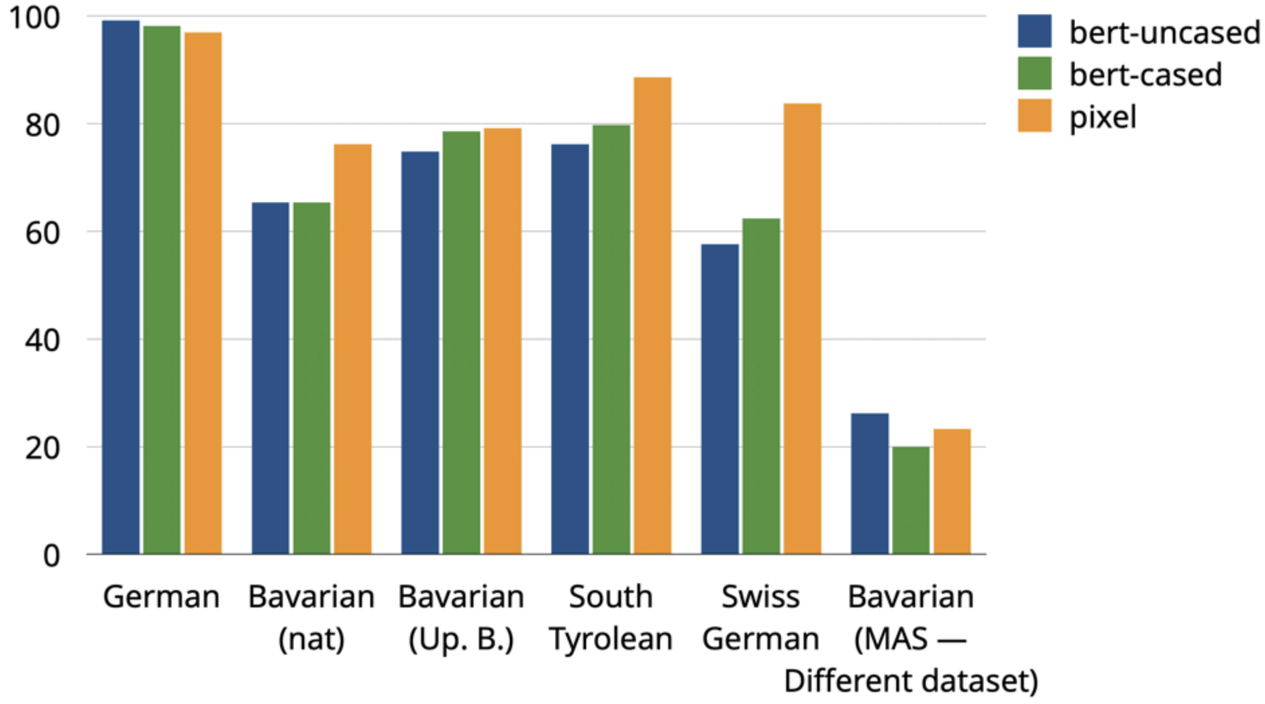
- Pixel models close on standard languages
- Works well on cross-dialectal settings: dialect syntax (POS, parsing) and intent detection, less for topic classification - worthwhile for other setups where tokenizers don't work well?

Syntax: POS



(similar results for parsing)

Intent detection



Slide by Verena Blaschke

Outline

Motivation: Beyond “standard" language

Part I - The Problem: Dialects & language variation

- Why are dialects challenging for NLP?
- What resources exist (for German dialects)?

Part II - The Toolbox: Transfer learning for dialect NLP

- Which transfer strategies exist across data, models, and representations?
- What do dialect speakers actually want?

Conclusion and Outlook

What do dialect speakers want?

Research Questions:

1. Which dialect technologies do respondents find especially useful?

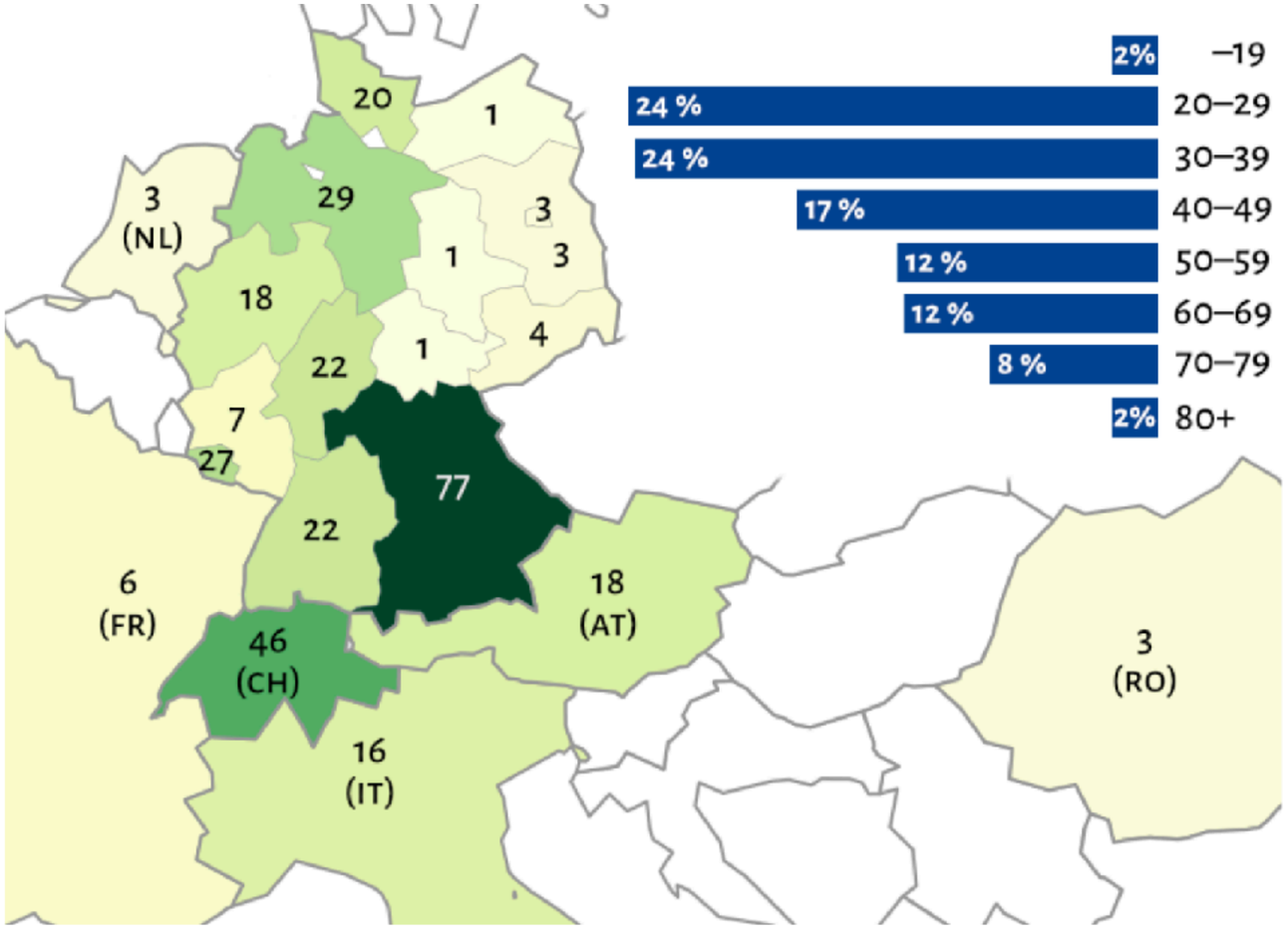
2. Does this depend on...
   - whether the input or output is dialectal?
   - whether the LT works with speech or text data?

(3. How does this reflect relevant sociolinguistic factors?) -> details in paper

(Blaschke et al., 2024 ACL)

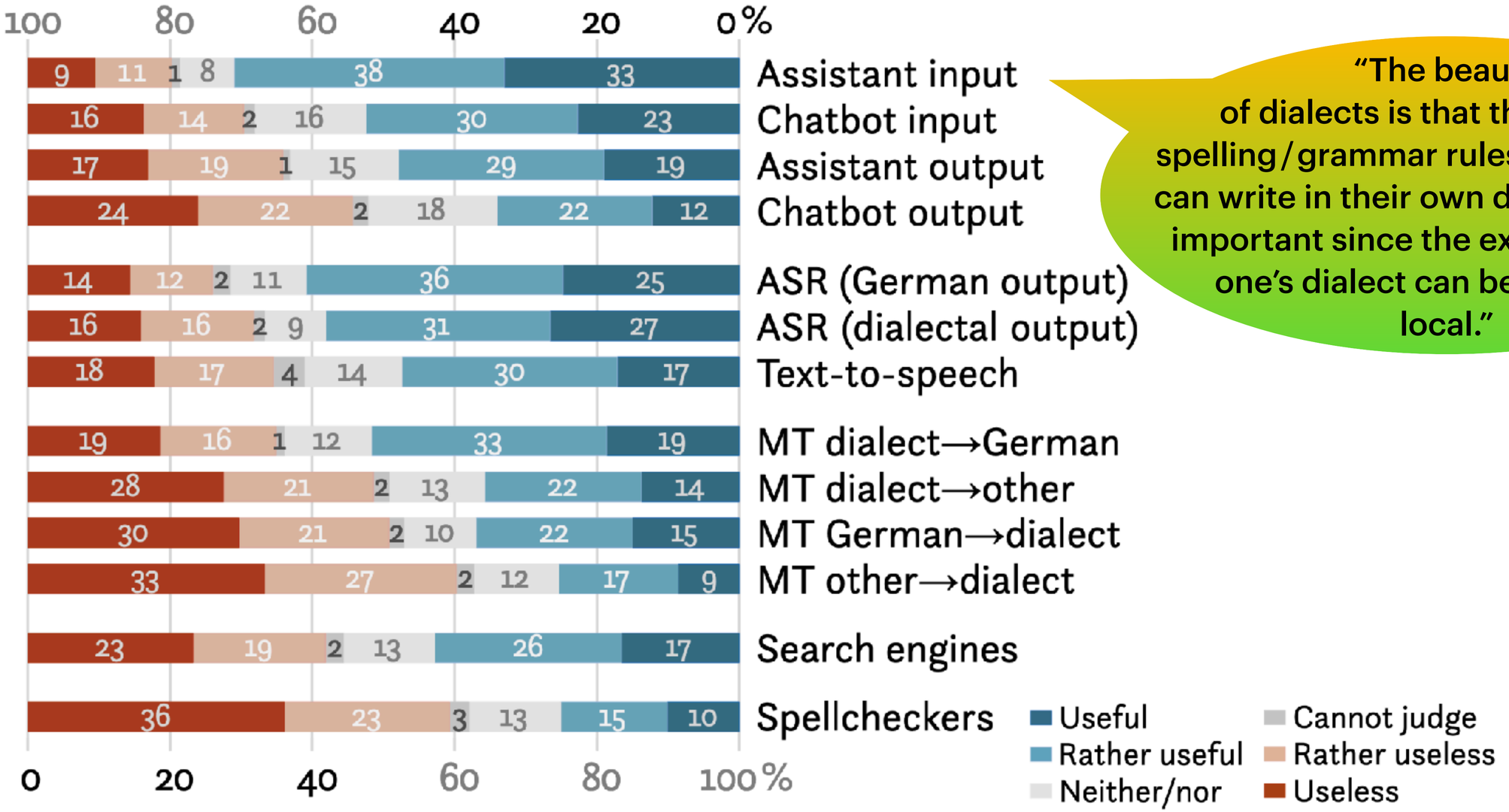
Survey respondents

- 441 respondents – **327** of whom speak a German dialect and finished the questionnaire



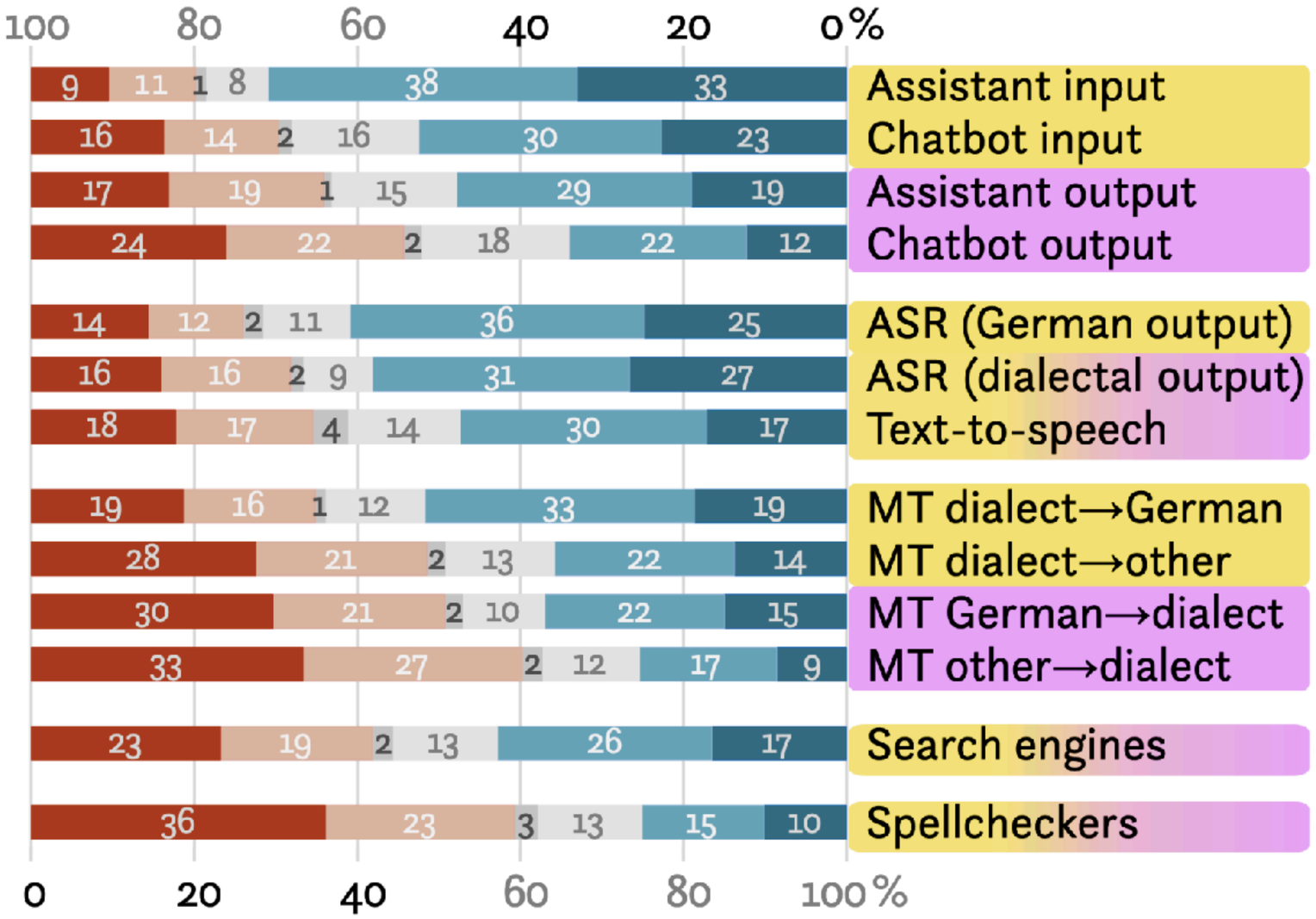
respondent age and location

Which dialect language technologies are deemed useful?

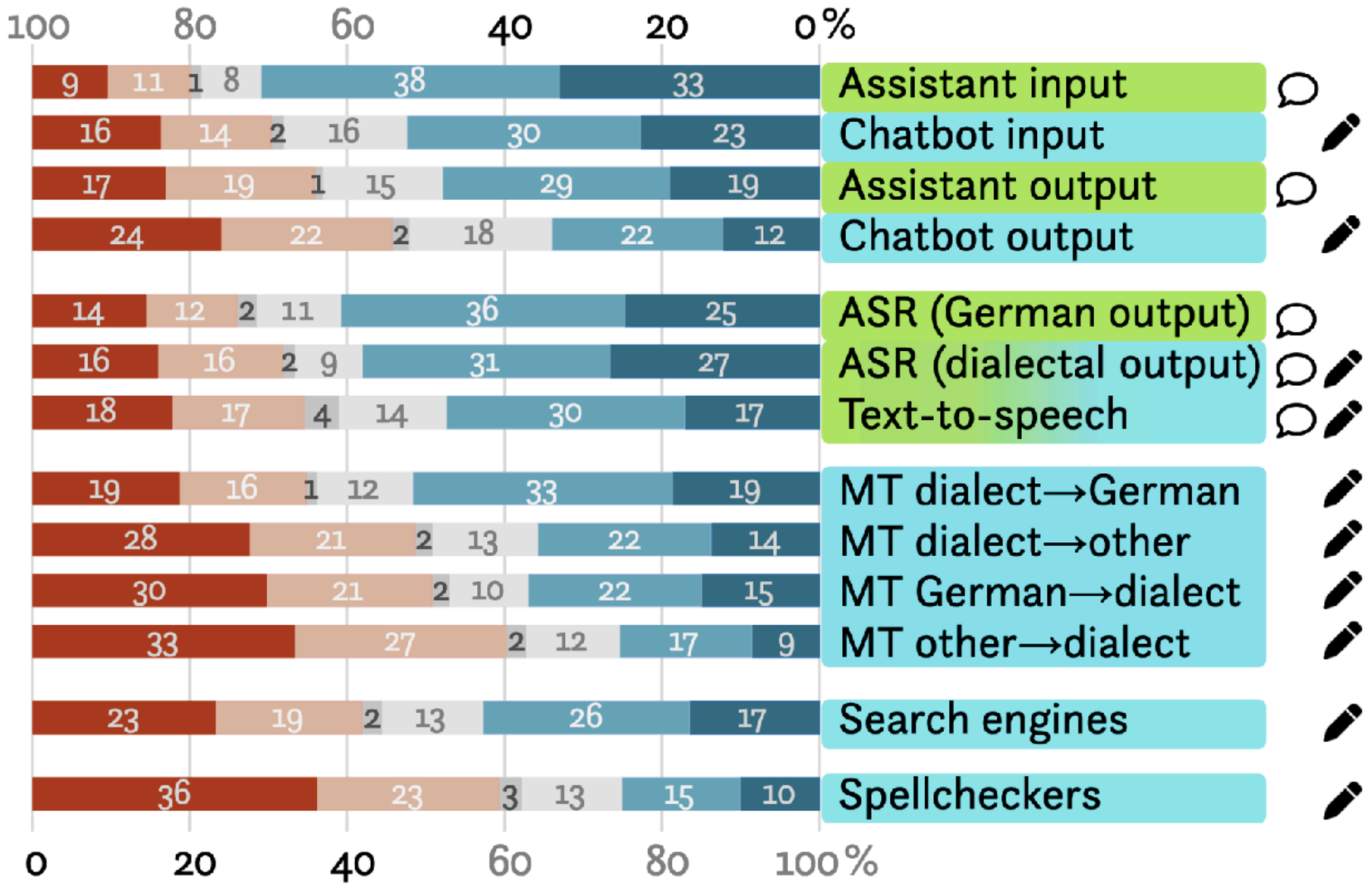


"The beauty of dialects is that there are no spelling/grammar rules and everyone can write in their own dialect, which is important since the exact version of one's dialect can be extremely local."

Dialect input vs output?



Spoken vs written dialect?



Outline

Motivation: Beyond “standard" language

Part I - The Problem: Dialects & language variation

- Why are dialects challenging for NLP?
- What resources exist (for German dialects)?

Part II - The Toolbox: Transfer learning for dialect NLP

- Which transfer strategies exist across data, models, and representations?
- What do dialect speakers actually want?

Conclusion and Outlook

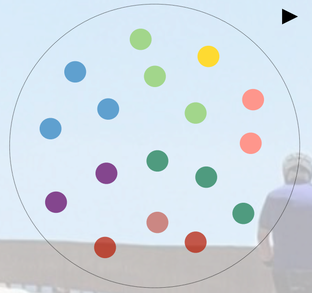
Conclusions and Outlook

Wrapping up

To wrap up:

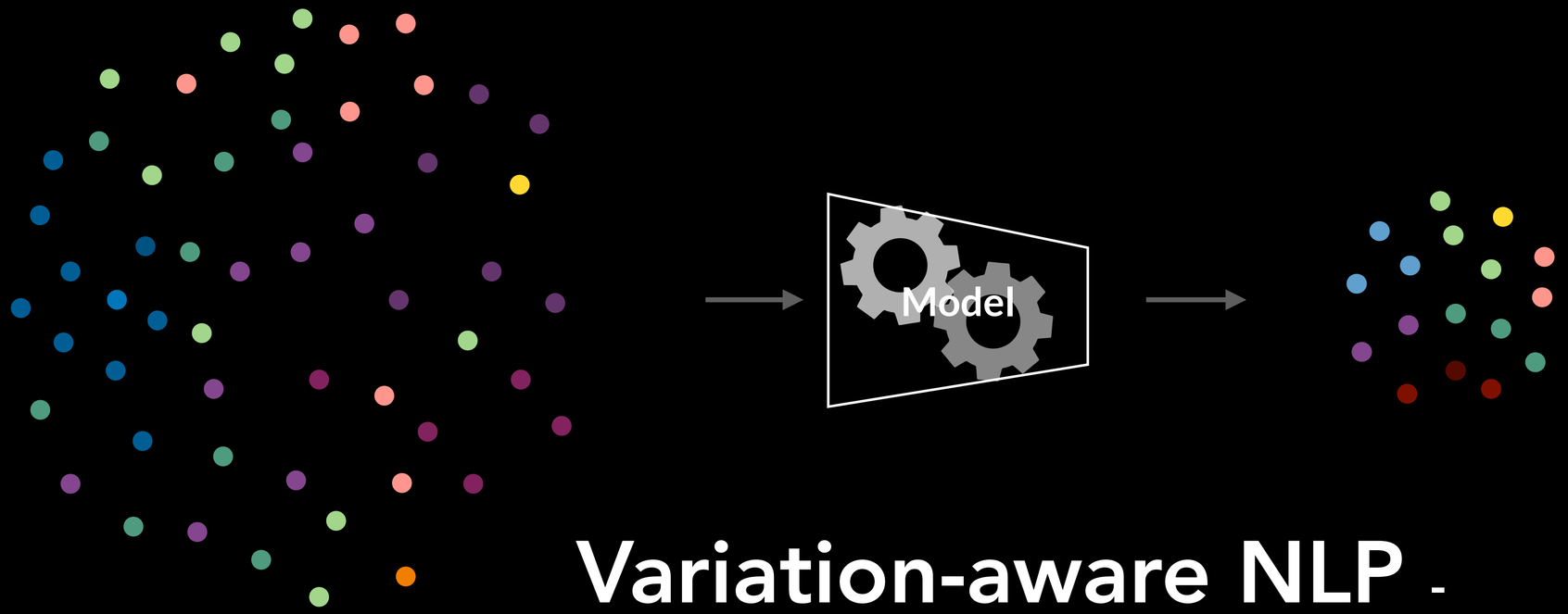
- **Dialects are not just "edge cases"**
  - Dialects force us to rethink assumptions about data, transfer, and evaluation
  - Importance to work on both written and spoken modalities

- **Transfer learning helps — but only if we respect variation**
  - No single method dominates: fine-tuning, adapters, noise, pixels, and auxiliary tasks all help in different ways



- **In future, we need more variety-aware NLP especially for low-resource language modelling**
  - Variation is the norm (see Lutgen et al., 2026 for sociolinguistic criteria and case study). Let’s embrace variation fully ❤️ and at the heart of NLP

Embrace the full spectrum of Variation : variation-aware NLP



Variation-aware NLP -

Join the Slack: https://join.slack.com/t/varnlpworkspace/shared_invite/zt-3y97wtn4k-uxKFcoKrWo9D2WKNqhqhFw

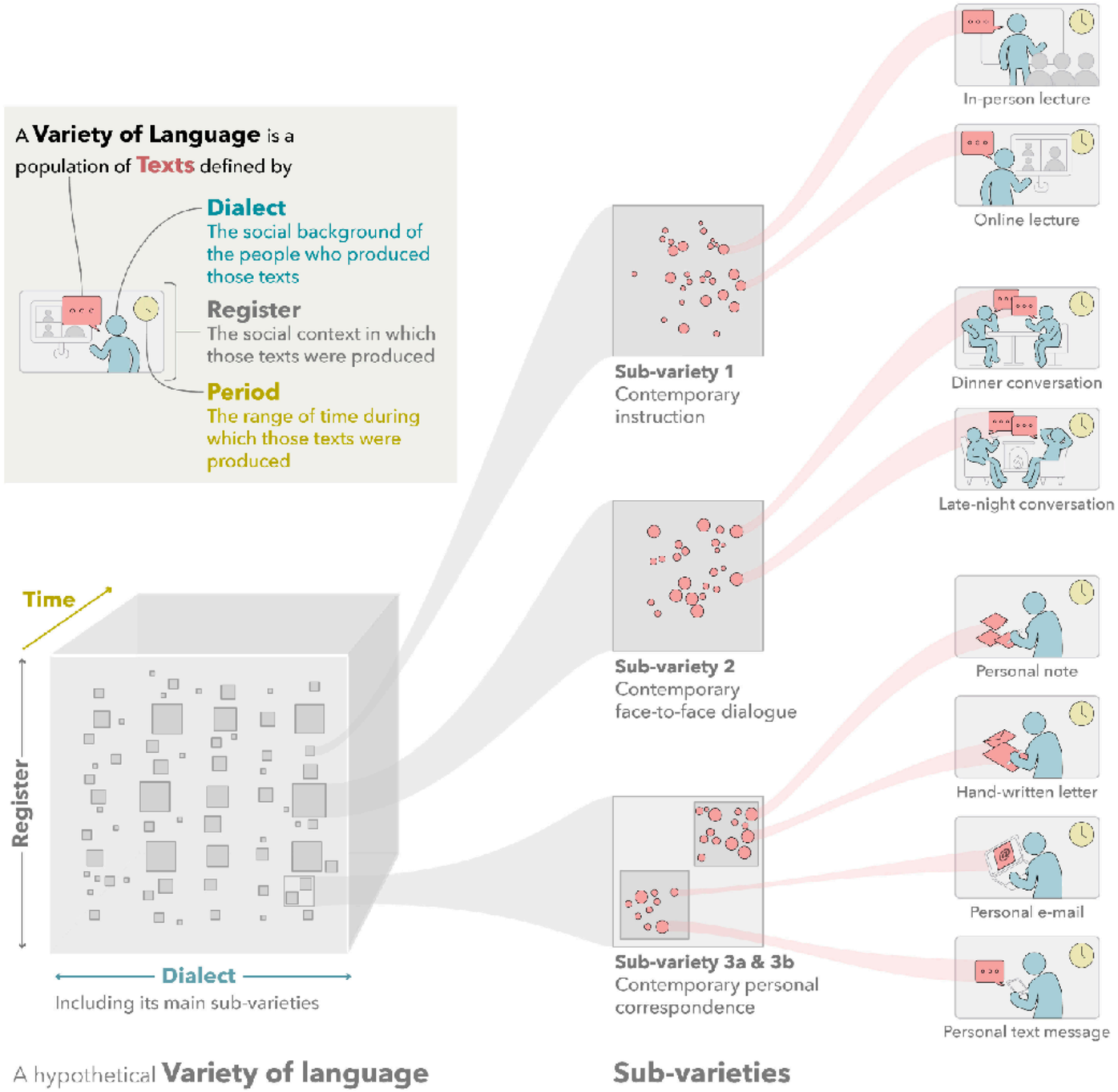
.. to connect, discuss, and
share updates on dialects and human label variation

Beyond the Standard.
Thanks to my team and collaborators.
Thank you for inviting me to
DialRes@LREC26.

Appendix

Backup Slides

Variety of Language (Grieve et al., 2025)

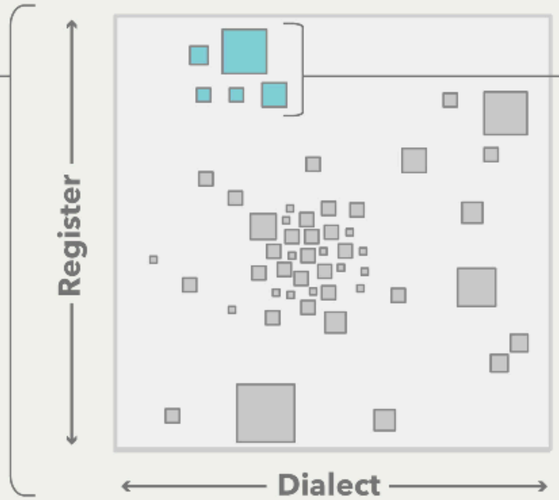


Jack Grieve et al. 2024. The sociolinguistic foundations of language modeling. Illustrator: Meike Latz

Variety of Language (Jack Grieve et al., 2025)

Variety of Language

A population of texts defined by one or more external factors including **Dialect** and **Register**



Corpus

A sample of texts that represents a variety of language, often **stratified** to represent its **sub-varieties**