MaiBaam

A Multi-Dialectal Bavarian Universal Dependency Treebank

Verena Blaschke, Barbara Kovačić, Siyao Peng,
Hinrich Schütze & Barbara Plank
LMU Munich

LREC–COLING 2024

Motivation

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

Abaza	1	<1K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari
Albanian	1	<1K		IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	3	456K		IE, Greek
Ancient Hebrew	1	39K		Afro-Asiatic, Semitic
Apurina	1	<1K		Arawakan
Arabic	3	1,042K		Afro-Asiatic, Semitic
Armenian	2	94K		IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Beja	1	1K		Afro-Asiatic, Cushitic
Belarusian	1	305K		IE, Slavic
Bengali	1	<1K		IE, Indic
Bhojpuri	1	6K		IE, Indic
Bororo	1	1K		Bororoan
Breton	1	10K		IE, Celtic
Bulgarian	1	156K		IE, Slavic
Buryat	1	10K		Mongolic

Bavarian

- North
- North/Central
- Central
- South/Central
- South

Bavarian

Ũ Trotzdean das’e’s moch, hairon tou’e’s niat.

Ũ Obwoi i’s mog, heirodn dua e’s ned.

Ũ Trotz dass i’s mog, hairatn tua i’s net.

DEU Obwohl ich sie mag, heirate ich sie nicht.

‘Although I like her I won’t marry her.’

Sentence via bar.wikipedia.org/wiki/Boarische_Grammatik_(Konjunktiona), CC BY-SA

Data

- 15k tokens
- 1 070 sentences
- Metadata: location/dialect area

Text genres and sources

- Wiki articles
- Wiki discussion pages
- Grammar examples
- Queries for virtual assistants
- Fairy tales

Annotation procedure

- 165 h pure annotation time
- Train an annotator on the existing German treebanks
- Weekly discussion of annotations and difficult cases
- Partially pre-annotate POS tags
- No normalization
- Validation

Tokenization

- Multi-word tokens: ADP+DET, PART+DET
- Token sequences with no whitespace in between
  (shortened DET/ADP/PRON and ‘normal’ VERB/AUX/NOUN/…)

Sentence via bar.wikipedia.org/wiki/Text:Hansl_und_Gretl, CC BY-SA

Tokenization

- Multi-word tokens: ADP+DET, PART+DET
- Token sequences with no whitespace in between
  (shortened DET/ADP/PRON and ‘normal’ VERB/AUX/NOUN/...)

Dann habnses ankent ...
‘Then they set it on fire...’

Dann	habn	se	s	ankent	...
Then	have.3PL	they	it	lighted	...
ADV	AUX	PRON	PRON	VERB	...

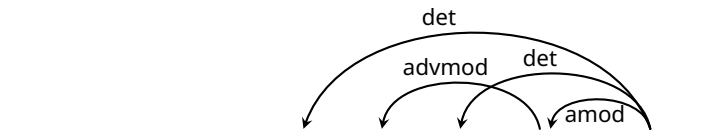
Sentence via bar.wikipedia.org/wiki/Text:Hansl_und_Gretl, CC BY-SA

Syntax

Differences to German

- Verbs (infinitive markers, auxiliaries)
- Noun phrases (word order in specific NP structures, personal names, possession)
- Subordinate clauses (relative markers, additional and/or inflected complementizers)
- And more! (dropped pronouns, negative concord, ...)

Syntax



‘The Lammer (river) has fairly clean water’

Sentence via bar.wikipedia.org/wiki/Låmma, CC BY-SA

Experiments

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

Out-of-the-box models

- UDPipe
- Stanza

Own models

- GBERT
- mBERT
- XLM-R

Additional experiments in paper, incl. tokenization-related experiments!

Experiments

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

Out-of-the-box models

- **UDPipe**
- **Stanza**

Own models

- **GBERT**
- mBERT
- XLM-R

Additional experiments in paper, incl. tokenization-related experiments!

Experiments

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
Additional experiments in paper, incl. tokenization-related experiments!

Experiments

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	42.30	24.89
GBERT	BAR	58.86	36.40
UDPipe	BAR	80.29	65.79

Acc = accuracy (part-of-speech tags); LAS = labelled attachment score
Additional experiments in paper, incl. tokenization-related experiments!

Experiments

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	42.30	24.89	Full words
GBERT	BAR	58.86	36.40	Subword tokens
UDPipe	BAR	80.29	65.79	Subword tok. + characters

Acc = accuracy (part-of-speech tags); LAS = labelled attachment score
Additional experiments in paper, incl. tokenization-related experiments!

Conclusion

POS tagging & dependency parsing isn't trivial, even when we have plenty of training data from a closely related language at our disposal!

- Preprint: `arxiv.org/abs/2403.10293`
- Data: `github.com/UniversalDependencies/UD_Bavarian-MaiBaam`
- Code: `github.com/mainlp/maibaam-code`
- Annotation guidelines: `arxiv.org/abs/2403.05902`

Conclusion

POS tagging & dependency parsing isn't trivial, even when we have plenty of training data from a closely related language at our disposal!

- Preprint: `arxiv.org/abs/2403.10293`
- Data: `github.com/UniversalDependencies/UD_Bavarian-MaiBaam`
- Code: `github.com/mainlp/maibaam-code`
- Annotation guidelines: `arxiv.org/abs/2403.05902`

More Bavarian NLP @ LREC–COLING 2024:

- *Sebastian, Basti, Wastl?! Recognizing Named Entities in Bavarian Dialectal Data* (Peng et al.)
- *Slot and Intent Detection Resources for Bavarian and Lithuanian: Assessing Translations vs. Natural Queries to Digital Assistants* (Winkler et al.)