Large language models and small language varieties

Challenges and current methods

Verena Blaschke & Barbara Plank
MaiNLP lab, LMU Munich
mainlp.github.io

Embracing variability in natural language processing
ICLaVE|12
July 10, 2024

Natural Language Processing

... but *which* languages?

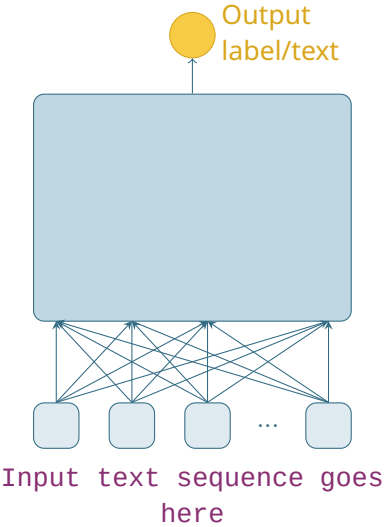
NLP – but which “language(s)”?

- Many speakers, abundant data, standardization

But how do we actually use language?

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

Overview – challenges & approaches

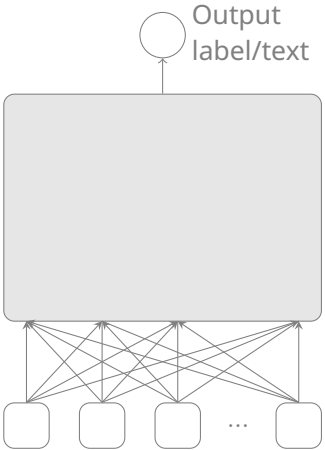


Human-centric NLP
(what tools and why?)

Modelling non-standard data

Available dialect data

Overview – challenges & approaches



Input text sequence goes here

Human-centric NLP
(what tools and why?)

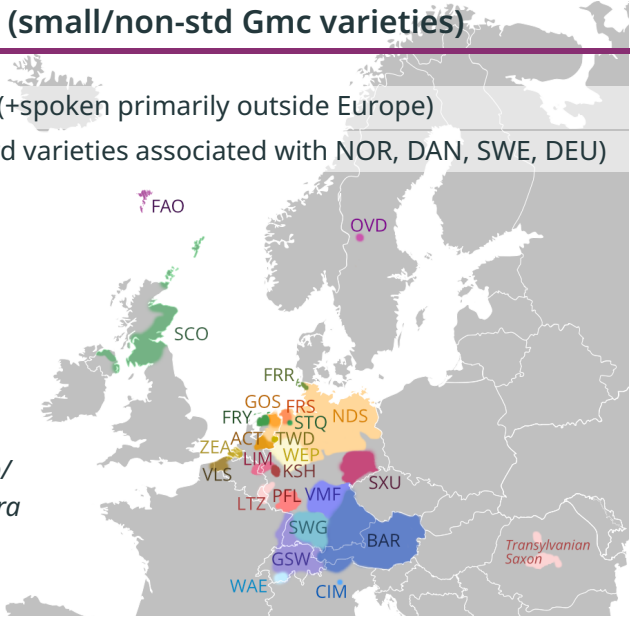
Modelling non-standard data

Available dialect data

Corpus overview (small/non-std Gmc varieties)

(+spoken primarily outside Europe)

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



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

Corpus overview (small/non-std Gmc varieties)

github.com/mainlp/germanic-lrl-corpora
100+ (mostly written) corpora for ∼30 language varieties

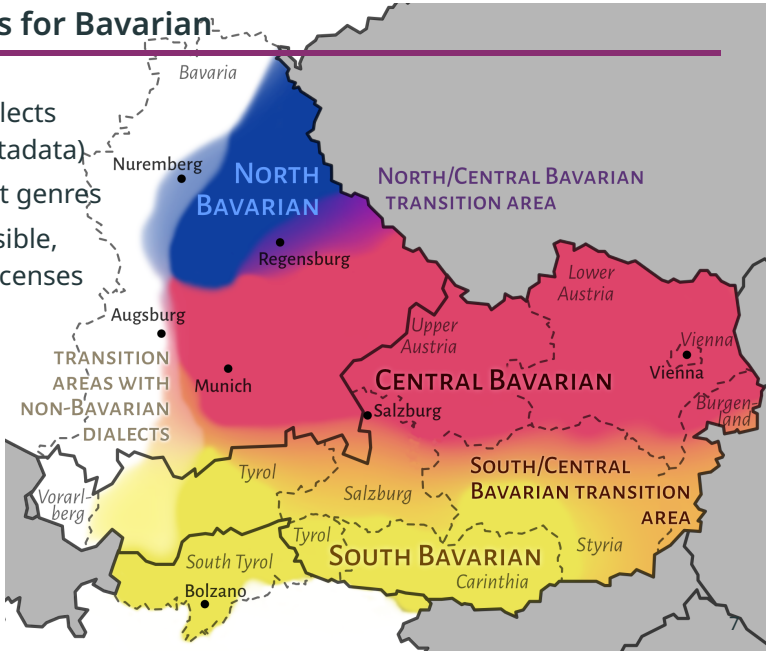
- Largely unannotated
- If annotated:
  - Geolocation, dialect group
  - Morphosyntax
  - Rarely: translations, content-related annotations

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

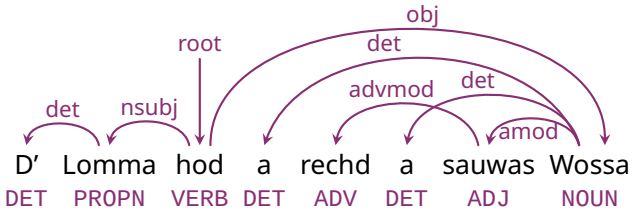
Blaschke/Schütze/Plank, NoDaLiDa 2023 "A survey of corpora for Germanic low-resource languages and dialects"

NLP resources for Bavarian

- Different dialects (location metadata)
- Different text genres
- Freely accessible, permissive licenses



NLP resources for Bavarian

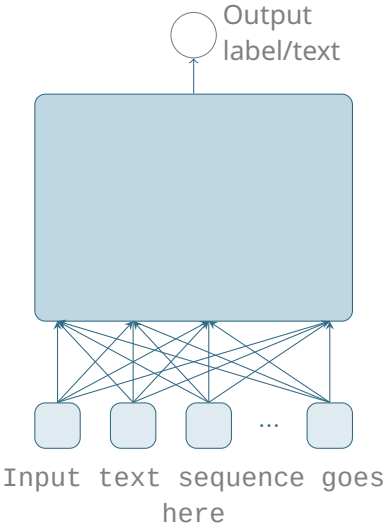


Da Rudoif hod 1365 de Universität Wean grindt
person *organization*

Wia hoass wearts heint?
weather attribute *datetime* → intent: *find weather*

MaiBaam (Blaschke+ 2024),
BarNER (Peng+ 2024),
xSID (van der Goot+ 2021, Aepli+ 2023, Winkler+ 2024)

Overview

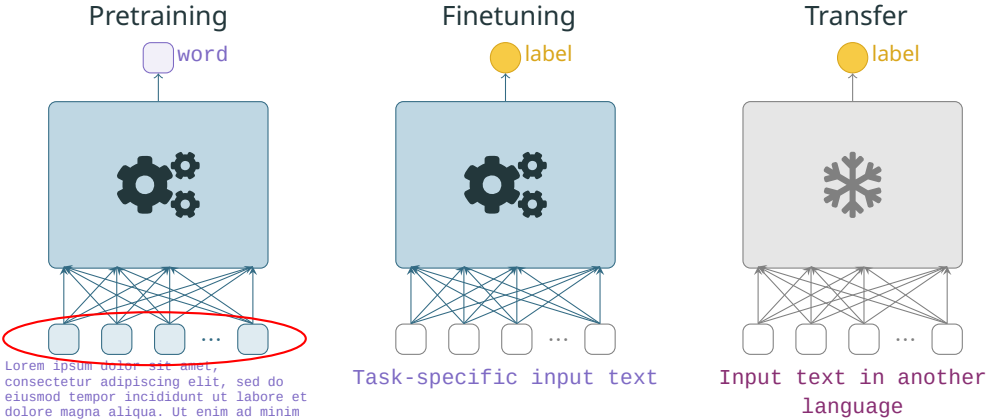


Human-centric NLP
(what tools and why?)

Modelling non-standard data

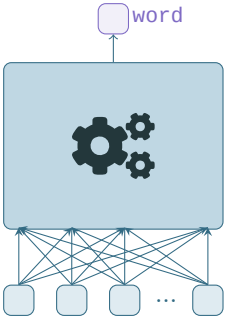
Available dialect data

LLMs: Pretrain – finetune – transfer



LLMs: 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 common character sequences
– “subword tokens” –
to numeric representations

Non-standard orthographies + tokenization

Subword tokenization with GBERT

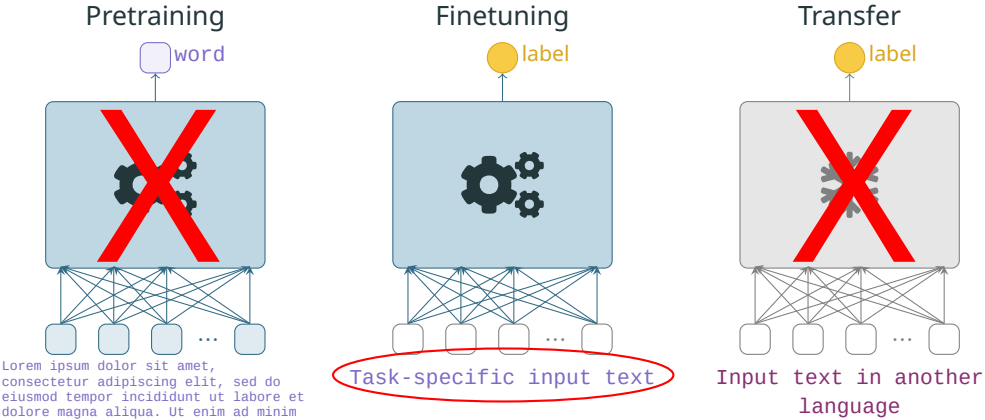
Die	Lammer	hat	ein	recht		sauberes	Wasser
`Die`	`Lamm` `–er`	`hat`	`ein`	`recht`		`sauber` `–es`	`Wasser`
D’	Lomma	hod	a	rechd	a	sauwas	Wossa
`D` `'`	`Lom` `–ma`	`ho` `–d`	`a`	`rech` `–d`	`a`	`sau` `–was`	`Wo` `–ssa`

“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?

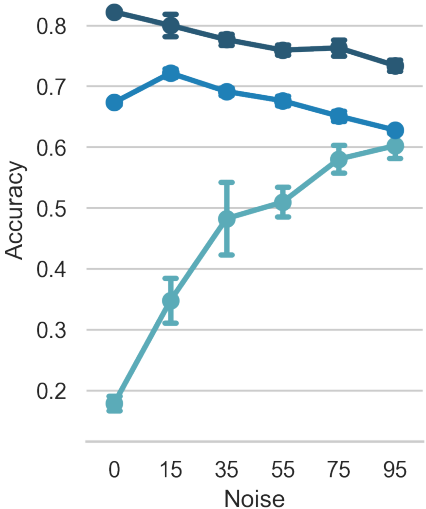


Character-level “noise”

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`

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

How much noise to add?



Finnish → Savonian Finnish

Nynorsk → North Norwegian
German → Low Saxon

Part-of-speech tagging

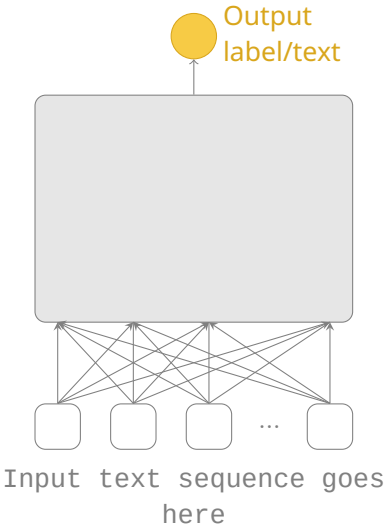
Blaschke/Schütze/Plank, VarDial 2023, "Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages"

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`

Overview



Human-centric NLP
(what tools and why?)

Modelling non-standard data

Available dialect data

What NLP tools and why?

Computational linguistics & machine learning research

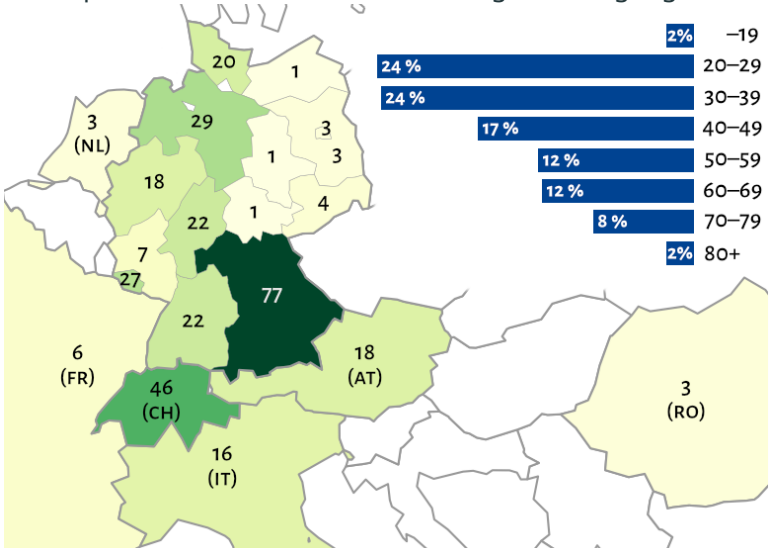
- How to learn from sparse + heterogeneous data?
- Quantitative patterns

NLP tools for linguists – we want to hear from you! :)

Applied language technologies for dialect speakers

Language technology for dialect speakers

>300 speakers of German dialects or regional languages



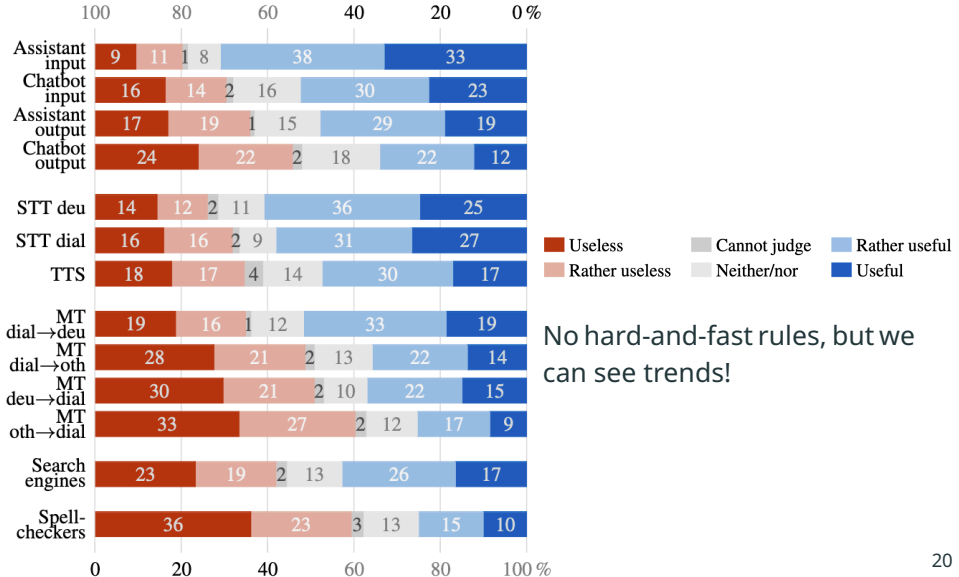
Questionnaire

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...

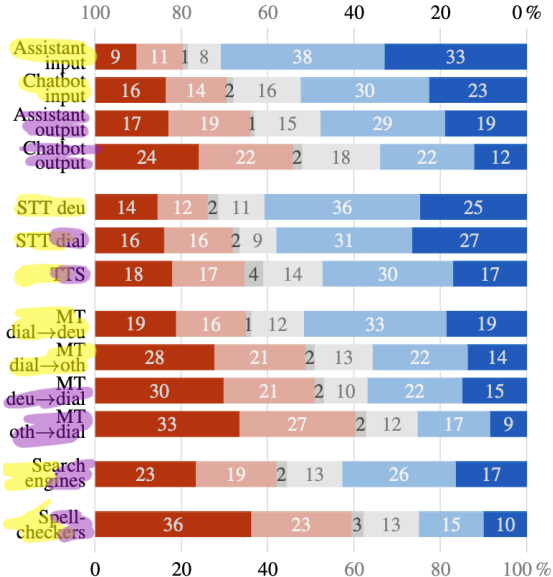
- ...that transcribes audio recorded in my dialect as written Standard German.
- ..that transcribes audio recorded in my dialect as written dialect.

Which dialect LTs are deemed useful?

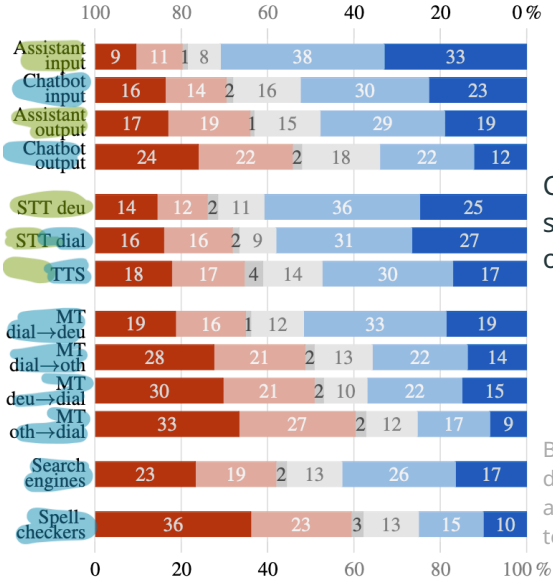


No hard-and-fast rules, but we can see trends!

Dialect input vs. output?



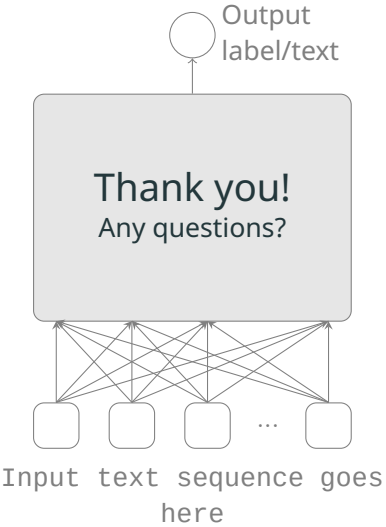
Text vs. speech?



Correlated with opinion on standardized dialect orthographies

Blaschke+, ACL 2024, “What do dialect speakers want? A survey of attitudes towards language technology for German dialects”

Summary – challenges & approaches



Reflecting on
what tools we build

Representing/modelling
non-standard data

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