

# Cross-lingual Syntactic Parsing

# Cross-lingual Syntactic Parsing

- e.g.: parse Lower Sorbian

# Cross-lingual Syntactic Parsing

- e.g.: parse Lower Sorbian
  - no training data

# Cross-lingual Syntactic Parsing

- e.g.: parse Lower Sorbian
  - no training data
  - use training data for other close languages
    - Upper Sorbian, Polish, Czech

# Cross-lingual Syntactic Parsing

- e.g.: parse Lower Sorbian
  - no training data
  - use training data for other close languages
    - Upper Sorbian, Polish, Czech
- our focus: dependency parsing, POS tagging
- probably adaptable to other NLP tasks
  - no/low annotated data for the “target” language
  - annotated data for some (close) “source” languages

# Methods, subtasks, problems

# Methods, subtasks, problems

- strong multilingualism
  - massively multiparallel corpora (Bible, Watchtower)
  - tens or hundreds of languages

# Methods, subtasks, problems

- strong multilingualism
  - massively multiparallel corpora (Bible, Watchtower)
  - tens or hundreds of languages
- word alignment, machine translation
  - low data
  - special needs (very literal, monotone, 1:1...)



# Methods, subtasks, problems

- strong multilingualism
  - massively multiparallel corpora (Bible, Watchtower)
  - tens or hundreds of languages
- word alignment, machine translation
  - low data
  - special needs (very literal, monotone, 1:1...)
- cross-lingually consistent annotation (UD)

# Methods, subtasks, problems

- strong multilingualism
  - massively multiparallel corpora (Bible, Watchtower)
  - tens or hundreds of languages
- word alignment, machine translation
  - low data
  - special needs (very literal, monotone, 1:1...)
- cross-lingually consistent annotation (UD)
- tokenization, encoding, transliteration...

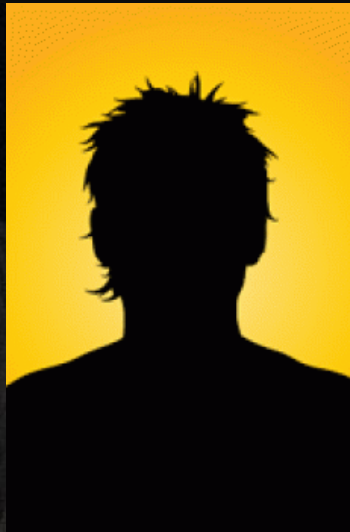
# At ÚFAL

- The “team”

- Dan Zeman

- Zdeněk Žabokrtský, David Mareček, Rudolf Rosa

- Loganathan, Adedayo Oluokun, Vinit Ravishankar...



# At ÚFAL

- The “team”
  - Dan Zeman
  - Zdeněk Žabokrtský, David Mareček, Rudolf Rosa
  - Loganathan, Adedayo Oluokun, Vinit Ravishankar...
- Best in the world (at least in shared tasks)
  - 2017 VarDial: 1<sup>st</sup> place
  - 2018 CoNLL: 1<sup>st</sup> place (for low-resource languages)



# At ÚFAL



- The “team”
  - Dan Zeman
  - Zdeněk Žabokrtský, David Mareček, Rudolf Rosa
  - Loganathan, Adedayo Oluokun, Vinit Ravishankar...
- Best in the world (at least in shared tasks)
  - 2017 VarDial: 1<sup>st</sup> place
  - 2018 CoNLL: 1<sup>st</sup> place (for low-resource languages)
- NPFL120 Multilingual NLP (Dan, Rudolf, oBo)
  - summer semester, everyone welcome!