Dialect NLP

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

Verena Blaschke
MaiNLP lab, LMU Munich

Talk @ Linguistics Graduate Colloquium,
University of Passau
December 10, 2024

Natural Language Processing

… but *which* languages?

NLP – but which “language(s)”?

- Many speakers, abundant data, standardization

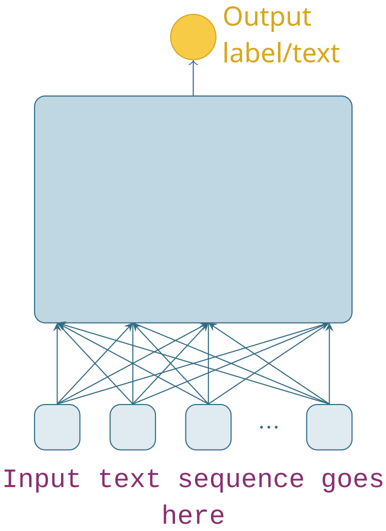
But does everyone use language this way?

- Also include minority languages, non-standard varieties
- Tricky for NLP!
  Modern methods learn from massive amounts of data

Why dialect NLP?

- Linguistics: research language variation, annotate data
- ML research: sparse and heterogeneous data
- Applied uses: automatic subtitling, voice-based assistants in cars, analyzing social media data, …

Overview – challenges & approaches

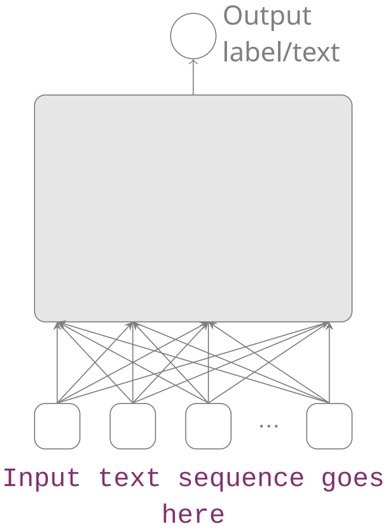


Human-centric NLP
(what tools and why?)

Modelling non-standard data

Available dialect data

Overview – challenges & approaches



Output label/text

…

Input text sequence goes here

Human-centric NLP (what tools and why?)

Modelling non-standard data

Available dialect data

Which relevant datasets are out there?

- Germanic dialects and low-resource languages
- (In)directly accessible for research (!!)
- Computer-friendly formats

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

“A survey of corpora for Germanic low-resource languages and dialects”
Blaschke, Schütze & Plank (NoDaLiDa 2023)

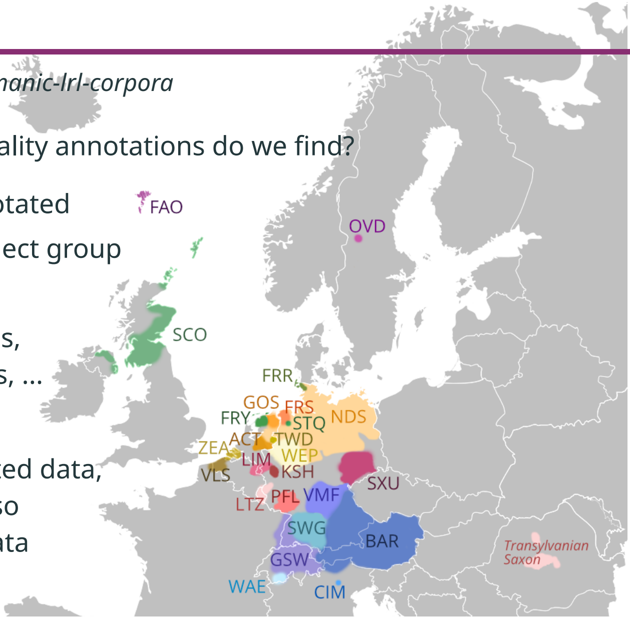
Annotations

github.com/mainlp/germanic-lrl-corpora

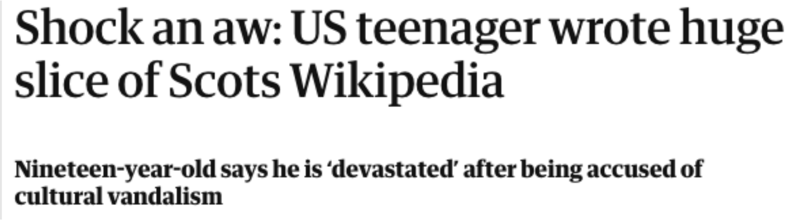
What, if any, high-quality annotations do we find?

- Mostly: not annotated
- Geolocation, dialect group
- Morphosyntax
- Rare: translations, sentiment, topics, ...

Many based on curated data, but more recently also on uncurated web data



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

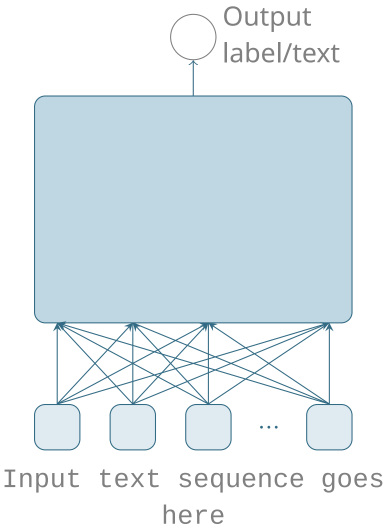
theguardian.com/uk-news/2020/aug/26/shock-an-aw-us-teenager-wrote-huge-slice-of-scots-wikipedia

Corpus overview: Conclusions

github.com/mainlp/germanic-lrl-corpora

- Two communities: variationists & NLP researchers – data exchange :)
- Findable; licenses allowing re-use
- Long-term storage + accessibility

Overview

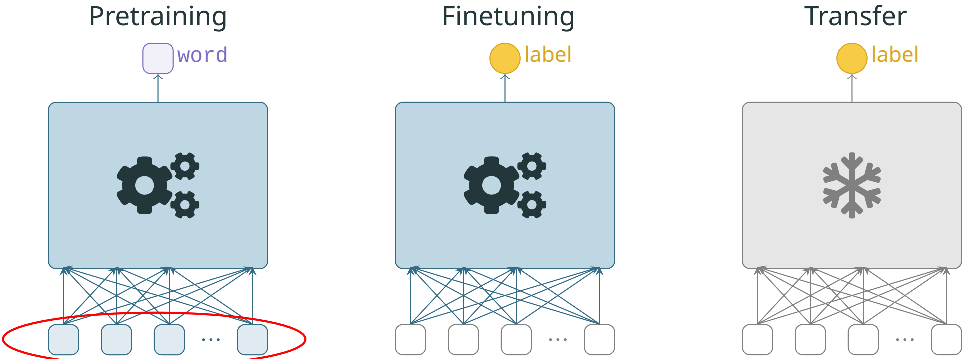


Human-centric NLP
(what tools and why?)

Modelling non-standard data

Available dialect data

Language models: Pretrain – finetune – transfer



Pretraining

word

Finetuning

label

Transfer

label

...

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, sed do eiusmod tempor incididunt ut

Task-specific input text

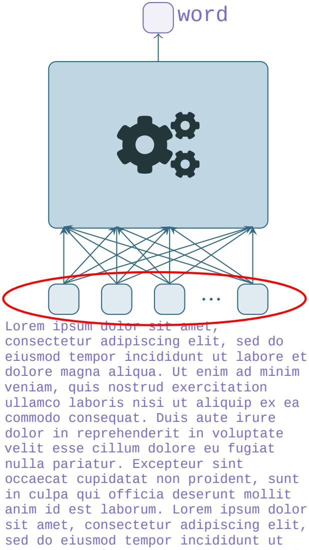
Input text in another (closely related) language

e.g., German

e.g., Bavarian

Language models: Pretrain – finetune – 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, sed do eiusmod tempor incididunt ut

Encoding input text

Map frequent character sequences – “subword tokens” – to numeric representations

Non-standard orthographies + tokenization

Subword tokenization with GBERT

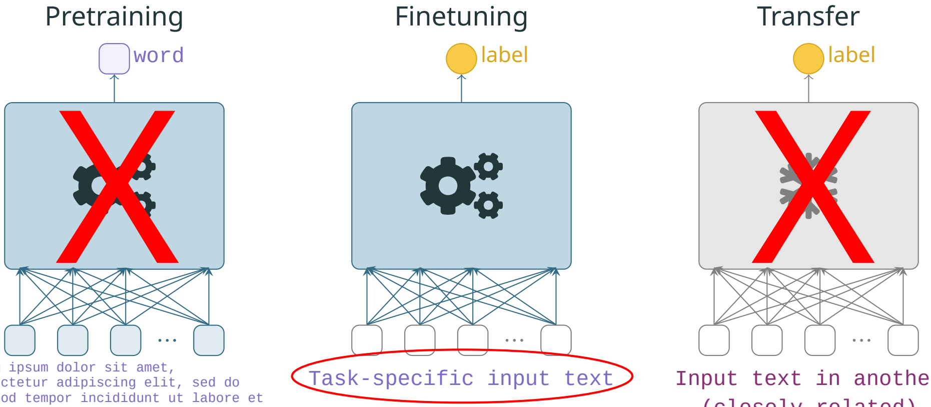
Die	Lammer	hat	ein	recht		sauberes	Wasser
Die	Lamm -er	hat	ein	recht		sauber -es	Wasser
D’	Lomma	hod	a	rechd	a	sauwas	Wossa
D '	Lom -ma	ho -d	a	rech -d	a	sau -was	Wo -ssa

“The Lammer (river) has fairly clean water”

ChatGPT & Co also rely on such tokenization

Sentence via bar.wikipedia.org/wiki/Låmma
GBERT: Chan/Schweter/Möller, COLING 2020, “German’s Next Language Model”

How to make models more robust?



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, sed do eiusmod tempor incididunt ut

Task-specific input text

Input text in another (closely related) language

e.g., German

e.g., Bavarian

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

Inject 15% of words with “noise”

Aepli/Sennrich, ACL Findings 2022 “Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise”

Character-level “noise”

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

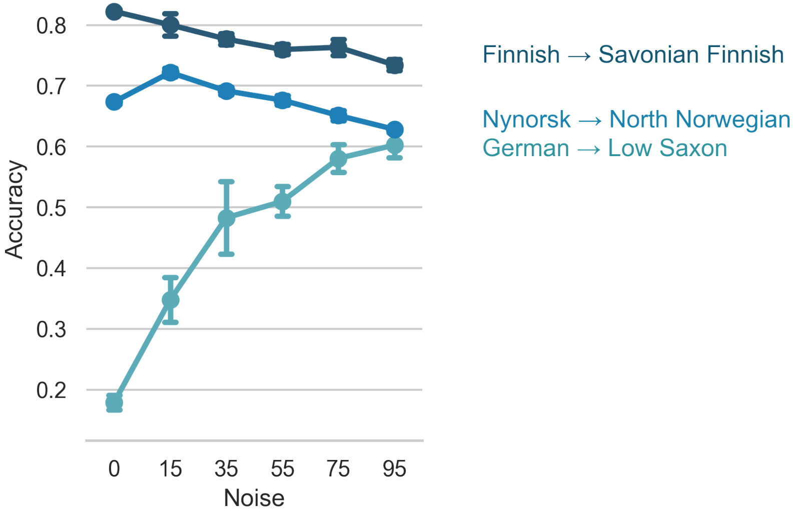
Inject 15% of words with “noise” (Aepli/Sennrich 2022)

How many words should we modify this way?

“Does manipulating tokenization aid cross-lingual transfer?
A study on POS tagging for non-standardized languages”
Blaschke, Schütze & Plank (VarDial 2024)

Part-of-speech tagging for dialects/regional languages related to German, Norwegian, French, Finnish, Arabic

How much noise to add?



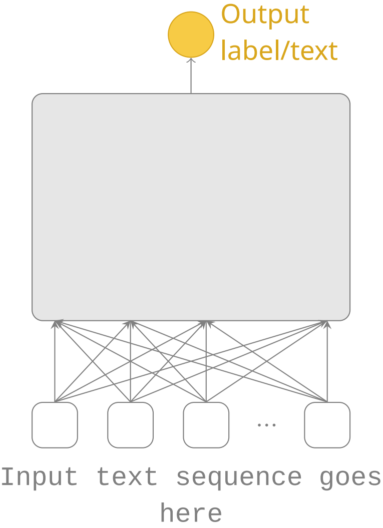
What explains this?

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

More details / langauge models / anlyses in the paper!

Overview



Human-centric NLP
(what tools and why?)

Modelling non-standard data

Available dialect data

Why dialect NLP?

Why, given the fact that these speakers also speak the standard language?

- Linguistics: research language variation
- ML research: sparse and heterogeneous data
- Applied reasons
  - Industry perspective (automatic subtitling, voice-based assistants in cars, analyzing social media data, ...)
  - Speaker perspective

"What do dialect speakers want? A survey of attitudes towards language technology for German dialects"
Blaschke, Purschke, Schütze & Plank (ACL 2024)

Motivation

Language technology (LT) – applied NLP systems

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

Research questions

1. Which dialect technologies do respondents find especially useful?
2. Does this depend on...
    - dialectal input vs. output?
    - speech- vs. text-based technologies?
3. How does this reflect relevant sociolinguistic factors?

Target audience: speakers of German dialects + regional languages
Questions

- Part I: about their dialect
- Part II: about attitudes towards LTs for their dialect

Questionnaire

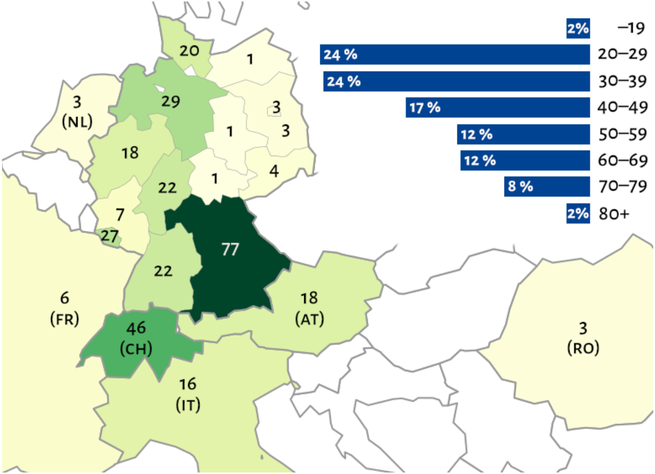
Speech-to-text systems transcribe spoken language. They are for instance used for automatically generating subtitles or in the context of dictation software.

Do you agree with the following statements?
There should be speech-to-text software...

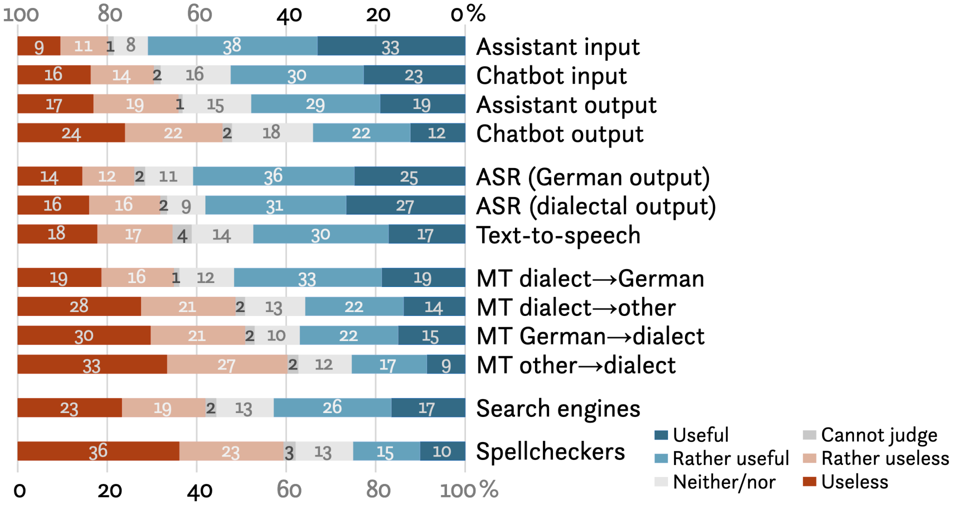
- ...that transcribes audio recorded in my dialect as written Standard German.
- ...that transcribes audio recorded in my dialect as written dialect.

Dialect background and attitudes

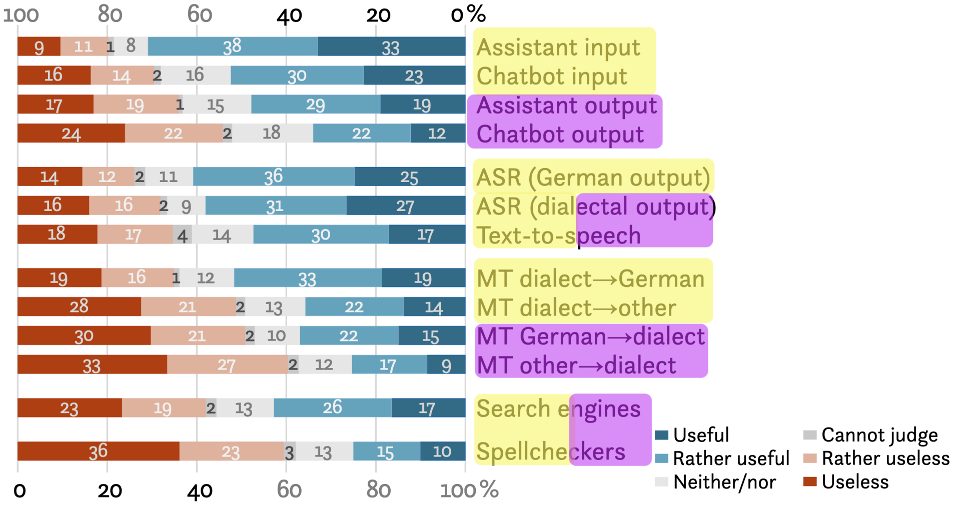
441 respondents – **327** of whom speak a German dialect and finished the questionnaire



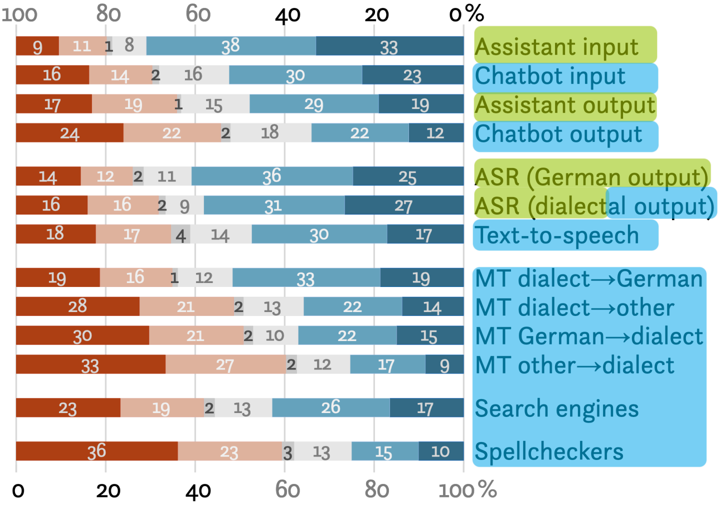
Which dialect LTs are deemed useful?



Dialect input vs. output?

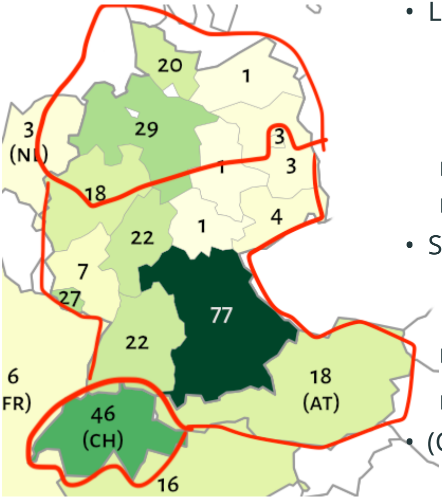


Spoken vs. written dialect?



Correlated with opinion on standardized dialect orthographies

Do attitudes reflect sociolinguistic factors?



- Low Saxon
  - Linguistically more distant
  - Recognized as language
  - Preservation efforts
  - 👍 Dialect LTs in general
  - 👍 Orthographies + spellcheckers
- Swiss German
  - High prestige
  - Spoken dialect, written Std German
  - 👎 Orthographies + spellcheckers
  - 👍 Spoken dialectal input
- (Central/Southern) Germany + Austria
  - Partially replaced by regiolects

Do attitudes reflect sociolinguistic factors?

Speaker(group)s aren’t monoliths!

Sociolinguistic background is
an important factor
(but individual opinions exist too)

- Low Saxon
  - Linguistically more distant
  - Recognized as language
  - Preservation efforts
  - 👍 Dialect LTs in general
  - 👍 Orthographies + spellcheckers
- Swiss German
  - High prestige
  - Spoken dialect, written Std German
  - 👎 Orthographies + spellcheckers
  - 👍 Spoken dialectal input
- (Central/Southern) Germany + Austria
  - Partially replaced by regiolects

Summary – challenges & approaches

Output label/text

Thank you!

Any questions?

…

Input text sequence goes here

Reflecting on what tools we build

Representing/modelling non-standard data

Data availability

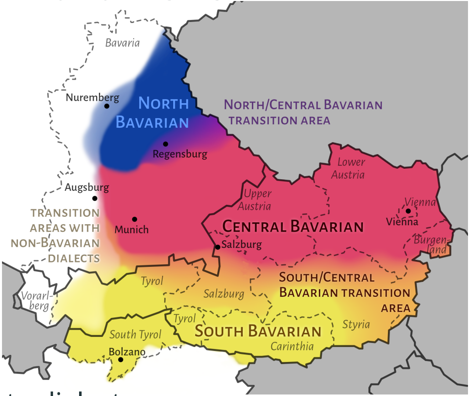
→ *github.com/mainlp/ germanic-lrl-corpora*

Appendix

Back-up slides

What do I mean with “dialects”?

Many definitions in linguistics, NLP & everyday language



Here:

- Non-standardized
- Closely related to a standard language
- Differences from the std language in
    - Pronunciation (spelling)
    - Lexicon
    - Morphology and syntax
- Often: continuum from standard to dialect
- Di(a)glossia; dialect speakers typically also write (and speak?) the standard

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

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

→ If you build a tool that works for one type of written representation, it doesn’t necessarily work for the others too

Survey: Dialect background and attitudes

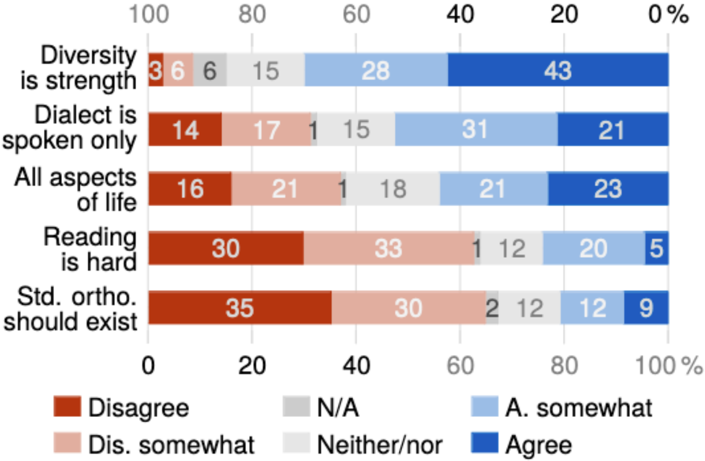
- 52 % speak their dialect daily
- 65 % against standardized orthography
- 66 % write their dialect (even if rarely)
- 35 % are actively involved in dialect preservation*
  - dialect preservation societies (13 %), teachers, dialectologists, ...
  - speaking the dialect in public, with children
- 14 % already familiar with an LT for their dialect

*“Language activists”

- 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

Survey: Dialect attitudes


Diversity is strength
Dialect is spoken only
All aspects of life
Reading is hard
Std. ortho. should exist
Disagree
N/A
A. somewhat
Dis. somewhat
Neither/nor
Agree