

Dialect NLP:

How (and why) to process non-standard language varieties

Verena Blaschke
LMU Munich & MCML

Data Science Group, University of Hamburg
20 October 2025

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)

Dialect NLP

- ? What: Dialects and language variation
- ? How: Dialect NLP – (Some) challenges and methods
- ? For whom: Speaker perspectives

Along the way: recommendations for working with dialect data

Dialect NLP

- ❓ **What: Dialects and language variation**
- ❓ How: Dialect NLP – (Some) challenges and methods
- ❓ For whom: Speaker perspectives

Along the way: recommendations for working with dialect data

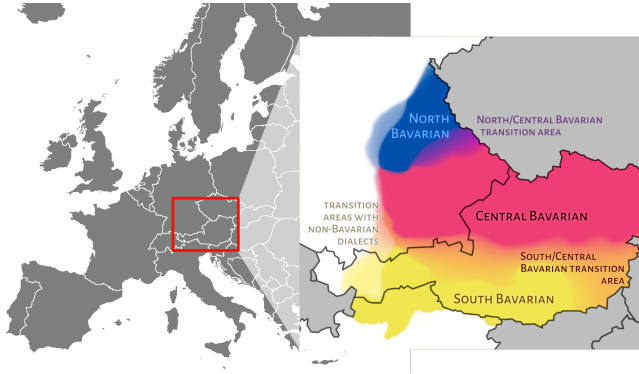
What do I mean with “dialects”?

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

Differences from the standard language

- Pronunciation (→ spelling)

[German]	Sie	haben	keine	Beine	
[Bavarian]	Se	hom	koane	Haxn	ned
	<i>They</i>	<i>have</i>	<i>no</i>	<i>legs</i>	<i>not</i>

“They [=fish] have no legs”

Linguistic differences

Differences from the standard language

- Pronunciation (→ spelling)
- Lexicon

[German]	Sie	haben	keine	Beine	
[Bavarian]	Se	hom	koane	Haxn	ned
	<i>They</i>	<i>have</i>	<i>no</i>	<i>legs</i>	<i>not</i>

“They [=fish] have no legs”

Linguistic differences

Differences from the standard language

- Pronunciation (→ spelling)
- Lexicon
- Grammar: morphology, syntax

[German]	Sie	haben	keine	Beine	
[Bavarian]	Se	hom	koane	Haxn	ned
	<i>They</i>	<i>have</i>	no	<i>legs</i>	not

“They [=fish] have no legs”

Linguistic differences

Differences from the standard language

- Pronunciation (→ 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

“They [=fish] have no legs”

Linguistic differences

Differences from the standard language

- Pronunciation (→ 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”

When do people use dialects?

- Spoken language
- Informal, written contexts (text messages, social media)
- Some literature, poetry, wikis

Losses da gut gehn in
Hamborsch & viel
Schbass!



WIKIPEDIA
Di frei Enzyklopedi

Suche (uf Hochdütsch) Suech

Hauptsite Diskussion Lese Quelltext anzeigen

Schwyzerdütsch Badisch Elsassisch Schwäbisch Vorarlbergisch

e Frichjahr fer unseri Sproch
www.kraftgute.com

Griß Godd älle midanand ond härzlich willkomma uf dr alemannische Wikipedia!
D freia Enzyklopedi, wo älle midmacha kenned.

Why dialect NLP?

- Annotate data for linguists, research variation
- Sparse & heterogeneous data for ML
- Downstream: systems for more robustly processing non-standard data
- (and more!)

Data

Challenges regarding dialect corpora

- Availability
- Quality
- Written representations

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

Verena Blaschke

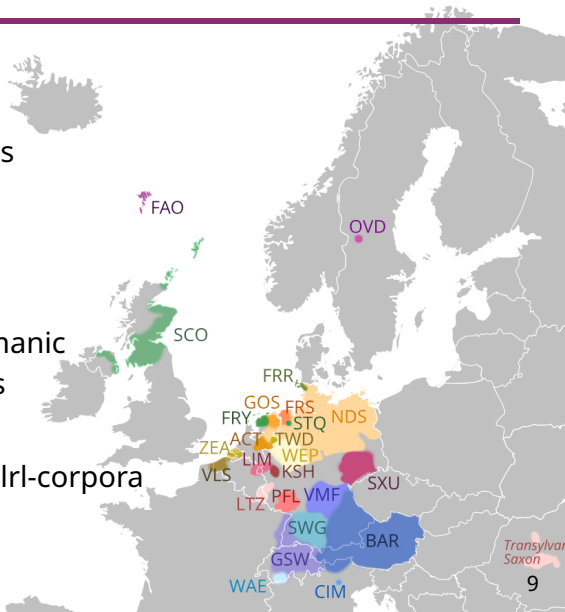
Hinrich Schütze

Barbara Plank

Datasets for Germanic low-resource varieties

- Accessible for research
- Computer-friendly formats
- Annotated + unannotated
- High-quality data (e.g., no OCR issues!)
- ➔ 100+ datasets for 35 Germanic dialects + small languages

github.com/mainlp/germanic-lrl-corpora



Annotations

What, if any, high-quality annotations do we find?

- Morphosyntax (POS tags, dependencies, phrase structure)
- Geolocation, dialect group
- Paraphrases, translations, sentiment, topics, slot and intent detection
 - Rare, but getting more popular
- Mostly: not annotated
 - ... and sometimes uncurated

Data quality: Uncurated data

Uncurated LRL data tend to be of rather low quality

– wrong language, bad data cleaning

(Kreutzer+, TACL 2022; Abadji+, LREC 2022)

OSCAR corpus (fixed subsequently)

⊙ **Scots language corpus is non linguistic?** lang:sco quality ver:21.09

#14 · Uinelj opened on Nov 4, 2021

⊙ **Quality warning: Neapolitan** lang:nap quality ver:2019 ver:21.09

#13 · Uinelj opened on Nov 4, 2021

⊙ **Quality warning: Somali** lang:so quality ver:2019 ver:21.09

#12 · Uinelj opened on Nov 4, 2021

⊙ **Quality warning: Northern Frisian** lang:frr quality ver:2019 ver:21.09

#11 · Uinelj opened on Nov 4, 2021

Data quality: Uncurated data

Uncurated LRL data tend to be of rather low quality

– wrong language, bad data cleaning

(Kreutzer+, TACL 2022; Abadji+, LREC 2022)

“West Flemish” QED OPUS corpus

```
<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>
```

Data quality: Low-status varieties prone to parodies?

Shock an aw: US teenager wrote huge slice of Scots Wikipedia

Nineteen-year-old says he is 'devastated' after being accused of cultural vandalism

Sprache [Am Gwëntext werkeln]

Das ist alles kein Bairisch, z. B. *Da ehemalige Generni Gerhard Graf von Schwarzen wurde am 24. Mai 1950 Konrad Adenauere*

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

Das das alles kein Bairisch
ersetzen und das Präteritum

"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."

Brooks/Hern, The Guardian, 2020
bar.wikipedia.org/wiki/Dischkrian:Bundeswehr

Dialectal wikis

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
pdz	PDC	2 k	27 %	92 %	6

Dialectal wikis

Wikipedia & Language		Articles	Manual edits (2022)	Manual edits (2021)	Monthly editors (2022)
Active editors! (but not that many)					
nds	NDS (Germany)* (🇩🇪)	84 k	44 %	99 %	30
lb	LTZ	41 k	19 %	85 %	56
fy	FRY	50 k	60 %	99 %	54
sco	SCO	11 k	33 %	100 %	70
als	GSW + SWG + WAB (🇨🇭)	30 k	69 %	100 %	58
bar	BAR (🇩🇪)	21 k	88 %	100 %	39
frr	FRR (🇩🇪)	17 k	79 %	85 %	16

Trend: more manual edits

Some articles: metadata regarding region, orthography

Various different policies regarding orthography

De Brukers vun de Wikipedia op Plattdüütsch hebbt utmaakt, dat se de **Sass-Schrievwies** na dat Wöörbook vun [Johannes Sass](#) (kiek ok ünner [Wikipedia:Wöörböker](#)) bruken doot.

vls	VLS (🇧🇪)	8 k	15 %	70 %	16
nds-nl	N	1 k	10 %	70 %	14
zea	Z	1 k	10 %	70 %	10
stq	S	1 k	10 %	70 %	8
ksh	K	1 k	10 %	70 %	6
pfl	PFL + OHL. RHEIN. FRANK., HESSIAN (🇩🇪)	5 k	50 %	100 %	6
pdc	PDC	2 k	27 %	92 %	6

Jrundsätzlich [[der Quälltäx ändere](#)]

Jeder schriev, wie em de Fingere jewaaße sin. |

How to represent a primarily spoken language?

- Normalized text (closely related standard language)

Etter litt godsnakk kom tre av kyrne ...

[After some coaxing, three of the cows came ...]

NB Tale

Norwegian

können sie ihre jugendzeit beschreiben

[Can you describe your youth?]

ArchiMob

Swiss German

How to represent a primarily spoken language?

- Normalized text (closely related standard language)
- Phone[m/t]ic transcriptions

""{t@4 l"it g""u:snAkk k"Om t4"e: "A:v C"y:n'@ ...

Etter litt godsnakk kom tre av kyrne ...

[After some coaxing, three of the cows came ...]

NB Tale

Norwegian

chönd sii iri jugendziit beschriibe

können sie ihre jugendzeit beschreiben

[Can you describe your youth?]

ArchiMob

Swiss German

How to represent a primarily spoken language?

- Normalized text (closely related standard language)
- Phone[m/t]ic transcriptions
- (More or less widely spread) orthographies

Nu leyt em de böyse vynd disse nacht ... UD LSDC
[Now, this night, the wicked enemy let them...] Low Saxon

How to represent a primarily spoken language?

- Normalized text (closely related standard language)
- Phone[m/t]ic transcriptions
- (More or less widely spread) orthographies
- Ad-hoc spellings

Nu leit em de baise Find düse Nacht ...

Nu leyt em de böyse vynd disse nacht ... UD LSDC

[Now, this night, the wicked enemy let them...] Low Saxon

How to represent a primarily spoken language?

- Normalized text (closely related standard language)
- Phone[m/t]ic transcriptions
- (More or less widely spread) orthographies
- Ad-hoc spellings

→ A tool that works for one type of written representation doesn't necessarily work for the others too

Recommendations

Blaschke, Schütze & Plank (NoDaLiDa 2023)

"A survey of corpora for Germanic low-resource languages and dialects"

... for *using* dialect corpora

- Check the quality!
- Suitable written representation for your purposes?
- Data scarcity – overlaps between (pre-)training, dev, test data?
- Data outside traditional NLP venues

... for *creating* dialect corpora

- Document the transcription guidelines / orthographies
- Share metadata like corpus size, data sources, license
- Archives for long-term storage (CLARIN, LRE Map, Zenodo)

Overview

- ! What: Dialects and language variation
- ? **How: Dialect NLP – (Some) challenges and methods**
- ? For whom: Speaker perspectives

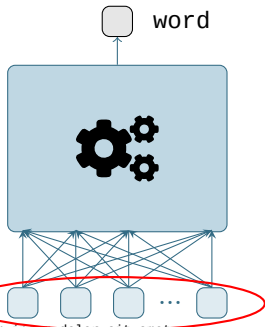
Linguistic differences

Differences from the standard language in

- **Pronunciation (→ spelling)**
- Lexicon
- Morphology
- Syntax
- Usage context

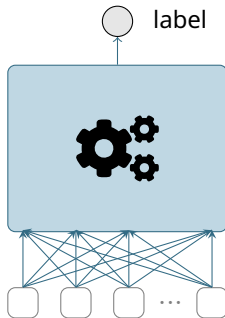
Cross-dialectal transfer

❌ Pretraining



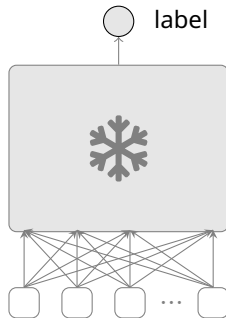
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum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

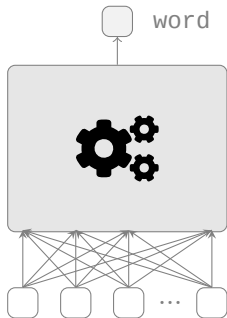
“The Lammer (river) has fairly clean water”

Sentence via bar.wikipedia.org/wiki/Låmma

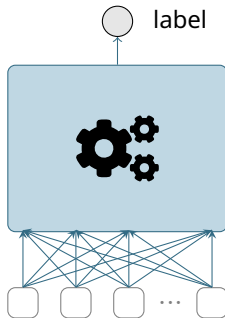
GBERT: Chan+, COLING 2020, “German’s next language model”

More robust input representations?

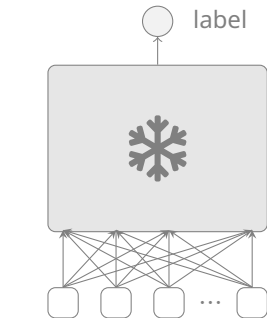
Diff. architecture



Change finetuning



Task-specific input
text



Input text in related
dialect

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,

Character-level “noise”

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



Inject 15% of fine-tuning words with “noise”

“Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise”

Aepli & Sennrich (ACL Findings 2022)

Character-level “noise”

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

 Inject 15% of words with “noise” (Aepli/Sennrich 2022)

How many words should we modify this way?

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

Verena Blaschke

Hinrich Schütze

Barbara Plank

Noise levels

0 %: The UHH DS group investigates Conversational AI systems with a focus on effective, fair, inclusive, and sustainable communication, incorporating interdisciplinary perspectives.

15 %: The UHH DS group investigates Conversational AI systems with a focus on effective, fair, inclusive, and sustainable communication, incorporating interdisciplinary perspectives.

...

95 %: The UHH DS group investigates Conversational AI systems with a focus on effective, fair, inclusive, and sustainable communication, incorporating interdisciplinary perspectives.

Dialect POS tagging

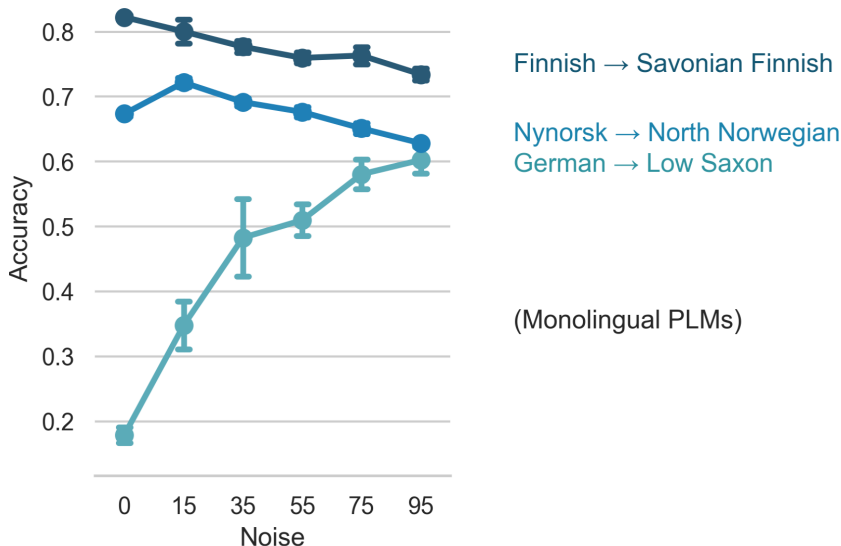
- Part-of-speech tagging
- Transfer from closely related standard languages to...
 - 3 dialects / regional languages of Germany
 - 3 Norwegian dialects
 - 2 regional languages of France
 - 6 Finnish dialects
 - 4 Arabic varieties

→ *Consistent performance drops (standard/dialect)*
- Monolingual BERTs/RoBERTas vs. XLM-R vs. mBERT

→ *Optimal choice varies across languages*
- Noise: Modify {0, 15, 35, 55, 75, 95}% of words

→ *Optimal choice varies across languages/models*

How much noise to add?

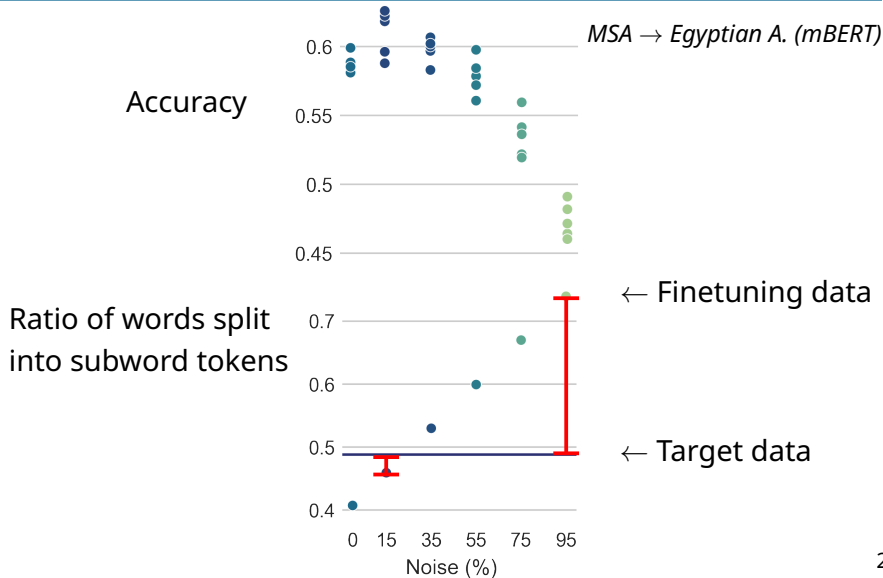


What explains this?

The more similar the word-splitting rates are, the better the results!

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		saubere s	Wasser
D (e	Lamm -er	hat	ein	recht		sau -ben -es	Wasser

Noise injection & subword tokenization



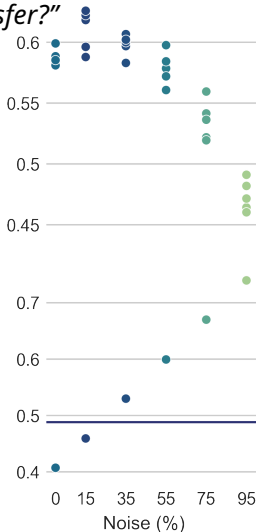
Recommendation

Blaschke, Schütze & Plank (VarDial @ EACL 2023)

"Does manipulating tokenization aid cross-lingual transfer?"

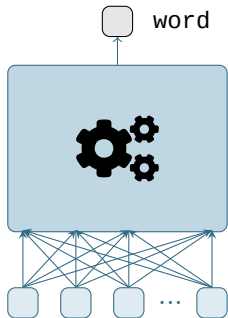
Spearman's rank correlation of > 0.8 for most models/languages (often > 0.95)!

- Don't want to tune noise level as a hyperparameter?
→ Compare the *split word ratios* for different noise levels
+ pick the noise level with the smallest difference
- Otherwise: start low + increase noise until dev accuracy drops

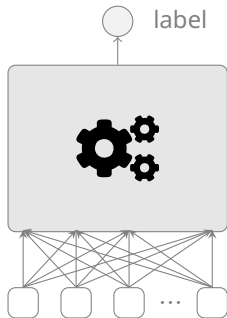


More robust input representations?

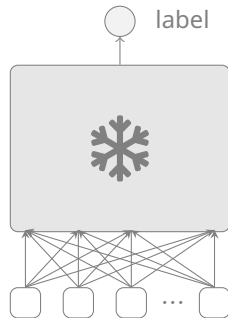
Diff. architecture



Change finetuning



Task-specific input
text

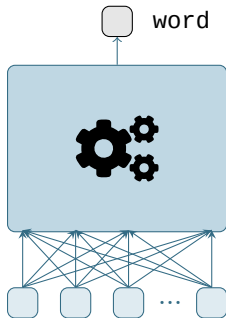


Input text in related
dialect

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,

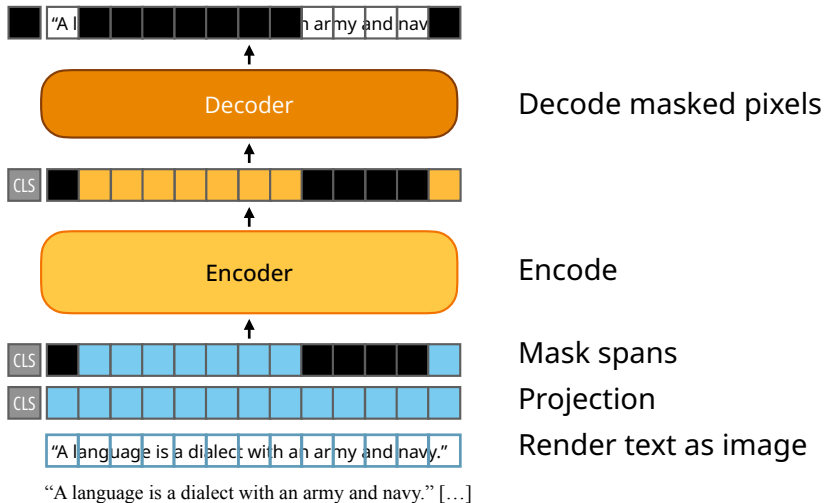
More robust input representations?

Diff. architecture

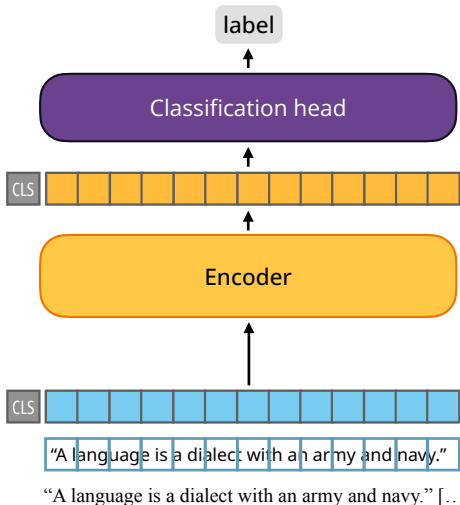


“Language modelling with pixels”
Rust, Lotz, Bugliarello, Salesky,
de Lhoneux & Elliott (ICLR 2023)

Pixel models (Rust+, 2023) – pretraining



Pixel models (Rust+, 2023) – fine-tuning



Text rendering can be adjusted for word-level tasks

"	A	language	is	a	dialect	wit
---	---	----------	----	---	---------	-----

Pixel models – robustness

(English) Pixel generally more robust against orthographic attacks than BERT

Attack	Sentence
NONE	Penguins are designed to be streamlined
CONFUSABLE	P enguin s are d esign e d to be s tre a mlin e d
SHUFFLE (INNER)	Pegnuins are dnesiged to be sieatrnmlnd
SHUFFLE (FULL)	ngePnius rae dsgednei to be etimaslernd
DISEMOVOWEL	Pngns r dsynd 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

Pixel models – robustness

Die	Lam	mer	hat	ein	recht	saub	eres	Wasser
-----	-----	-----	-----	-----	-------	------	------	--------

D'	Lomma	hod	a	rechd	a	sauwas	Wossa
----	-------	-----	---	-------	---	--------	-------

Evaluating Pixel Language Models on Non-Standardized Languages

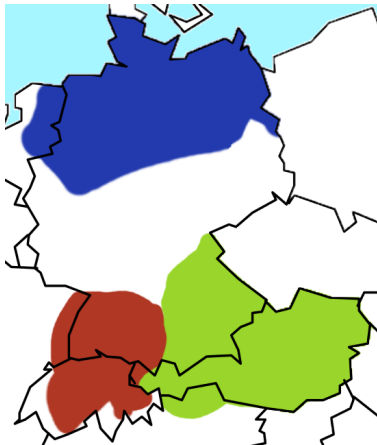
Alberto Muñoz-Ortiz 

Verena Blaschke  

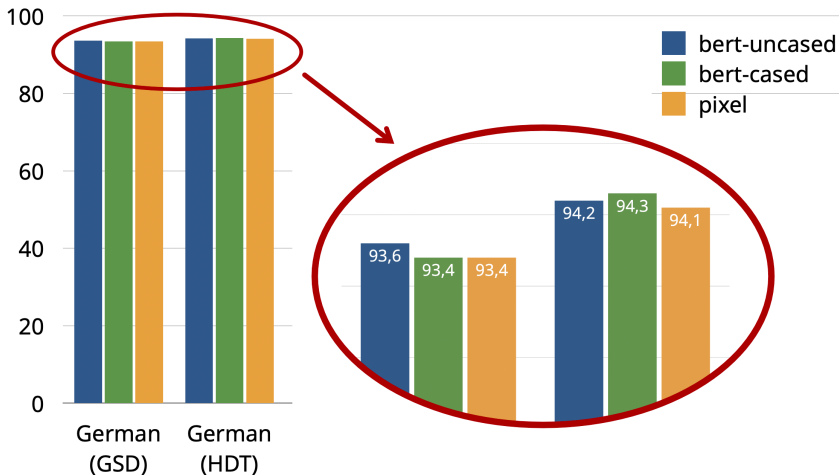
Barbara Plank  

German Pixel experiments

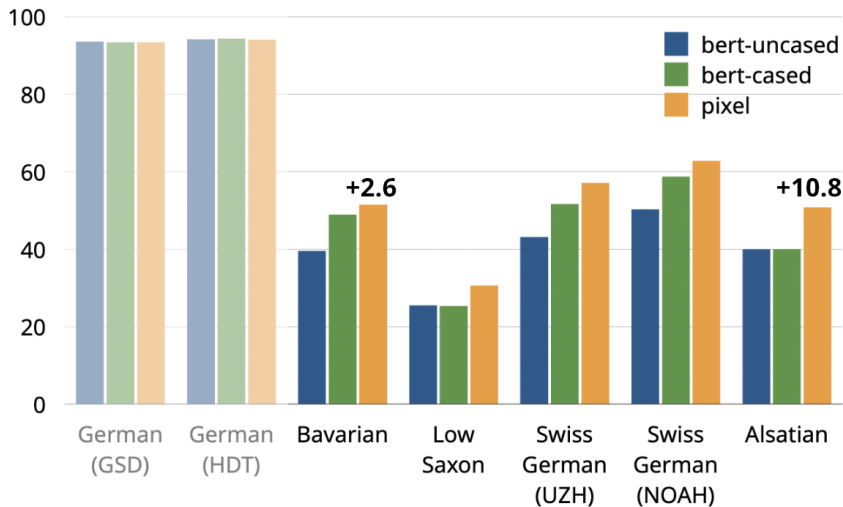
- German Pixel model (new!)
 - Same training data as a German BERT model
- Finetune on German, evaluate on dialects/regional languages
- 2 grammatical tasks:
POS tagging, parsing
- 2 semantic tasks:
intent classification (easy),
topic classification (harder)



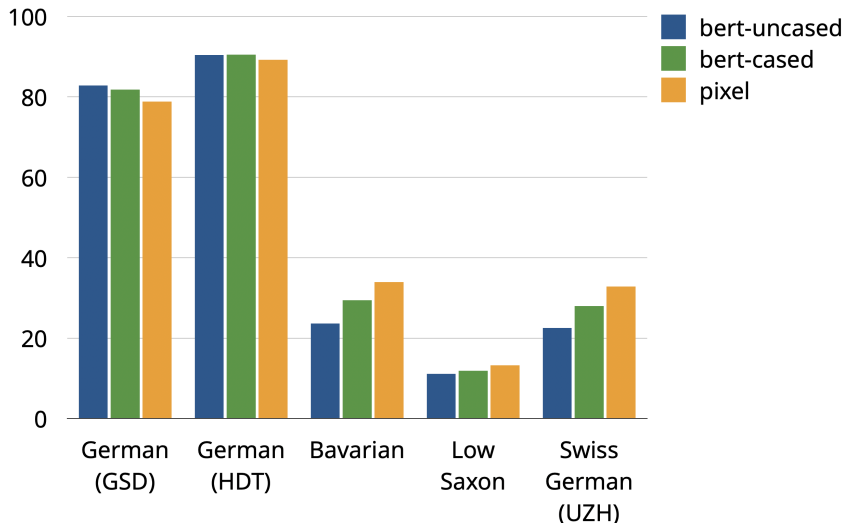
German Pixel: POS tagging (accuracy)



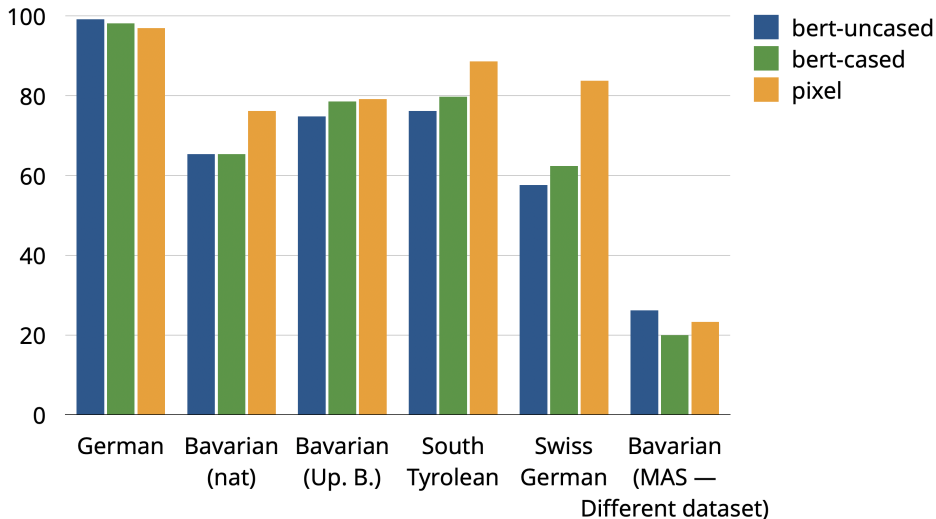
German Pixel: POS tagging (accuracy)



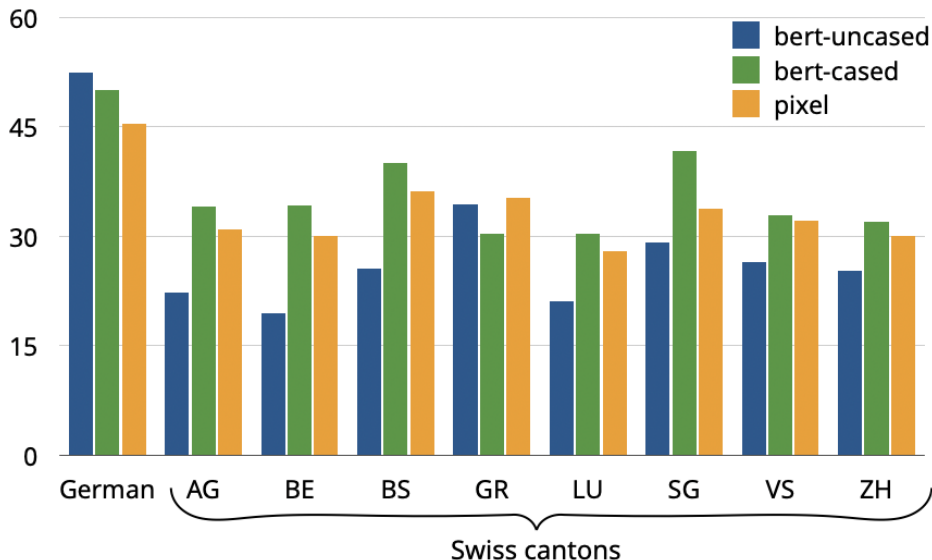
German Pixel: Parsing (LAS)



German Pixel: Intent classification (accuracy)



German Pixel: Topic classification (accuracy)



Pixel: Trade-off

Muñoz-Ortiz, Blaschke & Plank (COLING 2025)

"Evaluating pixel language models on non-standardized languages"

- More compute needed
 - On par with or worse than BERT in monolingual settings (+ where std language performance is bad)
 - Cross-dialectal settings / settings with less predictable spelling might be the place to shine
- Worthwhile for other settings where tokenizers don't work well?

Overview

- ! What: Dialects and language variation
- ! How: Dialect NLP – (Some) challenges and methods
- ? For whom: **Speaker perspectives**

Linguistic differences

Differences from the standard language in

- Pronunciation (→ spelling)
- Lexicon
- Morphology
- Syntax
- **Usage context**

Why dialect NLP?

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

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

What Do Dialect Speakers Want?

A Survey of Attitudes Towards Language Technology for German Dialects

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

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.

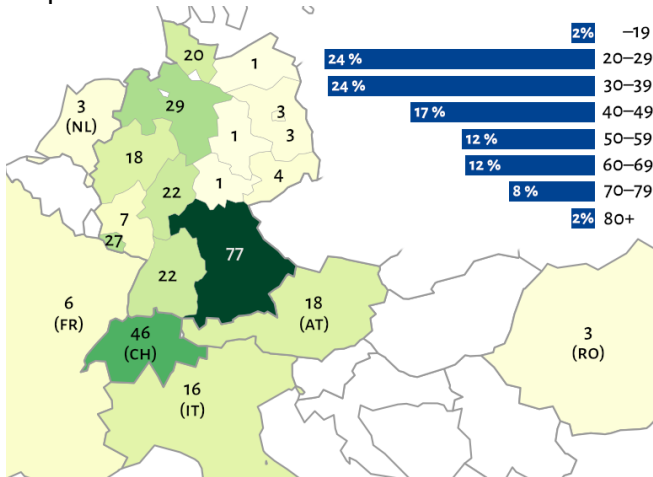
Do you agree with the following statements?

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.

Dialect background and attitudes

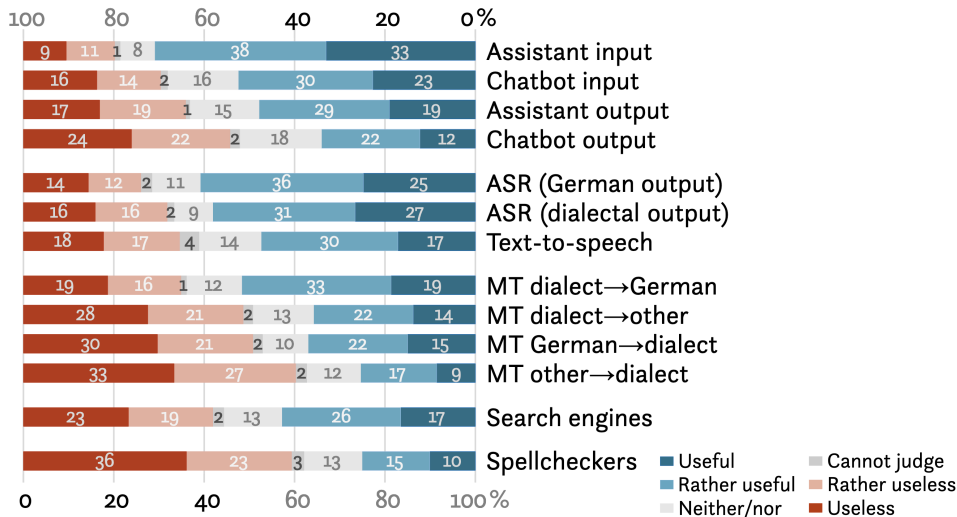
441 respondents – 327 of whom speak a German dialect and finished the questionnaire



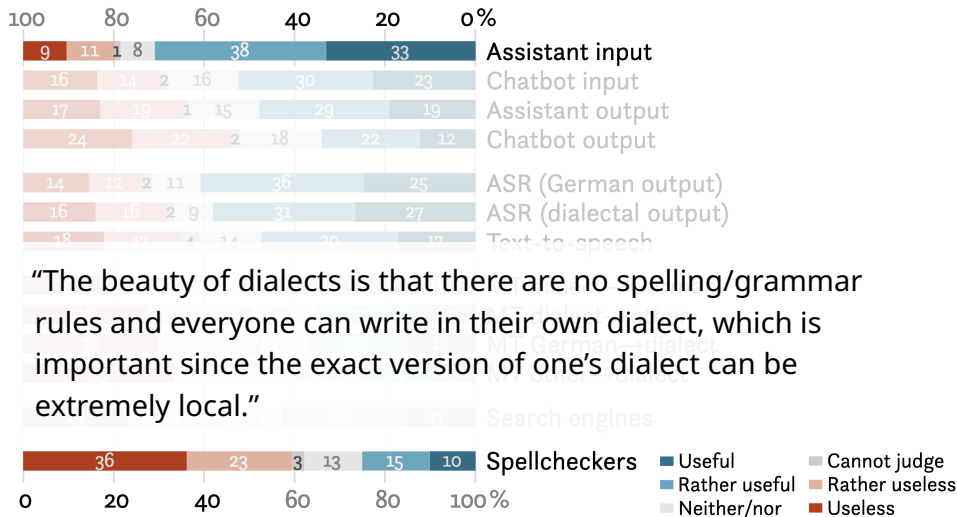
Dialect background and attitudes

- 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?

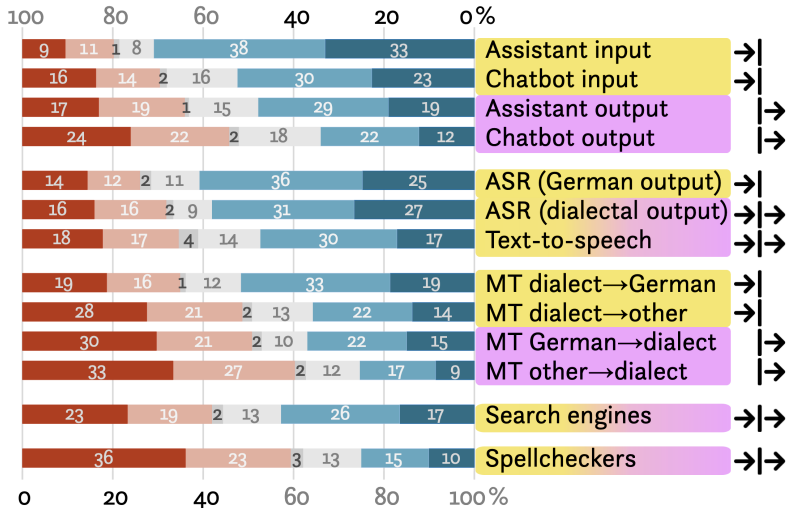


Which dialect LTs are deemed useful?

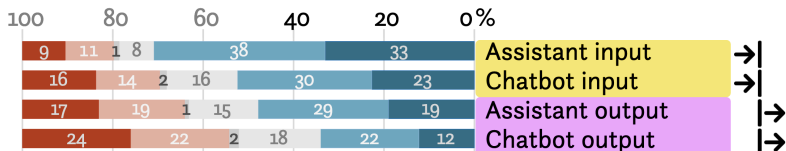


"The beauty of dialects is that there are no spelling/grammar rules and everyone can write in their own dialect, which is important since the exact version of one's dialect can be extremely local."

Dialect input vs. output?

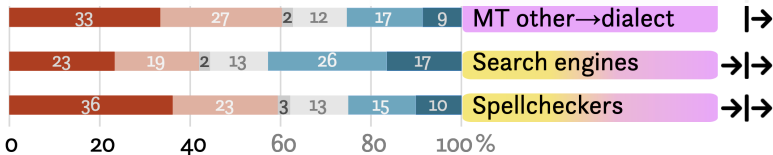


Dialect input vs. output?

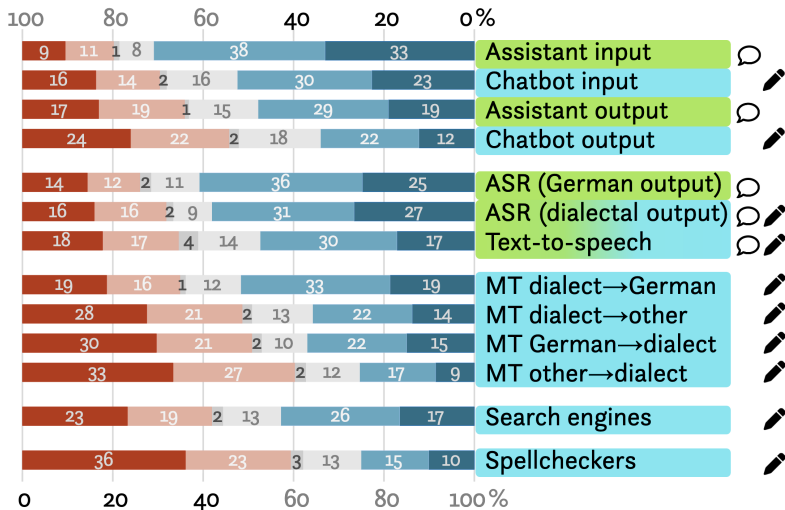


"It might be annoying if the output is slightly different from your own dialect."

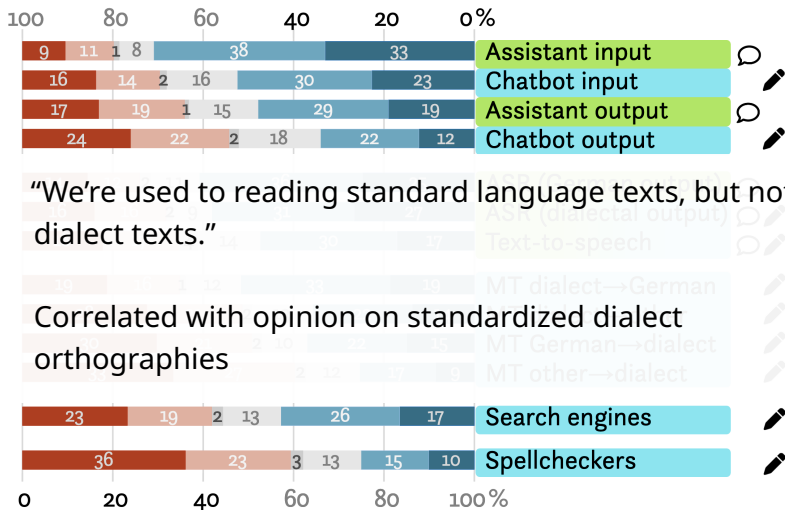
"Dialect is the language of the heart, not of a machine."



Spoken vs. written dialect?



Spoken vs. written dialect?



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
 - Linguistically more distant
 - 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

Blaschke, Purschke, Schütze & Plank (ACL 2024)

“What do dialect speakers want?”

- 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!

Personal takeaway: New projects with speech

Blaschke, Winkler, Förster, Wenger-Glemser & Plank (Interspeech 2025)

"A multi-dialectal dataset for German dialect ASR and dialect-to-standard speech translation"



🔊 D' Lomma hod a rechd a sauwas Wossa
✍️ Die Lammer hat ein recht sauberes Wasser

Phonetic + grammatical + lexical differences

→ Evaluation: what is a good dialect-to-standard transcription?

Personal takeaway: New projects with speech

Blaschke, Winkler & Plank (under review, ArXiv 2510.07890)

"Standard-to-dialect transfer trends differ across text and speech"

[ENG] Is it cold today?

[DEU] Ist es heute kalt?

→ Intent: weather query

[BAR] Is es heid koid?



Not just more realistic, but also brings useful insights (speech models generalize better to dialectal variation than text models do)!

Conclusion

- ! What: Dialects and language variation
- ! How: Dialect NLP – (Some) challenges and methods
 - Part of a longer line of work on robustness
 - Still not solved, still new methods!
 - Creative approaches welcome :)
- ! For whom: Speaker perspectives
 - Applied technologies & speaker communities
 - The next big topic? (not just in dialect NLP)

Ethical Considerations for Machine Translation of Indigenous Languages:

Giving a Voice to the Speakers

Manuel Mager^{♡*} Elisabeth Mager[‡]
Katharina Kann[♣] Ngoc Thang Vu[◇]

Not always about you: Prioritizing community needs when developing endangered language technology

Zoey Liu^{*} Crystal Richardson (Karuk)^{*}
Richard Hatcher Jr Emily Prud'hommeaux

Centering the Speech Community

Steven Bird

Dean Yibarbuk

Language Technologies as if People Mattered: Centering Communities in Language Technology Development

Nina Markl, Lauren Hall-Lew, Catherine Lai

What a Creole Wants, What a Creole Needs

Heather Lent¹, Kelechi Ogueji², Miryam de Lhoneux^{1,3,4}, Orevaghene Ahia⁵, Anders Sjøgaard¹

Conclusion

- ! What: Dialects and language variation
- ! How: Dialect NLP – (Some) challenges and methods
- ! For whom: Speaker perspectives

Papers & resources:
[dialect-erc.github.io](https://github.com/dialect-erc)
[verenablaschke.github.io](https://github.com/verenablaschke)
github.com/mainlp/germanic-lrl-corpora



Thank you to my collaborators!
Thanks for listening!