Natural dialect processing

NLP for non-standardized language varieties

Verena Blaschke & Barbara Plank
MaiNLP lab, LMU Munich

Quantitative approaches in dialectology
and variationist sociolinguistics
December 7, 2023

Natural Language Processing

Natural Language Processing

Natural Language Processing

... but *which* languages?

NLP – but which “language(s)”?

- Many speakers, abundant data, standardization

NLP – but which “language(s)”?

- Many speakers, abundant data, standardization

But how do we actually use language?

NLP – but which “language(s)”?

- Many speakers, abundant data, standardization

But how do we actually use language?

- Also include minority languages, non-standard varieties

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 –
  how can we learn from sparse + heterogeneous data?

NLP – but which “language(s)”?

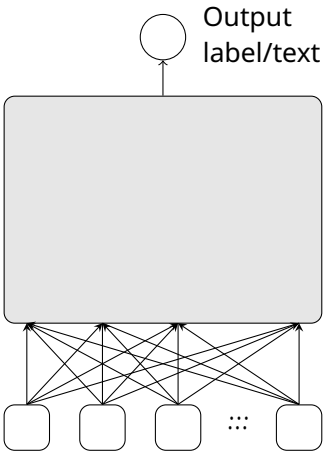
- Many speakers, abundant data, standardization

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 –
  how can we learn from sparse + heterogeneous data?

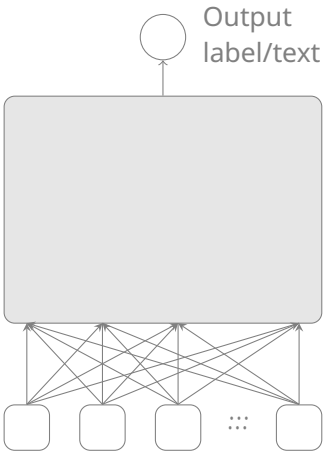
Overview of current challenges and approaches regarding NLP & dialects

Overview



Input text sequence goes here

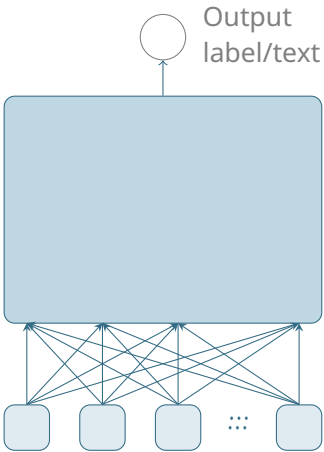
Overview



Input text sequence goes here

ż Available dialect data

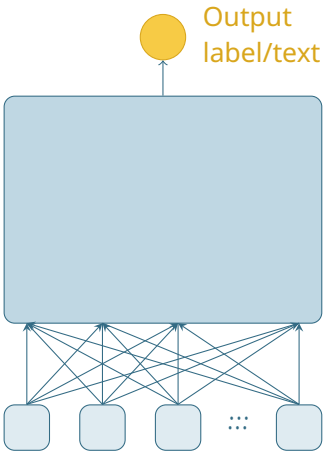
Overview



Input text sequence goes here

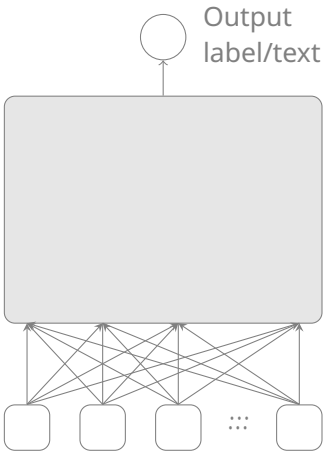
- Modelling non-standard data
- Available dialect data

Overview



- Human-centric NLP (what tools and why?)
- Modelling non-standard data
- Available dialect data

Overview



Input text sequence goes here

- Human-centric NLP (what tools and why?)
- Modelling non-standard data
- Available dialect data

(Lack of?) resources

Datasets for dialects and “small” languages

(Lack of?) resources

Datasets for dialects and “small” languages

- Two communities: variationists & NLP researchers

(Lack of?) resources

Datasets for dialects and “small” languages

- Two communities: variationists & NLP researchers
- Findable; licenses allowing re-use

(Lack of?) resources

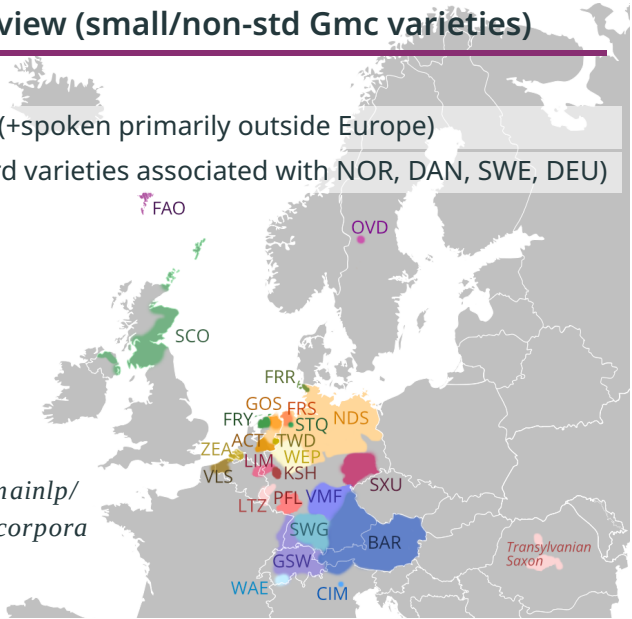
Datasets for dialects and “small” languages

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

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)

100+ (mostly written) corpora for 30 language varieties

Blaschke ea “A survey of corpora for Germanic low-resource languages and dialects” NoDaLiDa 2023

Corpus overview (small/non-std Gmc varieties)

100+ (mostly written) corpora for 30 language varieties

- Largely unannotated

Blaschke ea “A survey of corpora for Germanic low-resource languages and dialects” NoDaLiDa 2023

Corpus overview (small/non-std Gmc varieties)

100+ (mostly written) corpora for 30 language varieties

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

Blaschke ea “A survey of corpora for Germanic low-resource languages and dialects” NoDaLiDa 2023

Corpus overview (small/non-std Gmc varieties)

100+ (mostly written) corpora for 30 language varieties

- Largely unannotated
- If annotated:
  - Geolocation, dialect group
  - Morphosyntax
  - Rarely: translations, content-related annotations
- Mostly dedicated, curated corpora
  Recently: also uncurated, web-crawled ones

Blaschke ea "A survey of corpora for Germanic low-resource languages and dialects" NoDaLiDa 2023

Corpus overview (small/non-std Gmc varieties)

100+ (mostly written) corpora for 30 language varieties

- Largely unannotated
- If annotated:
  - Geolocation, dialect group
  - Morphosyntax
  - Rarely: translations, content-related annotations
- Mostly dedicated, curated corpora
  Recently: also uncurated, web-crawled ones

Data exchange between research communities?

Blaschke ea "A survey of corpora for Germanic low-resource languages and dialects" NoDaLiDa 2023

Bavarian dependency treebank

Work under review

- Different dialects
  (location metadata)

Map based on Wiesinger (1983, map 47.4)

Bavarian dependency treebank

Work under review

- Different dialects
  (location metadata)
- Different text genres

Map based on Wiesinger (1983, map 47.4)

Bavarian dependency treebank

Work under review

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

Map based on Wiesinger (1983, map 47.4)

Bavarian dependency treebank

Work under review

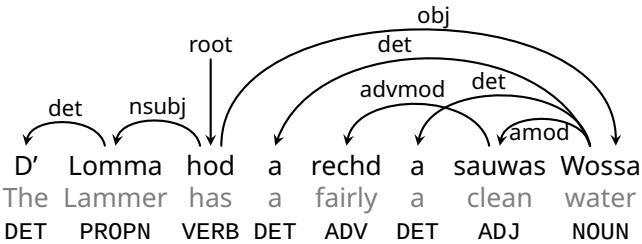
- Different dialects (location metadata)
- Different text genres
- Freely accessible, permissive license
- 11k words, 750 sentences

Map based on Wiesinger (1983, map 47.4)

Bavarian dependency treebank

Work under review

Universal Dependencies



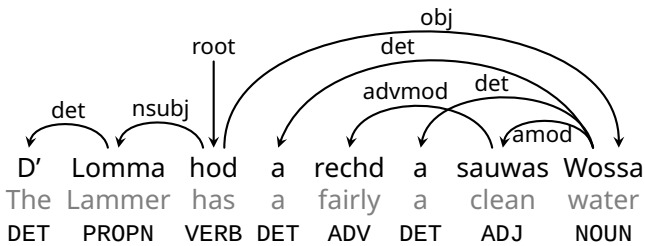
de Marneffe ea "Universal Dependencies" *Computational Linguistics* (2021)

Bavarian dependency treebank

Work under review

Universal Dependencies

- Cross-linguistic comparability (incl. DEU, GSW, NDS)



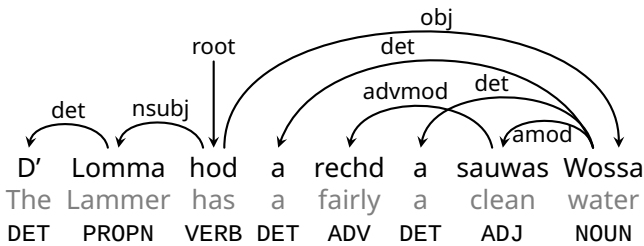
de Marneffe ea “Universal Dependencies” *Computational Linguistics* (2021)

Bavarian dependency treebank

Work under review

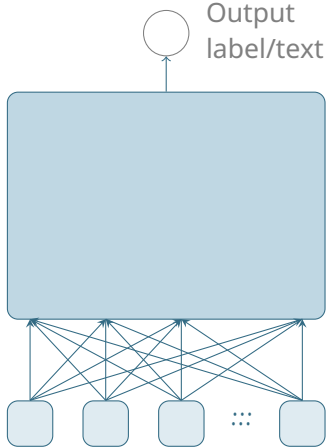
Universal Dependencies

- Cross-linguistic comparability (incl. DEU, GSW, NDS)
- Established for automatic annotation tasks



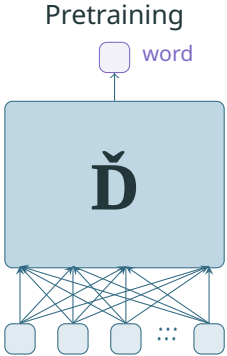
de Marneffe ea “Universal Dependencies” *Computational Linguistics* (2021)

Overview



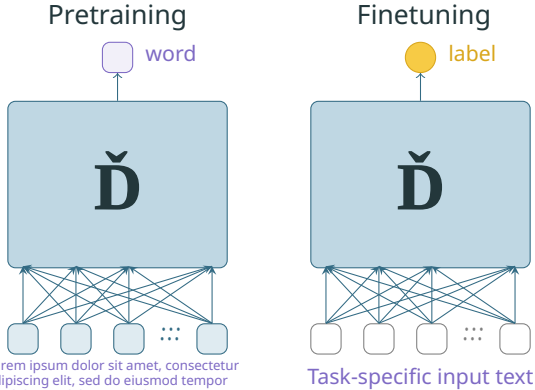
- Human-centric NLP (what tools and why?)
- Modelling non-standard data
- Available dialect data

Pretrain – finetune – transfer



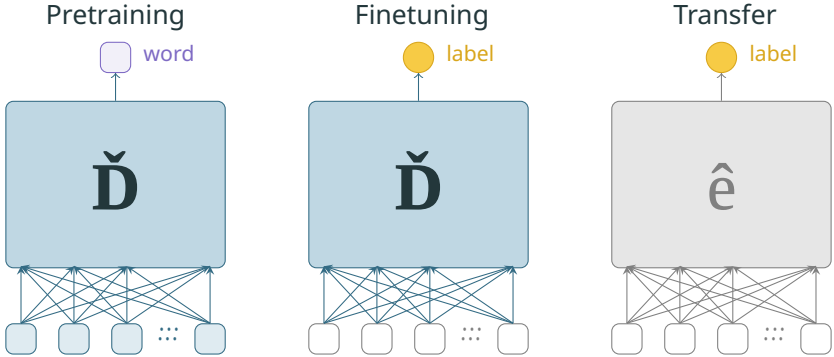
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum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

Pretrain – finetune – transfer



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

Pretrain – finetune – transfer

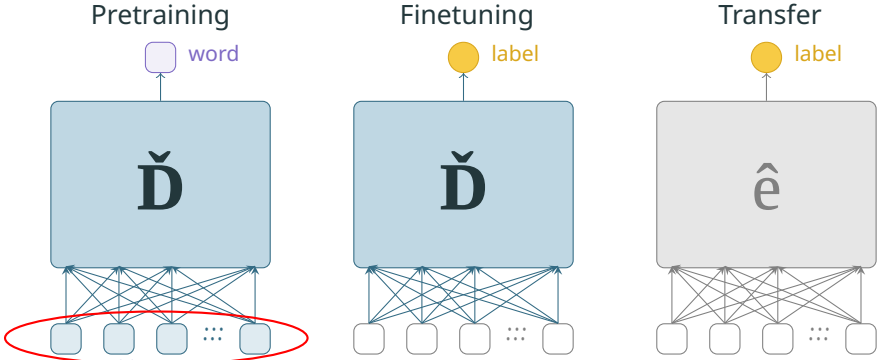


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum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

Task-specific input text

Input text in another language

Pretrain – finetune – transfer

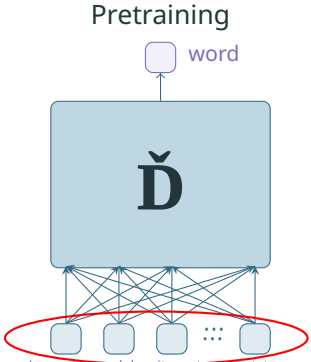


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum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

Task-specific input text

Input text in another language

Pretrain – finetune – transfer



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

Encoding input text

Map character sequences (“subword tokens”) to numeric representations

Non-standard orthographies + tokenization

Die	Lammer	hat	ein	recht		sauberes	Wasser
D’	Lomma	hod	a	rechd	a	sauwas	Wossa

Sentence via bar.wikipedia.org/wiki/L¬mma
Chan ea “German’s next language model” 2020

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

Sentence via `bar.wikipedia.org/wiki/L¬mma`
Chan ea “German’s next language model” 2020

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`

Sentence via `bar.wikipedia.org/wiki/L¬mma`
Chan ea “German’s next language model” 2020

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`

Sidenote:
ChatGPT/GPT-4/etc also use such kinds of tokenization

Sentence via `bar.wikipedia.org/wiki/L¬mma`
Chan ea “German’s next language model” 2020

Non-standard orthographies + tokenization

D’	Lomma	hod	a	rechd	a	sauwas	Wossa
D '	Lom –ma	ho –d	a	rech –d	a	sau –was	Wo –ssa

Solutions – ongoing work in the field

Non-standard orthographies + tokenization

D’	Lomma	hod	a	rechd	a	sauwas	Wossa
D '	Lom –ma	ho –d	a	rech –d	a	sau –was	Wo –ssa

Solutions – ongoing work in the field

- Move away from character-sequence-based representations altogether – requires training new models, experimental (overall performance might be worse)

Non-standard orthographies + tokenization

D’	Lomma	hod	a	rechd	a	sauwas	Wossa
D '	Lom –ma	ho –d	a	rech –d	a	sau –was	Wo –ssa

Solutions – ongoing work in the field

- Move away from character-sequence-based representations altogether – requires training new models, experimental (overall performance might be worse)
- (How to) make existing models more robust to variation

Brittleness towards uncommon structures

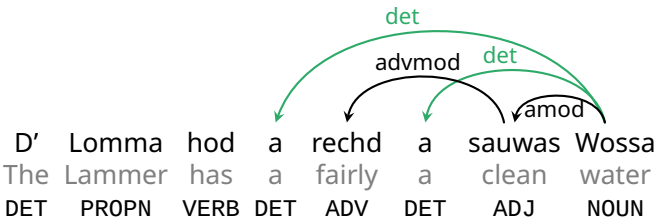
Work under review

D’	Lomma	hod	a	rechd	a	sauwas	Wossa
The	Lammer	has	a	fairly	a	clean	water
DET	PROPN	VERB	DET	ADV	DET	ADJ	NOUN

Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

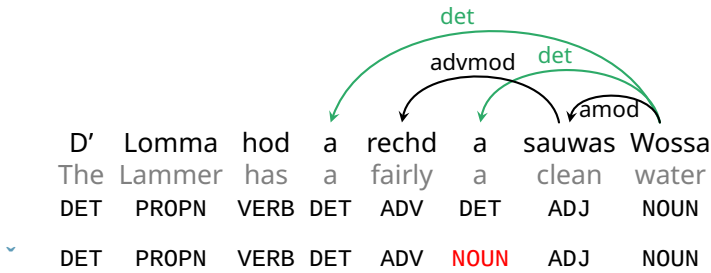
Work under review



Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

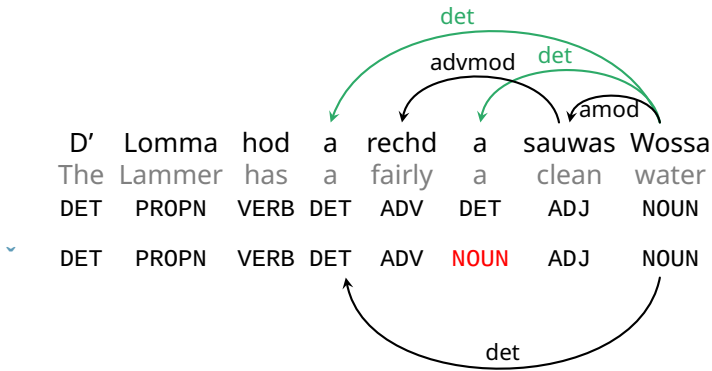
Work under review



Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

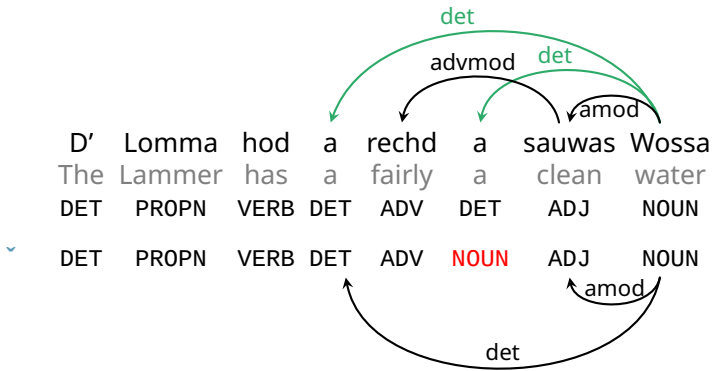
Work under review



Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

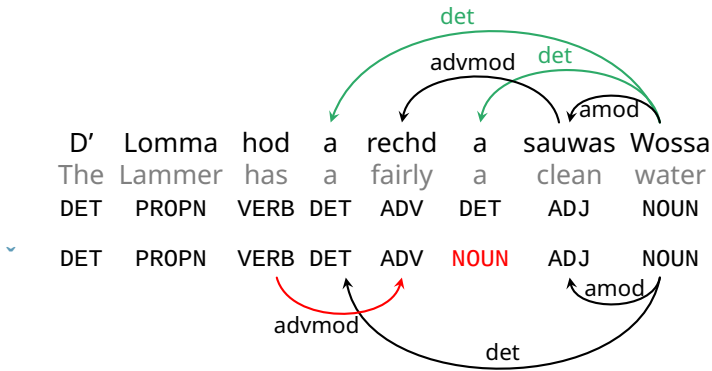
Work under review



Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

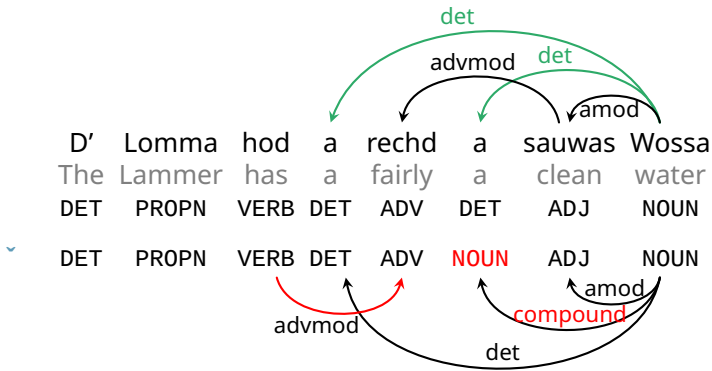
Work under review



Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

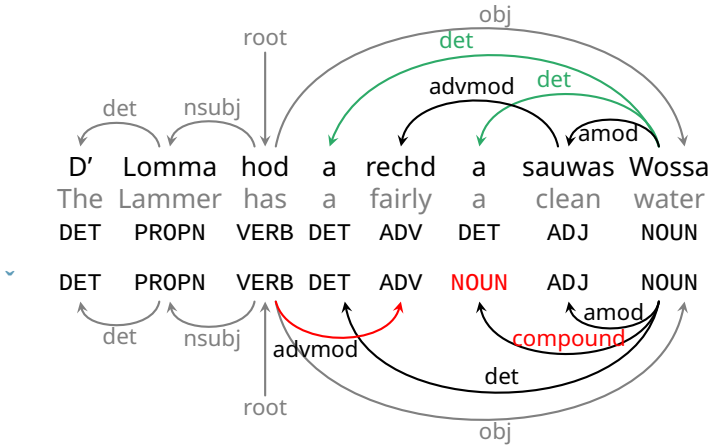
Work under review



Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

Work under review

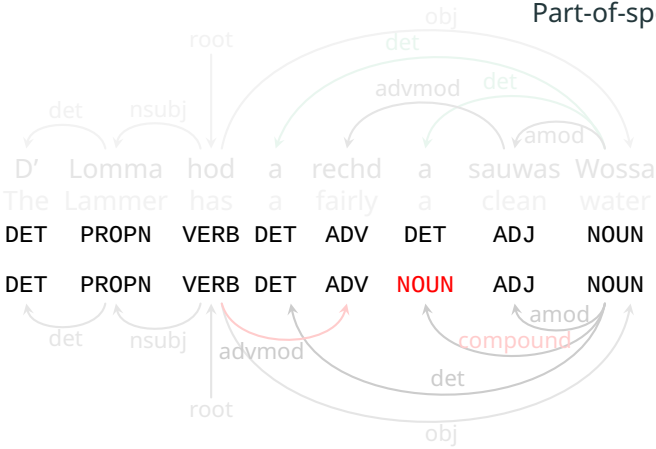


Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

Work under review

Part-of-speech tags:
Accuracy

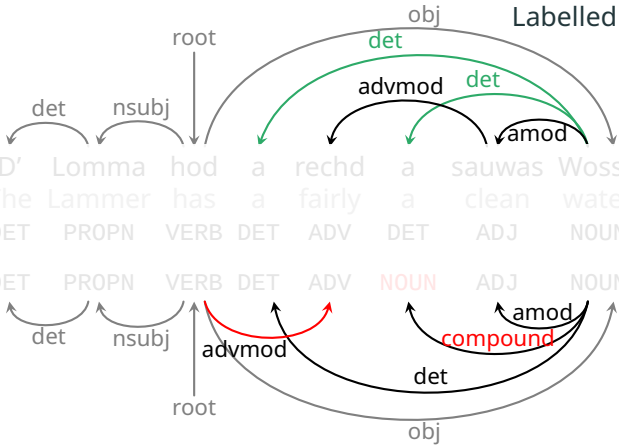


Sentence via bar.wikipedia.org/wiki/L¬mma

Brittleness towards uncommon structures

Work under review

Dependencies: Labelled attachment score



Sentence via bar.wikipedia.org/wiki/L¬mma

Automatic tagging/parsing

Work under review

Train on German data (there is no Bavarian training data!), test on German vs. Bavarian

Model	Test lang	Acc (%)	LAS (%)
Stanza	DEU	95.9	83.7
GBERT	DEU	96.8	83.1
UDPipe	DEU	96.5	84.9

Acc = accuracy (part-of-speech tags); LAS = labelled attachment score

Automatic tagging/parsing

Work under review

Train on German data (there is no Bavarian training data!), test on German vs. Bavarian

Model	Test lang	Acc (%)	LAS (%)
Stanza	DEU	95.9	83.7
GBERT	DEU	96.8	83.1
UDPipe	DEU	96.5	84.9
Stanza	BAR	40.9	23.1
GBERT	BAR	57.4	30.1
UDPipe	BAR	80.5	67.3

Acc = accuracy (part-of-speech tags); LAS = labelled attachment score

Automatic tagging/parsing

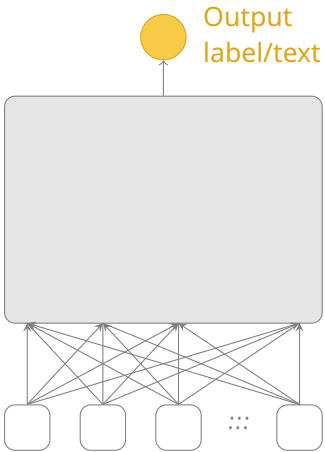
Work under review

Train on German data (there is no Bavarian training data!), test on German vs. Bavarian

Model	Test lang	Acc (%)	LAS (%)	Input representation
Stanza	DEU	95.9	83.7	
GBERT	DEU	96.8	83.1	
UDPipe	DEU	96.5	84.9	
Stanza	BAR	40.9	23.1	Full words
GBERT	BAR	57.4	30.1	Subword tokens
UDPipe	BAR	80.5	67.3	Subword tok. + characters

Acc = accuracy (part-of-speech tags); LAS = labelled attachment score

Overview



Input text sequence goes here

- Human-centric NLP (what tools and why?)
- Modelling non-standard data
- Available dialect data

What NLP tools and why?

Computational linguistics & machine learning research

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

Blokland ea “Language documentation meets language technology” *Septentrio conference series* (2015)

What NLP tools and why?

Computational linguistics & machine learning research

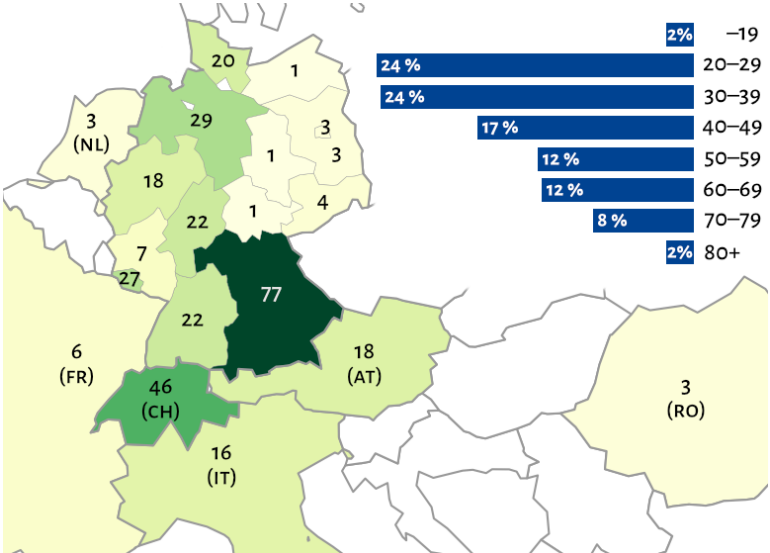
- Quantitative patterns
- How to learn from sparse + heterogeneous data?

NLP tools for linguists – also require dialogue between the communities!

Blokland ea “Language documentation meets language technology” *Septentrio conference series* (2015)

Language technology for dialect speakers

Work under review



Language technology for dialect speakers

Work under review

Self-identified dialect speakers

- How often / in what contexts do people speak/write their dialects?
- What do they think about various language-related technologies for their dialects? (e.g., spellcheckers, machine translation, digital assistants...)

Language technology for dialect speakers

Work under review

Self-identified dialect speakers

- How often / in what contexts do people speak/write their dialects?
- What do they think about various language-related technologies for their dialects? (e.g., spellcheckers, machine translation, digital assistants…)

Results

- Dialect input > dialect output

Language technology for dialect speakers

Work under review

Self-identified dialect speakers

- How often / in what contexts do people speak/write their dialects?
- What do they think about various language-related technologies for their dialects? (e.g., spellcheckers, machine translation, digital assistants…)

Results

- Dialect input > dialect output
- Processing speech > processing text

Language technology for dialect speakers

Work under review

Self-identified dialect speakers

- How often / in what contexts do people speak/write their dialects?
- What do they think about various language-related technologies for their dialects? (e.g., spellcheckers, machine translation, digital assistants…)

Results

- Dialect input > dialect output
- Processing speech > processing text
- Speaker communities vary in their attitudes!

Language technology for dialect speakers

Work under review

Self-identified dialect speakers

- How often / in what contexts do people speak/write their dialects?
- What do they think about various language-related technologies for their dialects? (e.g., spellcheckers, machine translation, digital assistants…)

Results

- Dialect input > dialect output
- Processing speech > processing text
- Speaker communities vary in their attitudes!
    - Between subgroups (e.g., GSW vs. NDS speakers)

Language technology for dialect speakers

Work under review

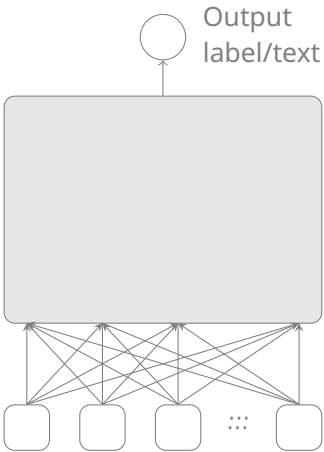
Self-identified dialect speakers

- How often / in what contexts do people speak/write their dialects?
- What do they think about various language-related technologies for their dialects? (e.g., spellcheckers, machine translation, digital assistants…)

Results

- Dialect input > dialect output
- Processing speech > processing text
- Speaker communities vary in their attitudes!
  - Between subgroups (e.g., GSW vs. NDS speakers)
  - Within subgroups

Summary



- Reflecting on what tools we build
- Representing/modelling non-standard data
- Data availability

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

Summary

- Reflecting on what tools we build
- Representing/modelling non-standard data
- Data availability
- ! *github.com/mainlp/germanic-lrl-corpora*