Creating a Universal Dependencies treebank

Verena Blaschke (LMU Munich)

Linguistic Annotation Frameworks, May 13, 2024

commons.wikimedia.org/wiki/File:Aussicht_vom_Baumturm,_Nationalpark_Bayerischer_Wald.jpg
(CC BY-SA 4.0)

We'll talk about a different kind of Bavarian forest

Today

MaiBaam (Bavarian treebank)

Considerations: data, preprocessing, annotations

Tools

Training/evaluating ML models

Today

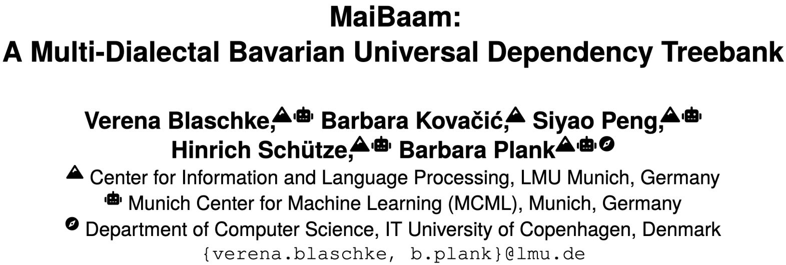
MaiBaam (Bavarian treebank)

Considerations: data, preprocessing, annotations

Tools

Training/evaluating ML models

MaiBaam



MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank

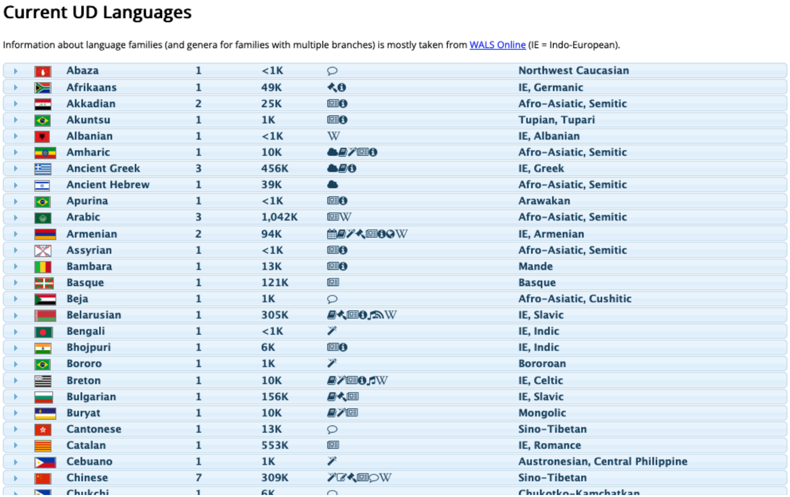
Verena Blaschke,▲🤖 Barbara Kovačić,▲ Siyao Peng,▲🤖
Hinrich Schütze,▲🤖 Barbara Plank▲🤖🧭
▲ Center for Information and Language Processing, LMU Munich, Germany
🤖 Munich Center for Machine Learning (MCML), Munich, Germany
🧭 Department of Computer Science, IT University of Copenhagen, Denmark
`{verena.blaschke, b.plank}@lmu.de`

Universal Dependencies

- Focus on cross-linguistic comparability (rather than perfectly capturing any one language's idiosyncracies)
    - Can be used for research on syntactic typology
- Simple to learn
- Established for automatic annotation tasks

"Universal Dependencies" (de Marneffe+, CL 2021)
"Multilingual gradient word-order typology from Universal Dependencies" (Baylor+, EACL 2024)

UD + language variation



Current UD Languages

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

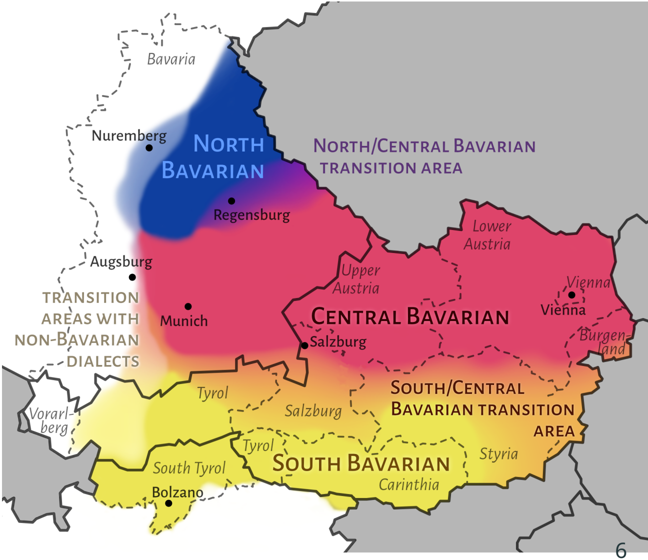
Abaza	1	<1K	💬	Northwest Caucasian
Afrikaans	1	49K	⚖ℹ	IE, Germanic
Akkadian	2	25K	📰ℹ	Afro-Asiatic, Semitic
Akuntsu	1	1K	📰ℹ	Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K	☁📖✎📰ℹ	Afro-Asiatic, Semitic
Ancient Greek	3	456K	☁📖ℹ	IE, Greek
Ancient Hebrew	1	39K	☁	Afro-Asiatic, Semitic
Apurina	1	<1K	📰ℹ	Arawakan
Arabic	3	1,042K	📰W	Afro-Asiatic, Semitic
Armenian	2	94K	📅📖✎⚖📰ℹ🌐W	IE, Armenian
Assyrian	1	<1K	📰ℹ	Afro-Asiatic, Semitic
Bambara	1	13K	📰ℹ	Mande
Basque	1	121K	📰	Basque
Beja	1	1K	💬	Afro-Asiatic, Cushitic
Belarusian	1	305K	📖⚖📰ℹ♬W	IE, Slavic
Bengali	1	<1K	✎	IE, Indic
Bhojpuri	1	6K	📰ℹ	IE, Indic
Bororo	1	1K	✎	Bororoan
Breton	1	10K	📖✎📰ℹ♬W	IE, Celtic
Bulgarian	1	156K	📖⚖📰	IE, Slavic
Buryat	1	10K	📖✎📰	Mongolic
Cantonese	1	13K	💬	Sino-Tibetan
Catalan	1	553K	📰	IE, Romance
Cebuano	1	1K	✎	Austronesian, Central Philippine
Chinese	7	309K	✎✍⚖📰💬W	Sino-Tibetan
Chukchi	1	6K	💬	Chukotko-Kamchatkan

Why UD for a dialect?

- Research on language variation
  - Investigate Bavarian morphosyntax
  - Compare annotated morphosyntactic structures with German, Swiss German treebanks
- ML research
  - How well does transfer from a standard language to a closely related non-standard variety work?
  - How well can we learn from sparse, heterogeneous data?

Bavarian

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



Data

- 15k tokens
- 1070 sentences
- Metadata:
    - Location/dialect area
    - Text genre & source

Annotation procedure

POS tags + syntactic dependencies

- Train an annotator on the existing German treebanks
- Weekly discussion of annotations and difficult cases
- 165 h pure annotation time
  (+ adjudication, training, literature research, corrections, ...)
- Partially pre-annotate POS tags

Today

MaiBaam (Bavarian treebank)

Considerations: data, preprocessing, annotations

Tools

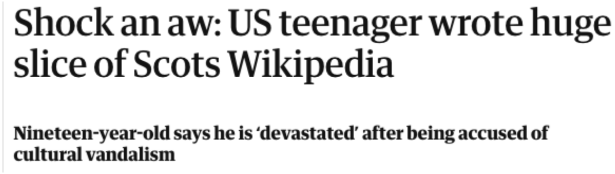
Training/evaluating ML models

General considerations

- Are there treebanks in related languages? Treebanks in the same language but using another annotation scheme?
    - Inspiration/help for tricky cases
    - Comparison afterwards (how does my dataset differ from XYZ?)
- Access to linguistic literature about this language?
    - Some terms are defined differently by UD than in some traditional grammars
- Expertise of native speakers, language experts?

Data selection

- Permissive licenses, e.g. Creative Commons
- Data quality



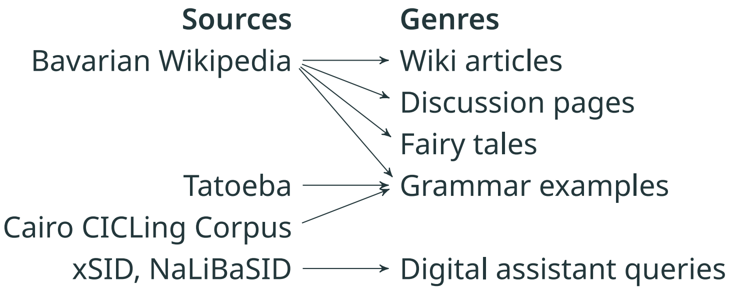
Shock an aw: US teenager wrote huge slice of Scots Wikipedia

Nineteen-year-old says he is ‘devastated’ after being accused of cultural vandalism

- What genres, styles, registers?
- What grammatical structures can we expect?

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

Data selection in MaiBaam



Sources

Bavarian Wikipedia

Tatoeba

Cairo CICLing Corpus

xSID, NaLiBaSID

Genres

Wiki articles

Discussion pages

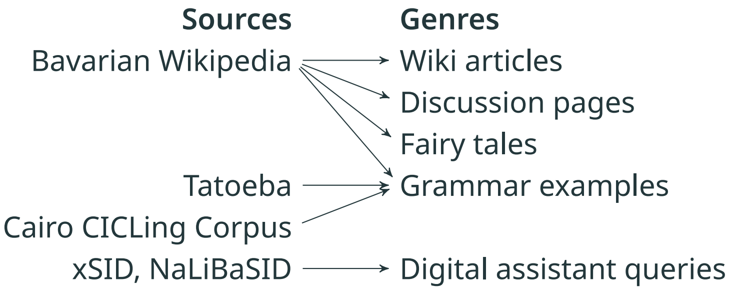
Fairy tales

Grammar examples

Digital assistant queries

bar.wikipedia.org, tatoeba.org, github.com/UniversalDependencies/cairo,
github.com/mainlp/xsid (van der Goot+ 2020)
github.com/mainlp/NaLiBaSID (Winkler+ 2024)

Data selection in MaiBaam



Sources

Bavarian Wikipedia

Tatoeba

Cairo CICLing Corpus

xSID, NaLiBaSID

Genres

Wiki articles

Discussion pages

Fairy tales

Grammar examples

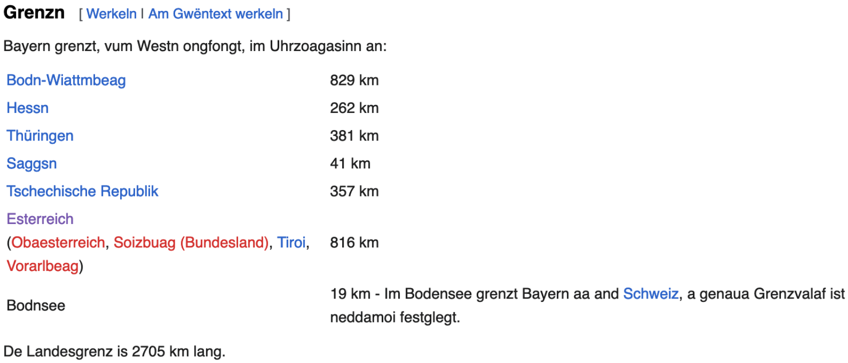
Digital assistant queries

- 1st, **2nd**, 3rd person
- Declarative, interrogative, imperative
- Present, **past**

bar.wikipedia.org, tatoeba.org, github.com/UniversalDependencies/cairo,
github.com/mainlp/xsid (van der Goot+ 2020)
github.com/mainlp/NaLiBaSID (Winkler+ 2024)

Sentences

- Sentence splitting
- Include section titles? Represent formatting?
- Normalization/typos



Grenzn [Werkeln | Am Gwëntext werkeln]

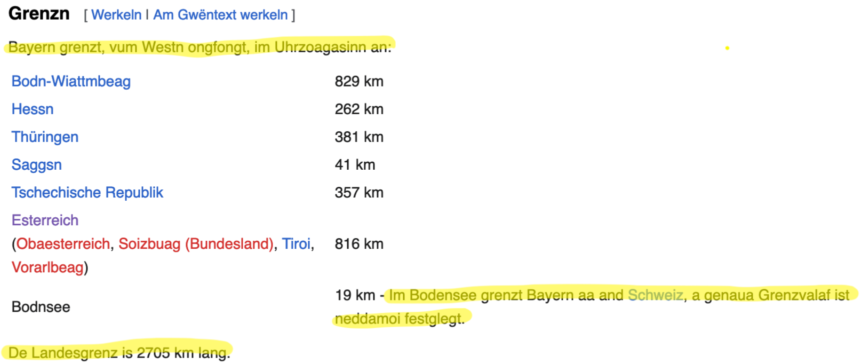
Bayern grenzt, vum Westn ongfongt, im Uhrzoagasinn an:

Bodn-Wiattmbeag	829 km
Hessn	262 km
Thüringen	381 km
Saggsn	41 km
Tschechische Republik	357 km
Esterreich (Obaesterreich, Soizbuag (Bundesland), Tiroi, Vorarlbeag)	816 km
Bodnsee	19 km - Im Bodensee grenzt Bayern aa and Schweiz, a genaua Grenzvalaf ist neddamoi festglegt.

De Landesgrenz is 2705 km lang.

Sentences

- Sentence splitting
- Include section titles? Represent formatting?
- Normalization/typos



Grenzn [Werkeln | Am Gwëntext werkeln]

Bayern grenzt, vum Westn ongfongt, im Uhrzoagasinn an:

Bodn-Wiattmbeag	829 km
Hessn	262 km
Thüringen	381 km
Saggsn	41 km
Tschechische Republik	357 km
Esterreich (Obaesterreich, Soizbuag (Bundesland), Tiroi, Vorarlbeag)	816 km
Bodnsee	19 km - Im Bodensee grenzt Bayern aa and Schweiz, a genaua Grenzvalaf ist neddamoi festglegt.

De Landesgrenz is 2705 km lang.

Tokenization

General tokenization considerations

- Abbrevations: *e.g., i.e.; z.B., bspw.*
- Hyphenated compounds: *left-handed; Dialekt-Forschung*

Bavarian, German, ...

- “Multi-word tokens”: fused preposition/determiner
    - *am* (“at the”)
      German UD: treat as *an dem*
      Bavarian UD: treat as *a m* (why no normalization?)

Tokenization II

Bavarian, German, ...

- “Multi-word tokens”: fused preposition/determiner
- Token sequences frequently without whitespace (shortened DET/ADP/PRON after 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

POS tags and syntactic dependencies

- 17 POS tags
- 37 dependency relations (+ subtypes)
- Often rather straight-forward
- Sometimes ambiguous

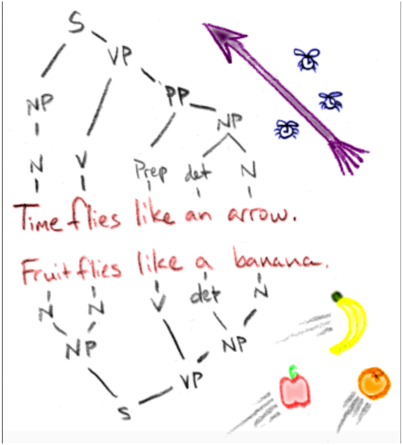


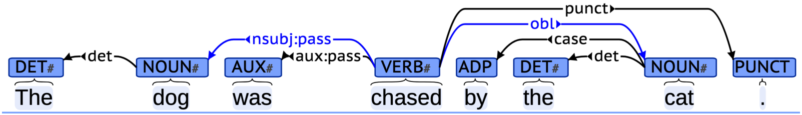
Image source: specgram.com/CLIII.4/08.phlogiston.cartoon.zhe.html

Syntactic dependencies

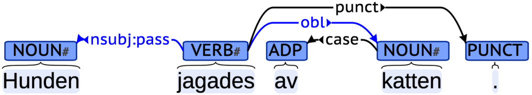
3 basic categories

- Nominals (∼ noun phrases)
- Clauses – verbs are heads of clauses
- Modifiers (∼ adjectives, adverbs)

Focus on content words:



punct
obl
case
nsubj:pass
det
aux:pass
det
DET# NOUN# AUX# VERB# ADP DET# NOUN# PUNCT
The dog was chased by the cat .



Example via universaldependencies.org/introduction.html

UD-specific definitions

E.g. for German + related languages:

- Modal particles (*ich sehe es* **ja**) → adverb
- *nicht* (“not”) → particle
- Possessive pronouns (*mein*, “my”) → determiner
- Indirect objects (iobj) reserved for verbs with 2 accusative objects (“the news cost [the CEO] [his job]”) – dative/genitive objects are considered oblique arguments (obl:arg)

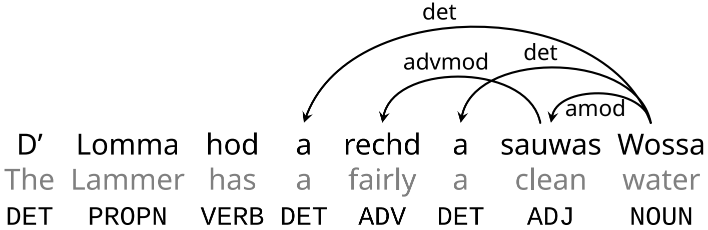
Example via universaldependencies.org/de/dep/iobj.html

Bavarian syntax

Differences to German

- Personal names
    - Anna Schmid – die Schmid Anna
- Additional complementizers
    - Ich möchte wissen, wie lange **dass** das noch dauert
      lit. “I want to.know how long **that** this still takes”
- Dropped 2nd person pronouns
    - Kannst [du] aufstehen?
      lit. “Can.2SG [you] get.up?”
- And more!

Syntax: crossing dependencies



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

‘The Lammer (river) has fairly clean water’

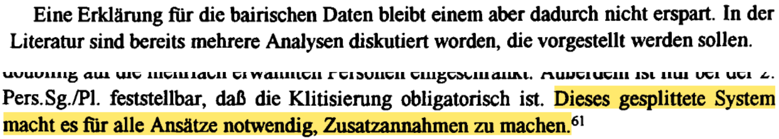
Sentence via bar.wikipedia.org/wiki/Låmma, CC BY-SA

Syntax: 2nd person complementizers

- Er will, dass ich rede.
- Er will, da**sst** (du) redst.

How to analyze this? → Dropped/doubled pronoun? Inflected *dass*?

What if there are conflicting established analyses by different linguists? Or if it’s not really mentioned by anyone?

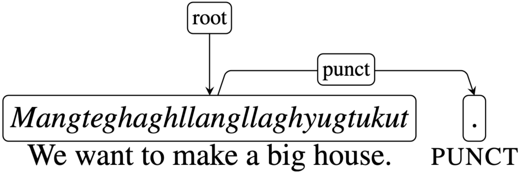


Eine Erklärung für die bairischen Daten bleibt einem aber dadurch nicht erspart. In der Literatur sind bereits mehrere Analysen diskutiert worden, die vorgestellt werden sollen.

doubling auf die mehrfach erwähnten Personen eingeschränkt. Außerdem ist nur bei der 2. Pers.Sg./Pl. feststellbar, daß die Klitisierung obligatorisch ist. Dieses gesplittete System macht es für alle Ansätze notwendig, Zusatzannahmen zu machen.[61]

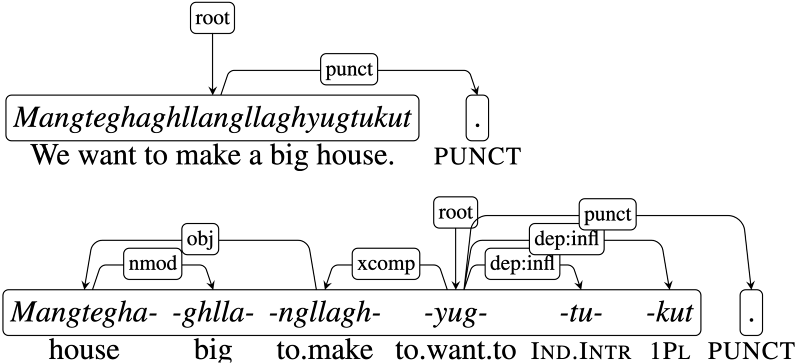
“Syntax des Bairischen” (Weiß 2013), pp. 123, 133

Sidenote: Are UD's rules ideal for all languages?



"Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik" (Park+ 2021)

Sidenote: Are UD's rules ideal for all languages?



"Expanding Universal Dependencies for Polysynthetic Languages:
A Case of St. Lawrence Island Yupik" (Park+ 2021)

Sidenote II: Possible extensions

Possible additional annotation layers

- Lemmas
- Glosses
- Morphological features
- Named entities
- …

Today

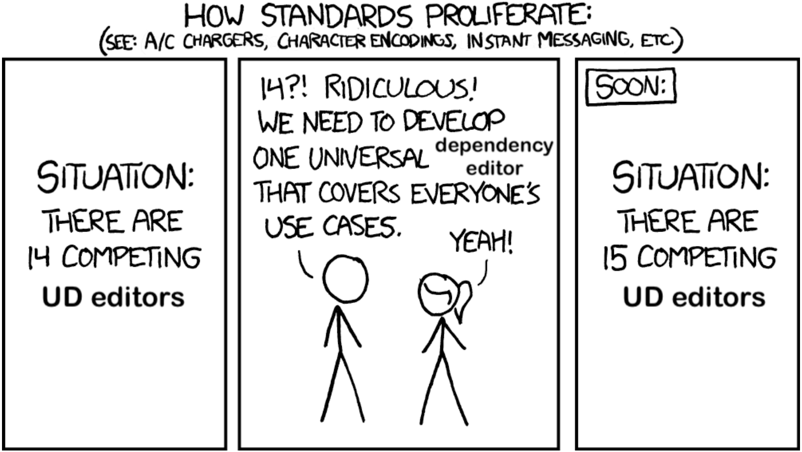
MaiBaam (Bavarian treebank)

Considerations: data, preprocessing, annotations

Tools

Training/evaluating ML models

Annotation tools


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING UD editors
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL dependency editor THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING UD editors

Original: xkcd.com/927 (CC BY-NC 2.5)

Consistency checks

- Does each sentence have a *root*?
- Does the dependent of the *auxiliary* relation have the POS tag AUX?
- Do the word forms tagged with a closed-class POS tag like PUNCT or DET make sense?

Partially implemented in out-of-the-box tools :)

github.com/universaldependencies/tools
udapi.github.io (Popel+ 2017)

Today

MaiBaam (Bavarian treebank)

Considerations: data, preprocessing, annotations

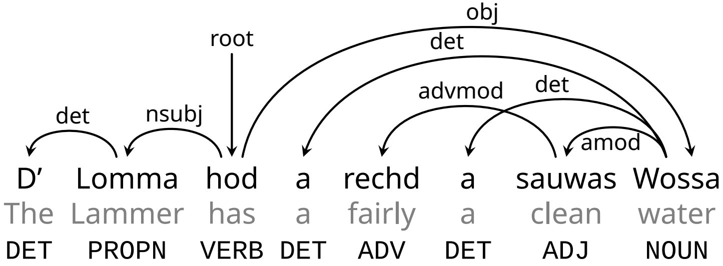
Tools

Training/evaluating ML models

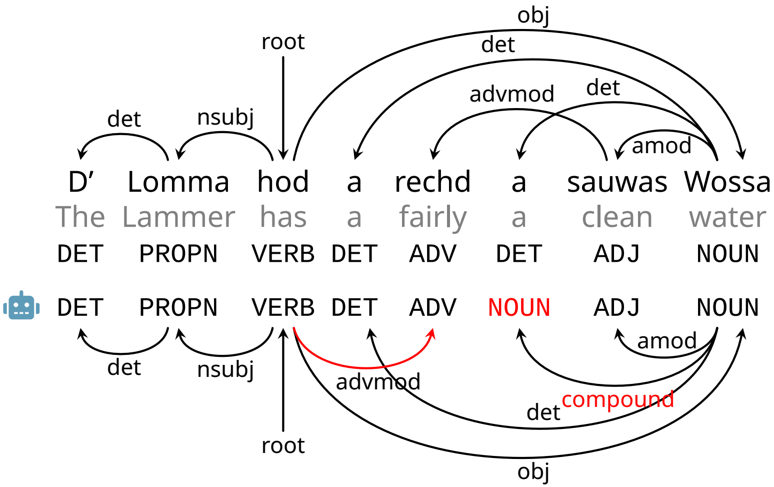
Modelling

- Enough data for train/dev/test splits or test-only?
- Transfer learning: train on another (ideally closely related) language, test on actual target language
- Baselines vs. engineering the Best Bavarian Parser Ever

Metrics


obj
root
det
advmod
det
det
nsubj
amod
D’ Lomma hod a rechd a sauwas Wossa
The Lammer has a fairly a clean water
DET PROPN VERB DET ADV DET ADJ NOUN

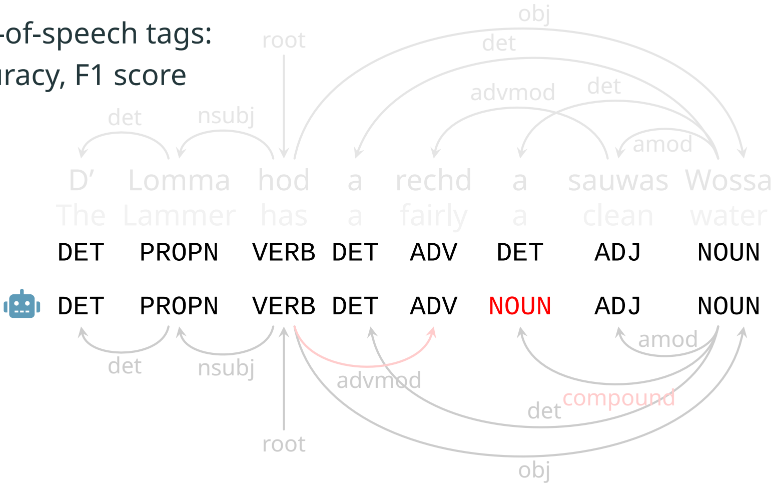
Metrics



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

Metrics

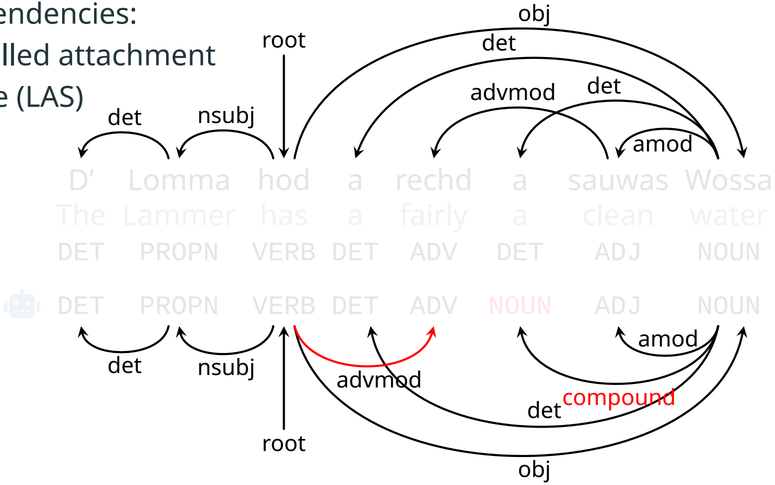
Part-of-speech tags:
Accuracy, F1 score



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

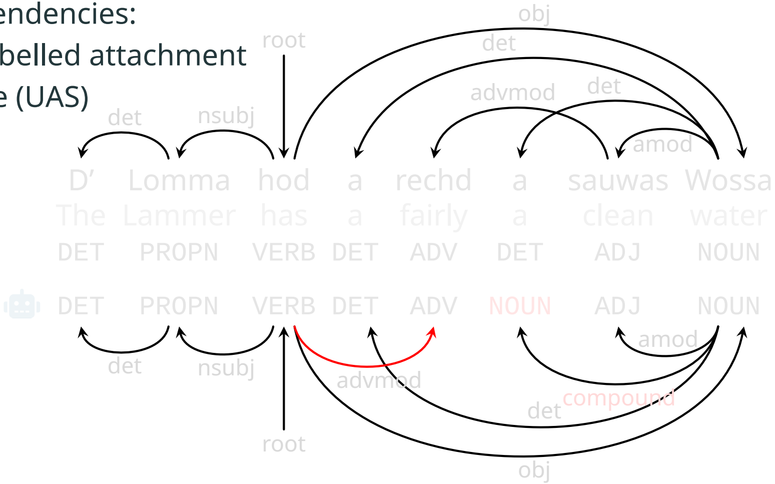
Metrics

Dependencies:
Labelled attachment score (LAS)


obj
root
det
advmod
det
det
nsubj
amod
D’ Lomma hod a rechd a sauwas Wossa
The Lammer has a fairly a clean water
DET PROPN VERB DET ADV DET ADJ NOUN
DET PROPN VERB DET ADV NOUN ADJ NOUN
amod
det
nsubj
advmod
det
compound
root
obj

Metrics

Dependencies:
Unlabelled attachment score (UAS)


obj
det
root
advmod
det
det
nsubj
amod
D' Lomma hod a rechd a sauwas Wossa
The Lammer has a fairly a clean water
DET PROPN VERB DET ADV DET ADJ NOUN
DET PROPN VERB DET ADV NOUN ADJ NOUN
amod
det
nsubj
advmod
compound
det
root
obj

Experiments

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

Out-of-the-box models

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

Own models

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

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

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	42.3	24.9
GBERT	BAR	58.9	36.4
UDPipe	BAR	80.3	65.8

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

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	42.3	24.9	Full words
GBERT	BAR	58.9	36.4	Subword tokens
UDPipe	BAR	80.3	65.8	Subword tok. + characters

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

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`

Automatic pre-annotation

Model	Test lang	Acc (%)	LAS (%)
UDPipe	BAR	80.3	65.8

Workable for POS tag pre-annotation?

How can we mitigate getting biased by our automatic pre-annotation?

Automatic pre-annotation II

Once you have already annotated some data (preliminary test set):

- Take 2 UDPipe parsers trained on different(!!) German treebanks
- Given the same input text, which tokens do they have the *same* predictions for?
- And which of those joint predictions involve POS tags predicted with >95% precision on the preliminary test data?

“Anchoring and agreement in syntactic annotations” (Berzak+, EMNLP 2016)

The end!

More resources

- UD guidelines universaldependencies.org/guidelines.html
- UD introduction (webinar) unidive.lisn.upsaclay.fr/doku.php?id=other-events:webinar-1

Term project and/or thesis topics ideas

- New UD/POS annotations (manual, automatic)
- New annotation layers for existing treebanks
- How to detect errors/inconsistencies?
- Better parsers

Questions?