Dialect NLP

For whom and with what data?

Verena Blaschke
MaiNLP, LMU Munich

M.Sc. seminar *Human-centric NLP*
May 2, 2024

Natural language processing

Natural language processing – which languages?

Natural language processing – which languages?

Who here speaks...

- a dialect or a regional language variety?

Natural language processing – which languages?

Who here speaks…

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

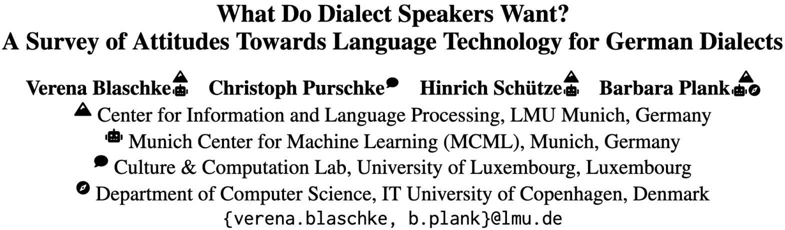
Structure

Dialect NLP

- For whom – attitudes of speakers
- With what data – LRL dataset overview/analysis
- Discussion

Why dialect NLP?

- Quantitative studies of linguistic variation
- How can we learn from sparse, heterogeneous data?
- Access to language technology



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

Verena Blaschke▲🤖 Christoph Purschke💬 Hinrich Schütze▲🤖 Barbara Plank▲🤖🧭
▲ Center for Information and Language Processing, LMU Munich, Germany
🤖 Munich Center for Machine Learning (MCML), Munich, Germany
💬 Culture & Computation Lab, University of Luxembourg, Luxembourg
🧭 Department of Computer Science, IT University of Copenhagen, Denmark
{verena.blaschke, b.plank}@lmu.de

Preprint/under review

Motivation

Language technology (LT) – applied NLP systems

- Machine translation
- Chatbots
- Virtual assistants
- Transcription (ASR/STT)
- Speech synthesis (TTS)
- Spellcheckers
- Search engines
- ...

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 and related 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

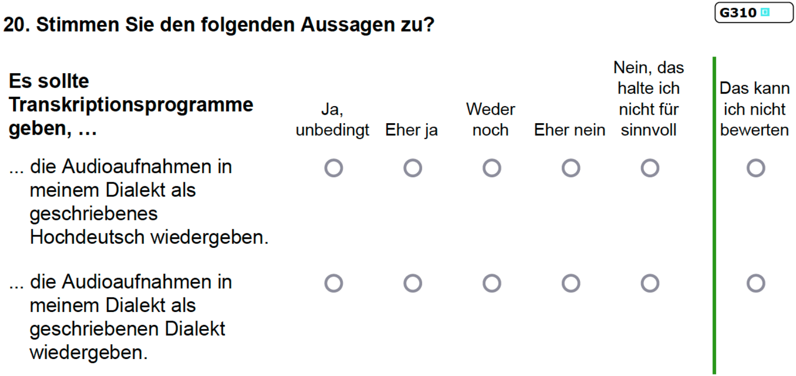
Questionnaire

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



G310 □

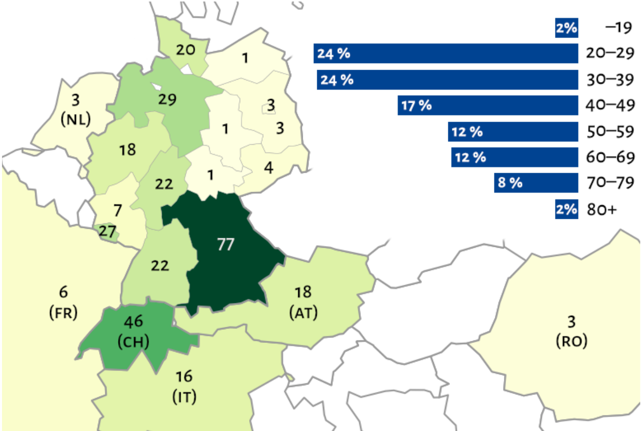
20. Stimmen Sie den folgenden Aussagen zu?

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

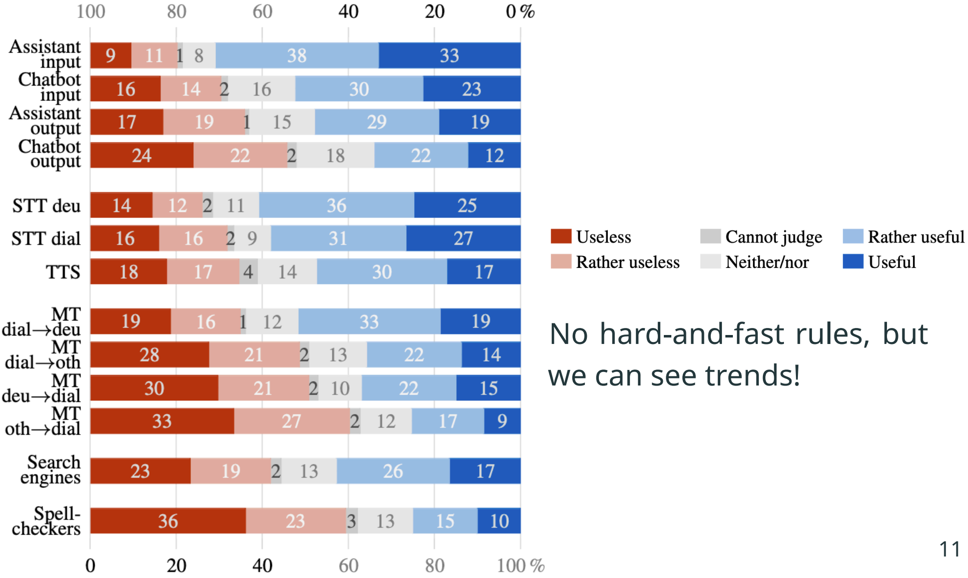
- Most speak the dialect fluently and frequently



Dialect background and attitudes

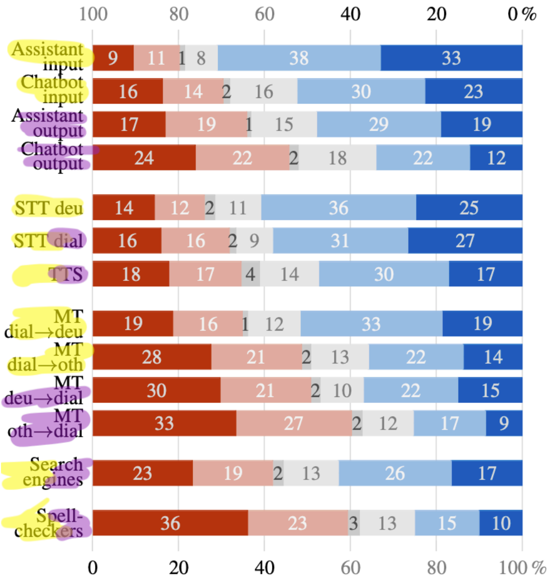
- 65 % against standardized orthography
- 66 % write their dialect (even if rarely)
- 35 % are actively involved in dialect preservation
    - dialect preservation societies (13 %), teachers, dialectologists, ...
    - speaking the dialect in public, with children
- 14 % already familiar with an LT for their dialect

Which dialect LTs are deemed useful?

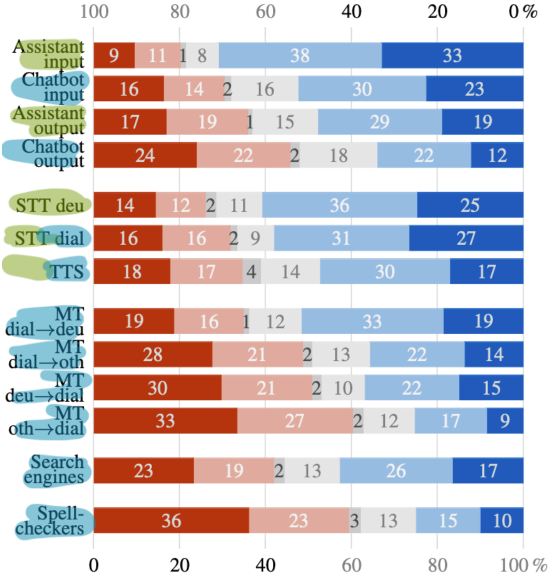


No hard-and-fast rules, but we can see trends!

Dialect input vs. output?



Text vs. speech?



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
- Nevertheless: removing the activists’ responses has very little impact on the order of preferred LTs

Do attitudes reflect sociolinguistic factors?

Three large dialect/language regions with different sociolinguistic realities

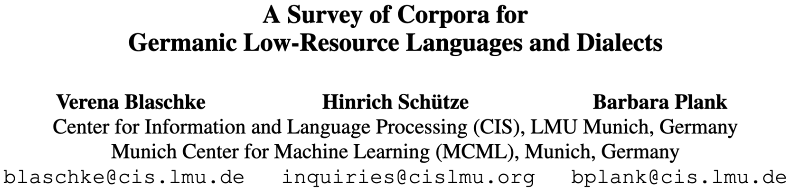
- Low Saxon (Plattdeutsch) – linguistically more distant from DEU, officially recognized language, speaker numbers in decline
- Swiss German – high prestige, spoken dialect + written German
- Non-Low-Saxon Germany + Austria – dialects partially replaced by regiolects

Do attitudes reflect sociolinguistic factors?

- Low Saxon: larger interest in orthographies + spellcheckers; Swiss: little interest
- Low Saxon: generally very in favour of dialect LTs, including text-based or with dialectal output
- Swiss: especially interested in LTs with spoken dialect input

Takeaways

- Generally: 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!



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

Verena Blaschke **Hinrich Schütze** **Barbara Plank**
Center for Information and Language Processing (CIS), LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
blaschke@cis.lmu.de inquiries@cislmu.org bplank@cis.lmu.de

Published at NoDaLiDa 2024

Challenges concerning LRL corpora

- Availability
- Quality
- How do you represent data from unwritten/newly written languages?

Which relevant datasets are out there?

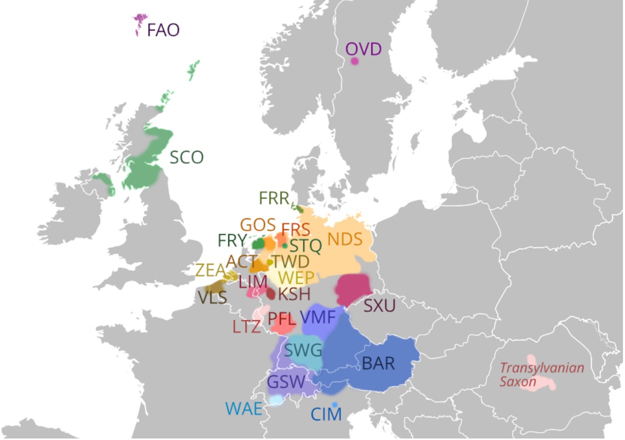
- Recent (rather than historic) data
- Full sentences/utterances (no word lists)
- (In)directly accessible for research (!!)
- Computer-friendly formats
- High-quality data (look out for, e.g., OCR issues!)

→100+ corpora!
github.com/mainlp/germanic-lrl-corpora

For which language varieties did we find datasets?

(+spoken primarily outside Europe)

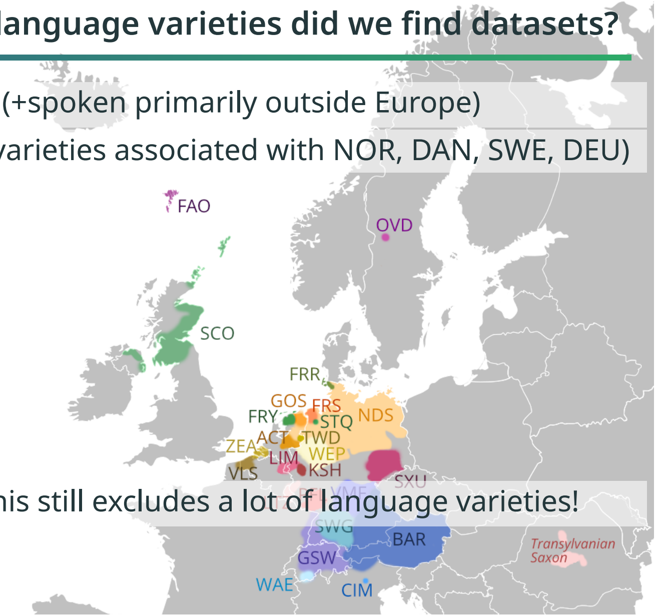
(+non-std varieties associated with NOR, DAN, SWE, DEU)



For which language varieties did we find datasets?

(+spoken primarily outside Europe)

(+non-std varieties associated with NOR, DAN, SWE, DEU)



...but this still excludes a lot of language varieties!

Annotations

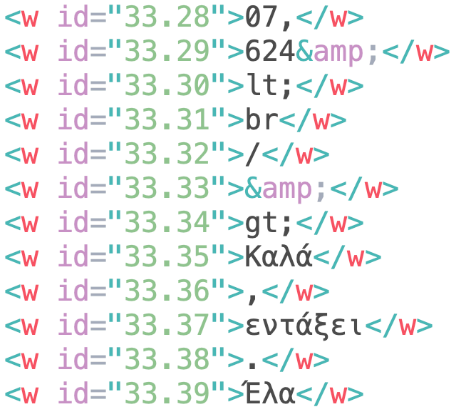
What, if any, high-quality annotations do we find?

- Mostly: not annotated
- Morphosyntax (POS tags, dependencies, phrase structure)
- Geolocation, dialect group
- Rare: paraphrases, translations, sentiment, topics, slot and intent detection

Data quality: uncurated data

Uncurated LRL data tend to be rather low quality (wrong language, bad data cleaning)

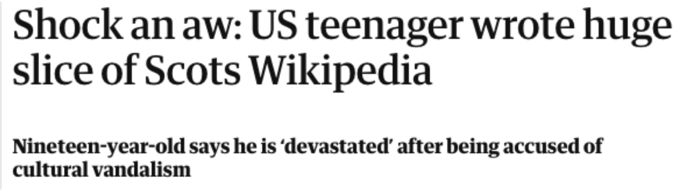
“West Flemish” QED OPUS corpus



```
<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>
<w id="33.37">εντάξει</w>
<w id="33.38">.</w>
<w id="33.39">Έλα</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

https://www.theguardian.com/uk-news/2020/aug/26/shock-an-aw-us-teenager-wrote-huge-slice-of-scots-wikipedia

How to represent a (mostly spoken) language?

- Privacy issues concerning audio material
- Wildly different transcription styles & orthographies

How to represent a (mostly spoken) language?

- Normalized text (closely related std language)

Etter litt godsnakk kom tre av kyrne ...	NB Tale
können sie ihre jugendzeit beschreiben	ArchiMob

How to represent a (mostly spoken) language?

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

Etter litt godsnakk kom tre av kyrne ... NB Tale
”“”{t@4 l”it g”“”u:snAkk k”Om t4”e: ”A:v C”y:n‘@ ...

können sie ihre jugendzeit beschreiben ArchiMob
chönd sii iri jugendziit beschriibe

How to represent a (mostly spoken) language?

- Normalized text (closely related std language)
- Phone[m/t]ic transcriptions
- (More or less accepted) LRL orthographies

Etter litt godsnakk kom tre av kyrne ...	NB Tale
””{t@4 l”it g””u:snAkk k”Om t4”e: ”A:v C”y:n‘@ ...	
können sie ihre jugendzeit beschreiben	ArchiMob
chönd sii iri jugendziit beschriibe	
Nu leyt em de böyse vynd disse nacht ...	UD LSDC

How to represent a (mostly spoken) language?

- Normalized text (closely related std language)
- Phone[m/t]ic transcriptions
- (More or less accepted) LRL orthographies
- Ad-hoc spellings

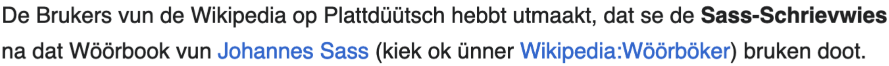
Etter litt godsnakk kom tre av kyrne ... NB Tale
”“”{t@4 l”it g”“”u:snAkk k”Om t4”e: ”A:v C”y:n‘@ ...

können sie ihre jugendzeit beschreiben ArchiMob
chönd sii iri jugendziit beschriibe

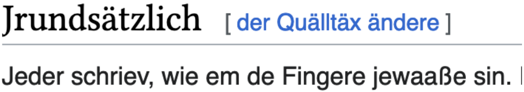
Nu leyt em de böyse vynd disse nacht ... UD LSDC
Nu leit em de baise Find düse Nacht ...

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-Schrievwies** 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.

Conclusion

Recommendations for *using* LRL corpora

- Check the quality!
- Is the written representation suitable for your purposes?
- Check whether your (pre-)training, dev, and test data are truly from independent sources (datasets overlap!)
- Also look for quantitative works by dialectologists and sociolinguists (outside traditional NLP venues)

Conclusion

Recommendations for *creating* LRL corpora

- Document the transcription guidelines / orthographies (if applicable)
- Use archives geared towards long-term storage (CLARIN, LRE Map, Zenodo)
- Share basic metadata like corpus size, data sources, annotation procedure
- Add license information

Further reading

- Not always about you: Prioritizing community needs when developing endangered language technology (Liu+, ACL 2022)
- Language Varieties of Italy: Technology Challenges and Opportunities (Ramponi, TACL 2024)
- What a Creole Wants, What a Creole Needs (Lent+, LREC 2022)
- Local Languages, Third Spaces, and other High-Resource Scenarios (Bird, ACL 2022)
- NLP systems for low resource languages – hype vs reality (Panel discussion, PML4DC @ ICLR 2023)
- Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets (Kreutzer+, TACL 2022)
- Evaluating the Diversity, Equity, and Inclusion of NLP Technology: A Case Study for Indian Languages (Khanuja+, EACL 2023 Findings)

Discussion, questions?

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 English?