Dialect NLP

Thinking outside the box when processing non-standard and low-resource languages

Verena Blaschke
MaiNLP lab, LMU Munich

M.Sc. seminar *Human-Centric NLP*
15 May 2025

Dialect NLP

Help, I'm working on a non-standardized low-resource language!

Verena Blaschke
MaiNLP lab, LMU Munich

M.Sc. seminar *Human-Centric NLP*
15 May 2025

Natural language processing – which languages?

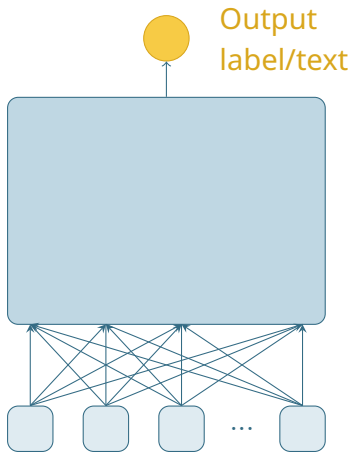
Who here speaks…

- a dialect or a regional language variety?
- a language with few or no NLP resources?

Today:

- Considerations and tips for processing data from dialects and other low-resource languages (LRLs)
- Specific examples (papers) + pointers to more literature

Overview – challenges & approaches



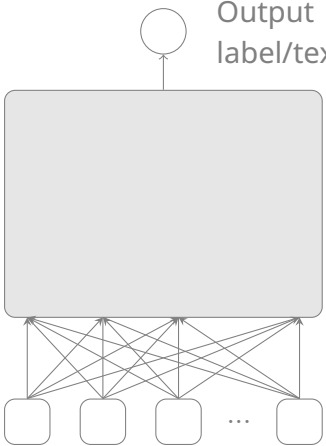
What tools and why?

Modelling non-standard data

Dialect/LRL data

- Dialects & language variation
- Data challenges

Overview – challenges & approaches



What tools and why?

Modelling non-standard data

Dialect/LRL data

- Dialects & language variation
- Data challenges

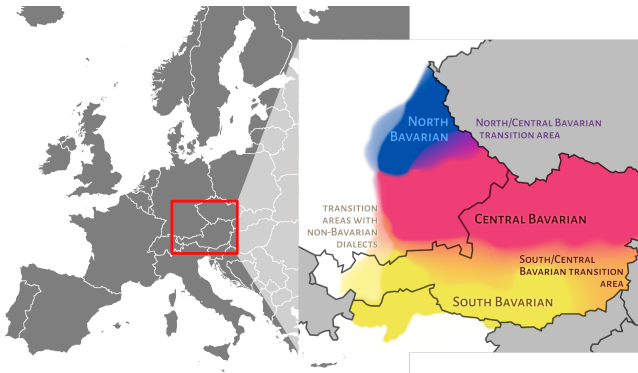
What do I mean with “dialects”?

Many definitions in linguistics, NLP & everyday language

- Any 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 ($\rightarrow$ spelling)

[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 ($\rightarrow$ spelling)
- Lexicon

[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 ($\rightarrow$ 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"

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*

“They [=fish] have no legs”

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"

When do people use dialects?

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

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 (& low-resource language) corpora

- Availability
- Quality
- Written representations

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

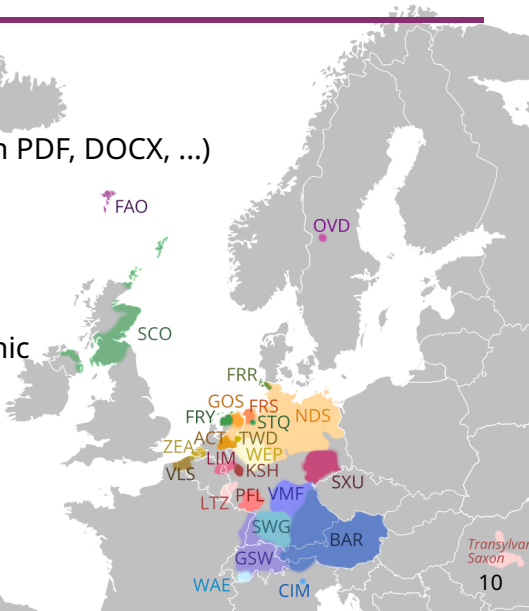
Verena Blaschke **Hinrich Schütze** **Barbara Plank**

Datasets for Germanic low-resource varieties

- Accessible for research
- Computer-friendly formats (XML, TSV, TXT, … rather than PDF, DOCX, …)
- Full sentences/utterances
- Annotated + unannotated
- High-quality data

➔ 100+ datasets for 35 Germanic dialects + small languages

github.com/mainlp/
germanic-lrl-corpora



How do you find corpora?

- Publications (ACL Anthology, arXiv, searching Google Scholar / Semantic Scholar) + a lot of manual digging…
- Google Dataset Search `datasetsearch.research.google.com`
- Data repositories
  - Zenodo `zenodo.org`
  - European Language Grid `live.european-language-grid.eu`
  - CLARIN Virtual Language Observatory `vlo.clarin.eu`
  - OpenSLR `openslr.org`
  - Text+ `text-plus.org`
  - OLAC `www.language-archives.org`
  - ORTOLANG `www.ortolang.fr/market/corpora`
  - Hamburg Centre for Language Corpora (HZSK)
  - OPUS `opus.nlpl.eu`

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: curated (elicited, transcribed, from books, manually checked web data, …), but not annotated
    - … and sometimes uncurated (e.g., webcrawls)

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)

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)

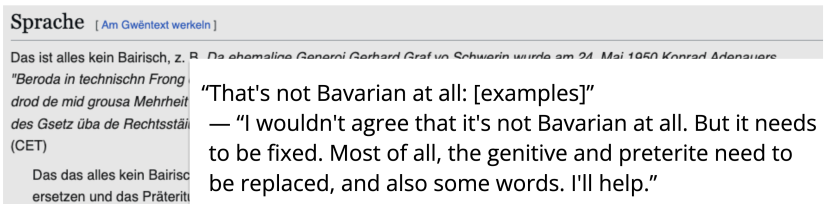
“West Flemish” QED OPUS corpus

```
<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



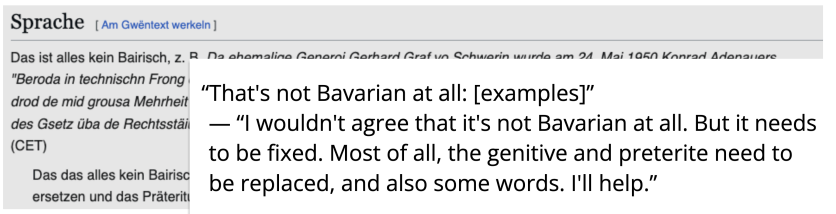
"That's not Bavarian at all: [examples]"
— "I wouldn't agree that it's not Bavarian at all. But it needs to be fixed. Most of all, the genitive and preterite need to be replaced, and also some words. I'll help."

Brooks/Hern, The Guardian, 2020

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



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

How to represent a primarily spoken language?

- Normalized text (closely related standard language)

Etter litt godsnakk kom tre av kyrne …	NB Tale
[After some coaxing, three of the cows came …]	*Norwegian*
können sie ihre jugendzeit beschreiben	ArchiMob
[Can you describe your youth?]	*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 … NB Tale
[After some coaxing, three of the cows came …] *Norwegian*

chönd sii iri jugendziit beschriibe
können sie ihre jugendzeit beschreiben ArchiMob
[Can you describe your youth?] *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

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

How to represent a (mostly spoken) language?

Speakers themselves can have different attitudes towards orthography!

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

Jrundsätzlich [der Quälltäx ändere]

Jeder schriev, wie em de Fingere jewaaße sin.

Data scarcity & overlaps

Bavarian Wikipedia

- In mBERT's pretraining data
- In Bavarian named entity dataset (BarNER)
- In Bavarian treebank (MaiBaam)

Swiss German part-of-speech corpora

- Overlap between NOAH dataset and UZH Universal Dependencies dataset

Recommendations

Blaschke, Schütze & Plank (NoDaLiDa 2023)

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

… for *using* dialect/LRL corpora

- Check the quality!
  - Any obvious issues? Language ID correct? Likely produced by actual speakers?
- Suitable written representation for your purposes?
- Overlaps between (pre-)training, dev, test data?
- Data outside traditional NLP venues
  - E.g., works by linguists

Recommendations

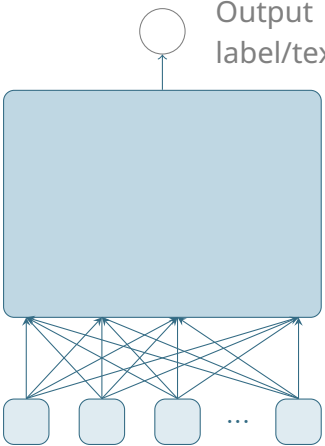
Blaschke, Schütze & Plank (NoDaLiDa 2023)

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

… for *creating* dialect/LRL corpora

- Document the transcription guidelines / orthographies
- Share metadata like corpus size, data sources, annotation procedure; specify a license or access conditions (!)
- Used archives geared towards long-term storage (CLARIN, LRE Map, Zenodo, …)

Overview



What tools and why?

Modelling non-standard data

Dialect data

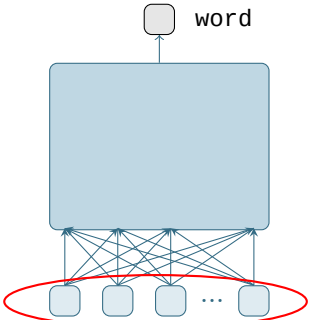
Linguistic differences

Differences from the standard language in

- **Pronunciation ($\rightarrow$ spelling)**
- Lexicon
- Morphology
- Syntax
- Usage context

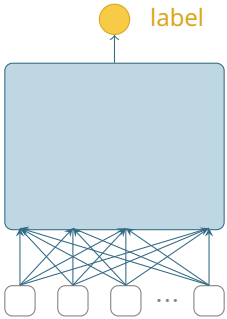
Cross-dialectal transfer

✗ Pretraining



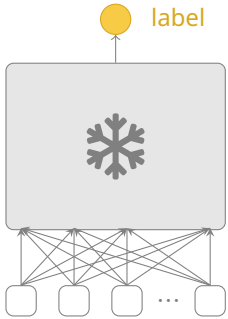
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum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

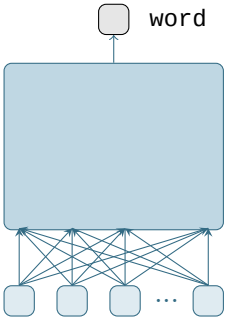
Die	Lammer	hat	ein	recht		sauberes	Wasser
Die	Lamm –er	hat	ein	recht		sauber –es	Wasser

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”

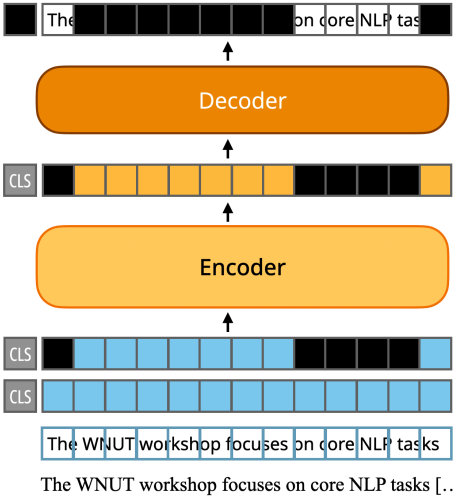
More robust input representations?



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,

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

Pixel models (Rust+, 2023) – pretraining



Decode masked pixels

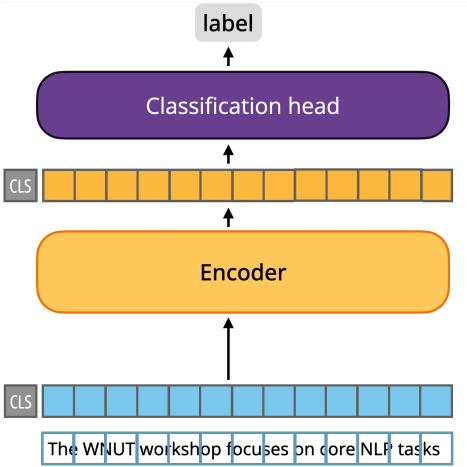
Encode

Mask spans

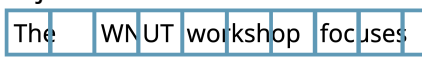
Projection

Render text as image

Pixel models (Rust+, 2023) – finetuning



Text rendering can be adjusted for word-level tasks



Pixel models – robustness

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

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

Table via *Rust+ “Language modelling with pixels” (ICLR 2023)*

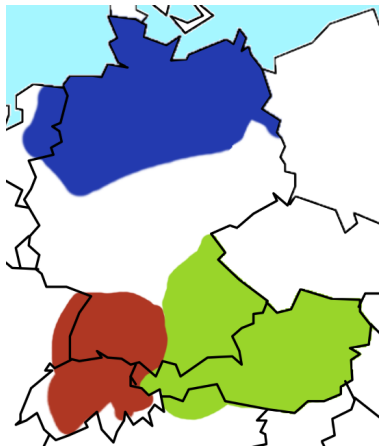
Pixel models – robustness

Evaluating Pixel Language Models on Non-Standardized Languages

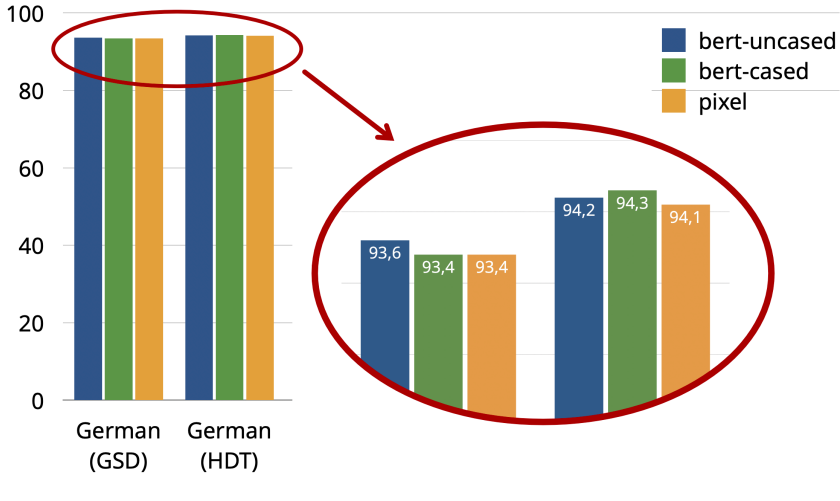
Alberto Muñoz-Ortiz **Verena Blaschke** **Barbara Plank**

German Pixel experiments

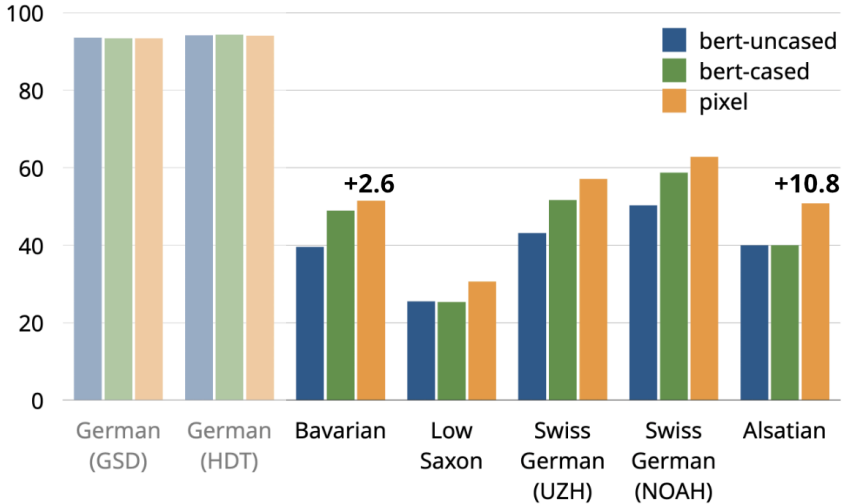
- German Pixel model (new!)
  - Same training data as a German BERT model
- Finetune on German, evaluate on dialects/regional languages
- 2 grammatical tasks: POS tagging, parsing
- 2 semantic tasks: intent classification (easy), topic classification (harder)



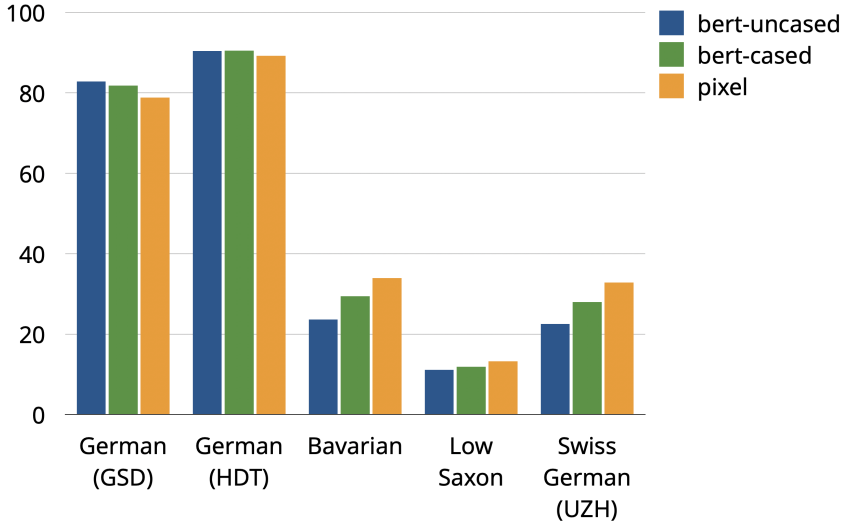
German Pixel: POS tagging (accuracy)



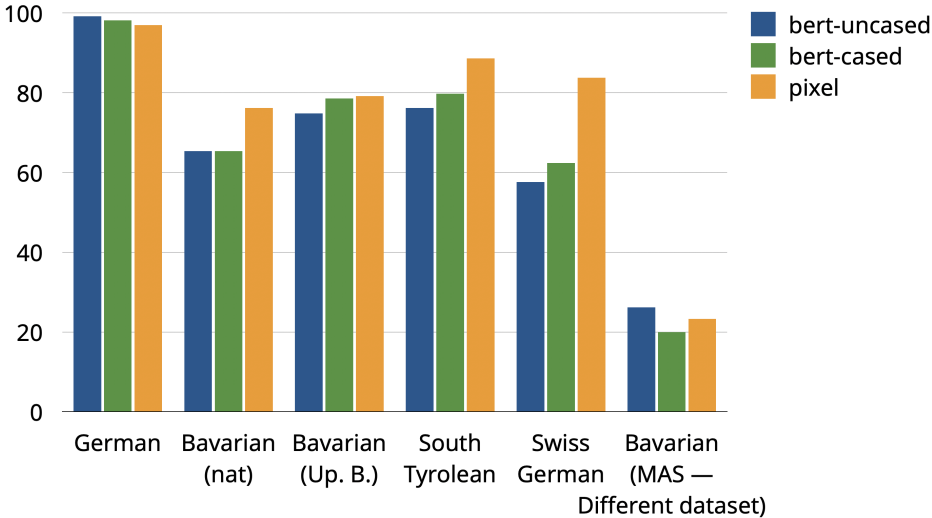
German Pixel: POS tagging (accuracy)



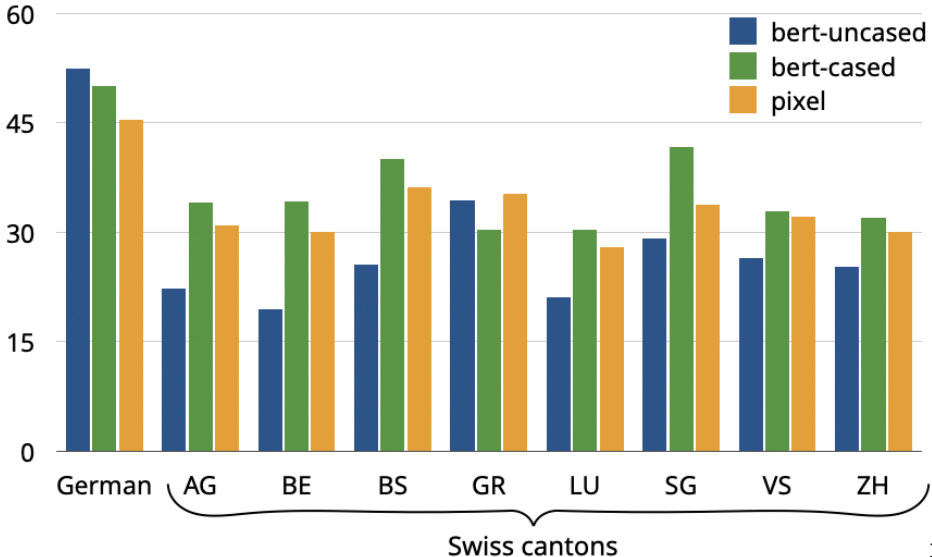
German Pixel: Parsing (LAS)



German Pixel: Intent classification (accuracy)



German Pixel: Topic classification (accuracy)



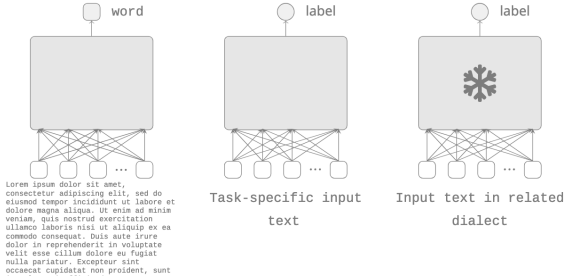
Pixel: Trade-off

Muñoz-Ortiz, Blaschke & Plank (COLING 2025)

"Evaluating pixel language models on non-standardized languages"

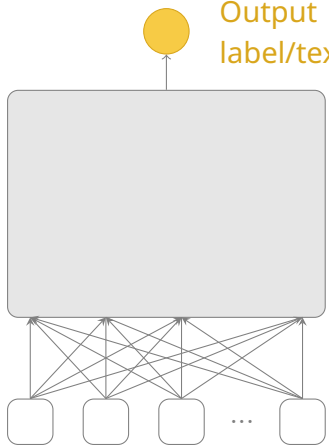
- More compute needed
- On par with or worse than BERT in monolingual settings (+ where std language performance is bad)
- Cross-dialectal settings / settings with less predictable spelling might be the place to shine
- Worthwhile for other "noisy" settings? (typos, ...)

Modelling non-standard data



- Different model architectures
- Changing the tokenizer
- Making the fine-tuning data more like the target data
- Normalizing the target data
- …

Overview



What tools and why?

Modelling non-standard data

Dialect data

Linguistic differences

Differences from the standard language in

- Pronunciation ($\rightarrow$ 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 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
- 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.

Questionnaire

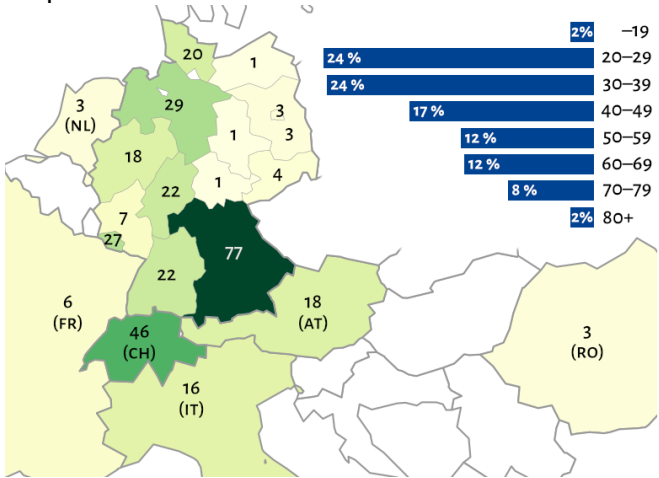
20. Stimmen Sie den folgenden Aussagen zu?

G310

Es sollte Transkriptionsprogramme geben, …	Ja, unbedingt	Eher ja	Weder noch	Eher nein	Nein, das halte ich nicht für sinnvoll	Das kann ich nicht bewerten
... die Audioaufnahmen in meinem Dialekt als geschriebenes Hochdeutsch wiedergeben.	○	○	○	○	○	○
... die Audioaufnahmen in meinem Dialekt als geschriebenen Dialekt wiedergeben.	○	○	○	○	○	○

Dialect background and attitudes

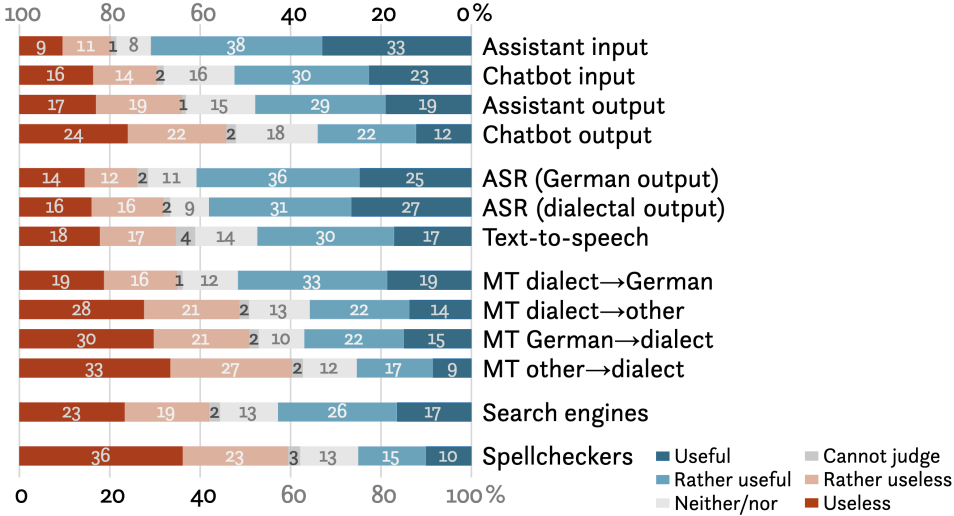
441 respondents – **327** of whom speak a German dialect and finished the questionnaire



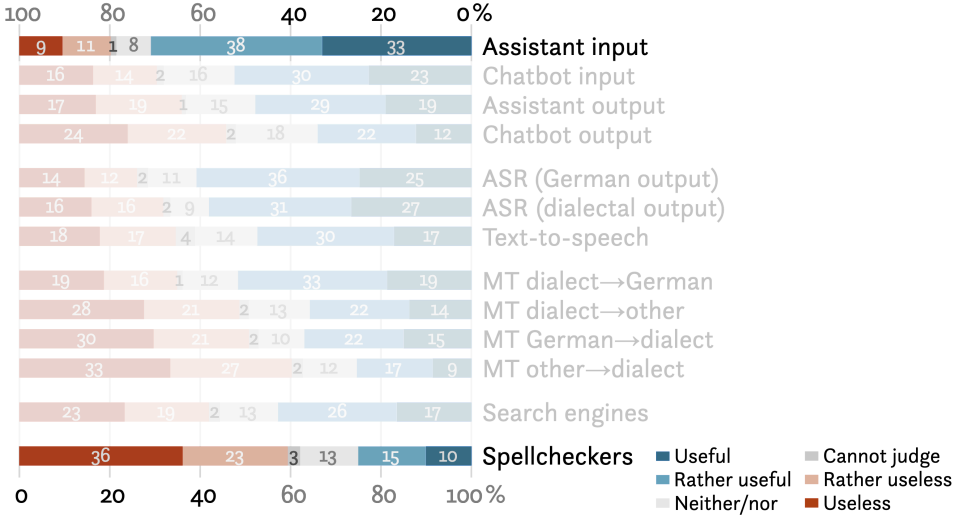
Dialect background and attitudes

- 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

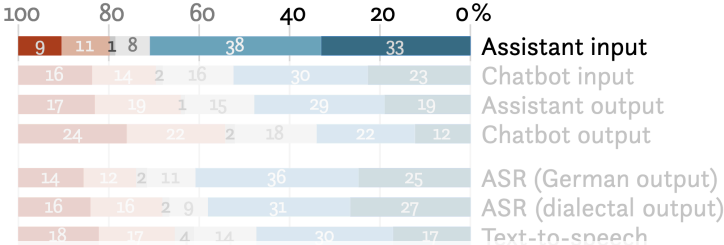
Which dialect LTs are deemed useful?



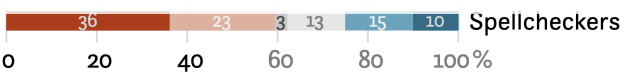
Which dialect LTs are deemed useful?



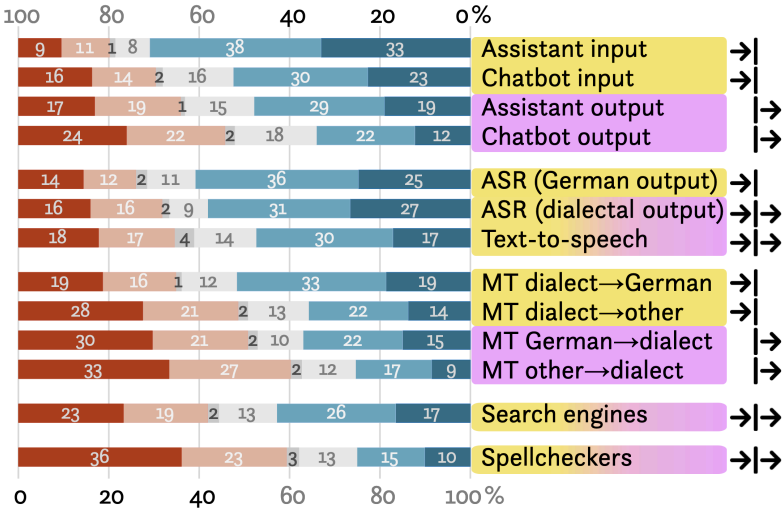
Which dialect LTs are deemed useful?



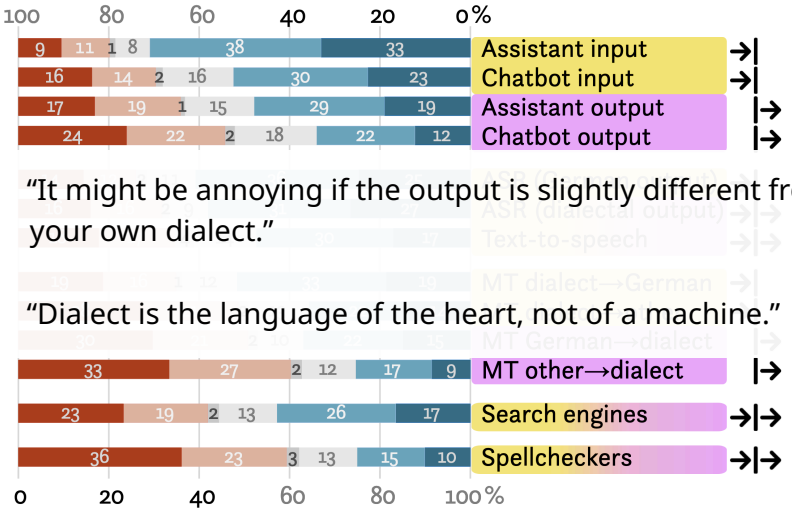
"The beauty of dialects is that there are no spelling/grammar rules and everyone can write in their own dialect, which is important since the exact version of one's dialect can be extremely local."



Dialect input vs. output?



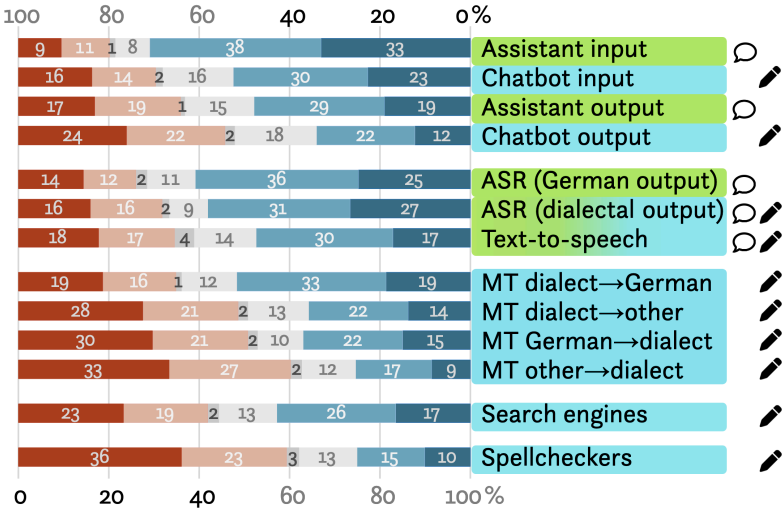
Dialect input vs. output?



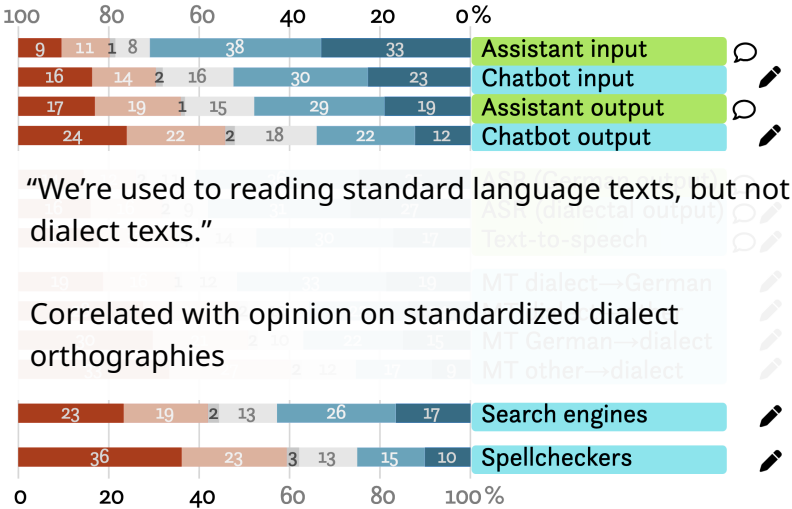
"It might be annoying if the output is slightly different from your own dialect."

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

Spoken vs. written dialect?



Spoken vs. written dialect?



"We're used to reading standard language texts, but not dialect texts."

Correlated with opinion on standardized dialect orthographies

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

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

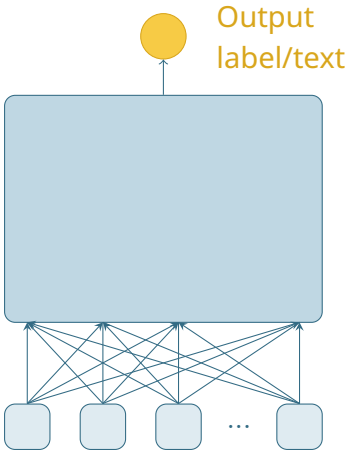
Takeaways

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!

Summary – challenges & approaches



Which tools and why?

- Consider the speaker perspectives

Modelling non-std data

- Be creative!

Data availability & quality

- Which language varieties are currently included in research & language tech?
- How trustworthy and generalizable are your data?

Further reading – dialect NLP

- Natural language processing for similar languages, varieties, and dialects: A survey (Zampieri+, Natural Language Engineering 2020)
- Quantifying the Dialect Gap and its Correlates Across Languages (Kantharuban+, EMNLP Findings 2023)
- DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages (Faisal+, ACL 2024)

Other types of variation, e.g., syntactic variation:

- Multi-VALUE: A Framework for Cross-Dialectal English NLP (Ziems+, ACL 2023)

Other kinds of methods, e.g., modifying the standard-language fine-tuning data:

- Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise (Aepli/Sennrich, ACL Findings 2022)

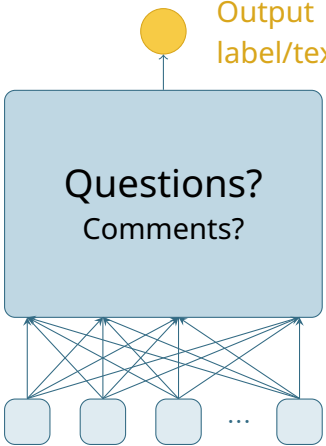
Further reading – low-resource language NLP

- NLP systems for low resource languages – hype vs. reality (Panel discussion, PML4DC @ ICLR 2023)
- Language Varieties of Italy: Technology Challenges and Opportunities (Ramponi, TACL 2024)
- Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets (Kreutzer+, TACL 2022)
- Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models (Ahia+, EMNLP 2023)
- A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios (Hedderich+, NAACL 2021)
- The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi+, ACL 2020)
- Evaluating the Diversity, Equity, and Inclusion of NLP Technology: A Case Study for Indian Languages (Khanuja+, EACL Findings 2023)

Further reading – LTs & speaker communities

- Not always about you: Prioritizing community needs when developing endangered language technology (Liu+, ACL 2022)
- What a Creole Wants, What a Creole Needs (Lent+, LREC 2022)
- Local Languages, Third Spaces, and other High-Resource Scenarios (Bird, ACL 2022)
- Ethical Considerations for Machine Translation of Indigenous Languages: Giving a Voice to the Speakers (Mager+, ACL 2023)
- Language Technologies as If People Mattered: Centering Communities in Language Technology Development (Markl+, LREC-COLING 2024)
- My LLM might Mimic AAE [African American English] – But When Should It? (Sandoval+, NAACL 2025)

Questions/discussion



Which tools and why?

Representing/modelling non-standard data

Data availability & quality

Discussion

1. [Your question/comment here :)]
2. Should we *only* do NLP research for technologies that speakers immediately deem useful?
3. If you speak a LRL: which NLP advancements should be a priority for your language?
4. Problem solved if everybody just becomes fluent in a standard language / high-resource language / English?