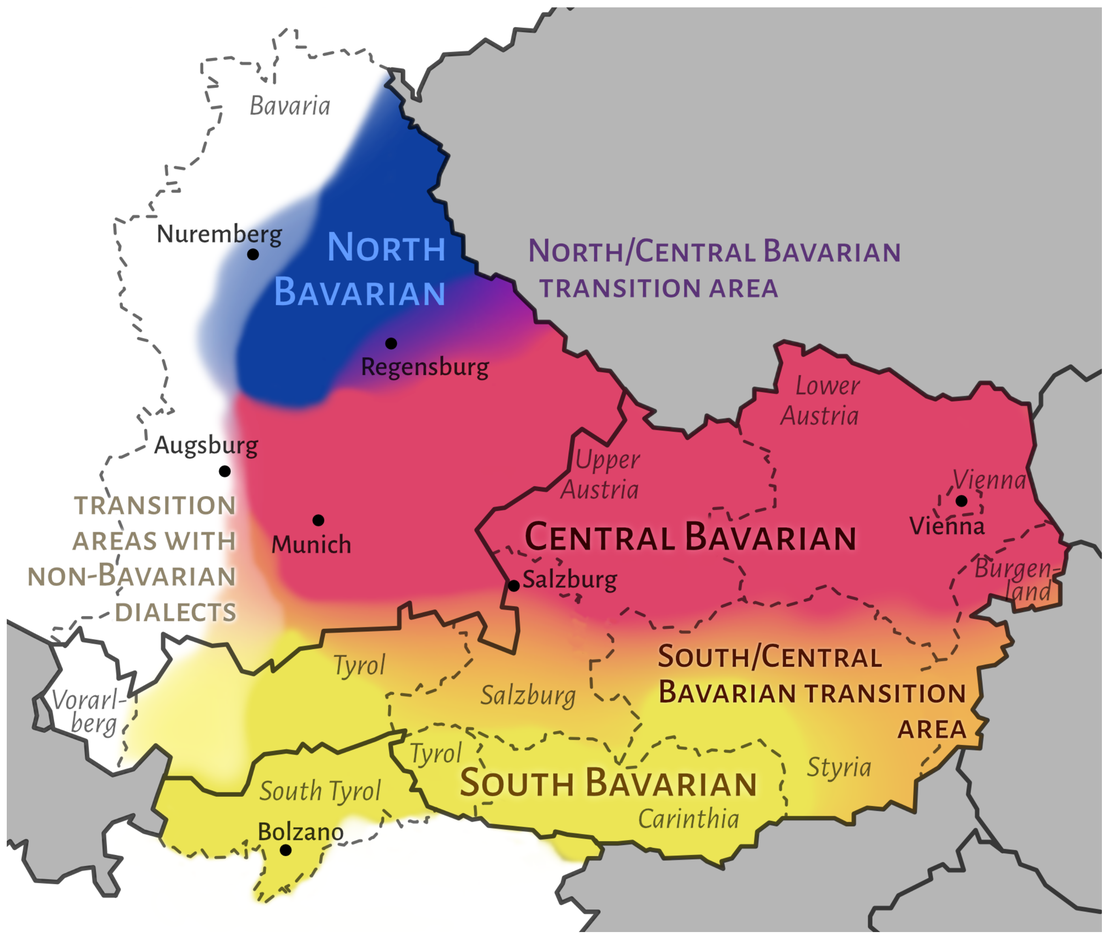
MaiBaam – A multi-dialectal Bavarian Universal Dependency treebank

Verena Blaschke, Barbara Kovačić, Siyao (Logan) Peng,
Hinrich Schütze, Barbara Plank

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MCML Munich Center for Machine Learning



- Morphosyntactic & orthographic **variation**
- Sentence-level **dialect & genre metadata**
- **Manually annotated** POS tags & dependencies
- **15k tokens**, 1k sentences
- Now on UD: **UD_Bavarian_MaiBaam**

Data, paper & annotation guidelines →

Bavarian

- Spoken by 10M+ people in DE, AT, IT
- No orthography
- 3 main dialect groups (+ transition areas)
- Closely related to German, yet morphosyntactic differences

● I mecht wissn, wej lång das'sd nu brauchst.

● I mechad gern wissn, wia lang du no brauchst.

● I mecht gearn wissn, wia long du no brausch.

DE Ich möchte wissen, wie lange du noch brauchst.

EN I want to know how long you'll still take.

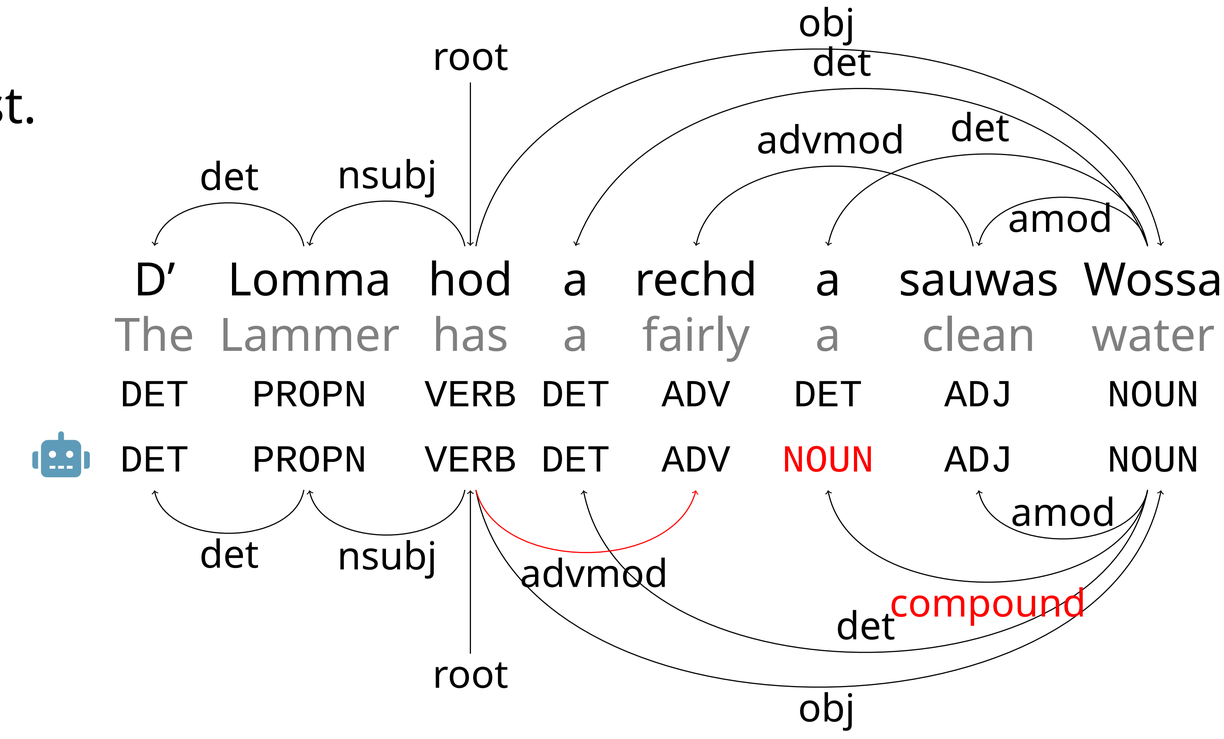
Treebank stats

Dialect group	Toks
● North	833
● North/Central	793
● Central	3 303
● South/Central	1 130
● South	995
? Underspecified	7 969
Total	15 023

Genre	Toks
Wiki	7 988
Grammar ex.	2 485
Non-fiction	2 019
Social	1 599
Fiction	932

Parsing baselines

- Best system: UDPipe–GSD
  80.3% POS accuray, 65.8% LAS
  → Ample room for improvement!
  Can you beat it?
- Input representations seem to be an important factor in parsing/tagging performance
  (full words vs. subwords vs. characters)
- More experiments in paper



Gold (top) vs.
prediction by best system (bottom)
"The Lammer (river) has fairly clean water"

This research is supported by the ERC Consolidator Grant DIALECT 101043235.