# Improving UD processing via satellite resources for morphology

Kaja Dobrovoljc Tomaž Erjavec Nikola Ljubešić

Jozef Stefan Institute
Ljubljana, Slovenia

cjvt Centre for Language Resources and Technologies

UDW 2019, Paris, August 30

# Motivation

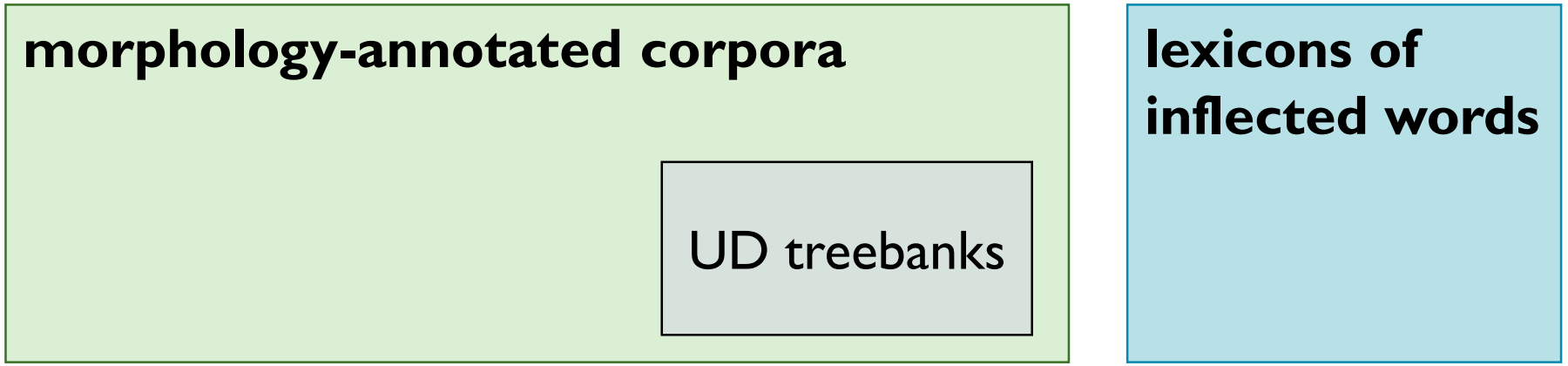
- many treebanks and tools available for UD-based NLP tasks
  - state-of-the-art results for dependency parsing and lower layers
- UD treebanks require both morphological and syntactic annotation (costly)
- UD tools could benefit from other existing language resources, as well
  - esp. for lemmatization, POS tagging, feature prediction of languages with complex morphology
- e.g. language resources available for Croatian and Slovenian morphology

# Motivation

- many treebanks and tools available for UD-based NLP tasks
  - state-of-the-art results for dependency parsing and lower layers
- UD treebanks require both morphological and syntactic annotation (costly)
- UD tools could benefit from other existing language resources, as well
  - esp. for lemmatization, POS tagging, feature prediction of languages with complex morphology
- e.g. language resources available for Croatian and Slovenian morphology

# Motivation

- many treebanks and tools available for UD-based NLP tasks
  - state-of-the-art results for dependency parsing and lower layers
- UD treebanks require both morphological and syntactic annotation (costly)
- UD tools could benefit from other existing language resources, as well
  - esp. for lemmatization, POS tagging, feature prediction of languages with complex morphology
- e.g. language resources available for Croatian and Slovenian morphology



# Our goal

- present the conversion of existing morphology resources to UD scheme
- explore their contribution to UD processing on different linguistic levels

**Talk outline**

1. Conversion of the resources
2. Experiments
3. Results
4. Conclusions

# CONVERSION OF THE RESOURCES

# Language resources for morphology

- two reference training corpora with morphology annotations
  - ssj500k (Krek et al. 2013-) for Slovenian
  - hr500k (Ljubešić et al. 2018-) for Croatian

- two reference lexicons of inflected forms
  - Sloleks morphological lexicon (Dobrovoljc et al. 2013-) for Slovenian
  - hrLex inflection lexicon (Ljubešić 2016-) for Croatian

- similar, but developed within different projects
  - including the conversion to UD

# Morphology-annotated corpora

| ssj500k | hr500k |
| --- | --- |
| • The largest manually annotated corpus of Slovenian (~580,000 tokens)<br>• Fully annotated for segmentation, lemmatization, morphosyntax (JOS/MULTEXT-East).<br>• Partially annotated for named entities, JOS dependency syntax, semantic roles, multi-word expressions, Universal Dependencies.<br>• Rule-based conversion to UD (Dobrovoljc et al. 2015) → **25% of the corpus** | • The largest manually annotated corpus of Croatian (~500,000 tokens).<br>• Fully annotated for segmentation, lemmatization, morphosyntax (MULTEXT-East), named entities.<br>• Partially annotated for semantic roles, Universal Dependencies.<br>• Rule-based conversion to UD morphology, manual annotation of UD syntax (Agić and Ljubešić 2015) → **40% of the corpus** |

# Morphology-annotated corpora

## ssj500k

- Context-independent JOS-to-UD conversion rules for morphology.
- Exception 1: list of DET (lexicon-based)
- Exception 2: *biti* as AUX/VERB (syntax-based)
- Released as part of ssj500k 2.2, in CONLL-U and TEI XML

CLARIN.SI



## hr500k

- Context-independent MTE-to-UD conversion of morphology.
- Exception: abbreviations
- Released as part of hr500k 1.0, in CONLL-U and TEI XML

CLARIN.SI



MTE

**`Numeral, Form=letter, Type=ordinal`**

e.g. *prvi* 'first', *drugi* 'second', *tretji* 'third' …

UD

**`ADJ, NumType=Ord`**

# Lexicons of inflected forms

## Sloleks

- The largest manually compiled collection of inflected forms for Slovenian (~2.7M forms, 100k lemmas).
- Additional information on lemma, grammatical features (JOS/MTE), pronunciation, frequency of usage.
- Conversion using the JOS-to-UD mappings from ssj500k.
- Lexicon with UPOS and FEATS released as part of Sloleks 2.0 (CLARIN.SI), tab-separated list only.

CLARIN.SI



## hrLex

- The largest semi-automatically compiled collection of inflected forms for Croatian (~6.4M forms, 170k lemmas).
- Additional information on lemma, grammatical features (MTE), frequency of usage.
- Conversion using the MTE-to-UD mappings from hr500k.
- Lexicon with UPOS and FEATS released as part of hrLex 1.3 (CLARIN.SI), tab-separated list.

CLARIN.SI

# EXPERIMENTS

# Tool

- StanfordNLP tool (Qi et al. 2018)
  - full neural network pipeline for robust text analytics on various levels
  - https://stanfordnlp.github.io/stanfordnlp/
- one of the best-performing systems in CoNLL Shared Task 2017-18
  - top-three for all metrics for Slovenian and Croatian
- pipeline architecture
  - morphological tagging + features → lemmatization → parsing

# Experiment setup

1. extended training corpora for morphology
   - training on the official UD (baseline) vs. training on full ssj500k/hr500k

   →tagging + lemmatization + parsing

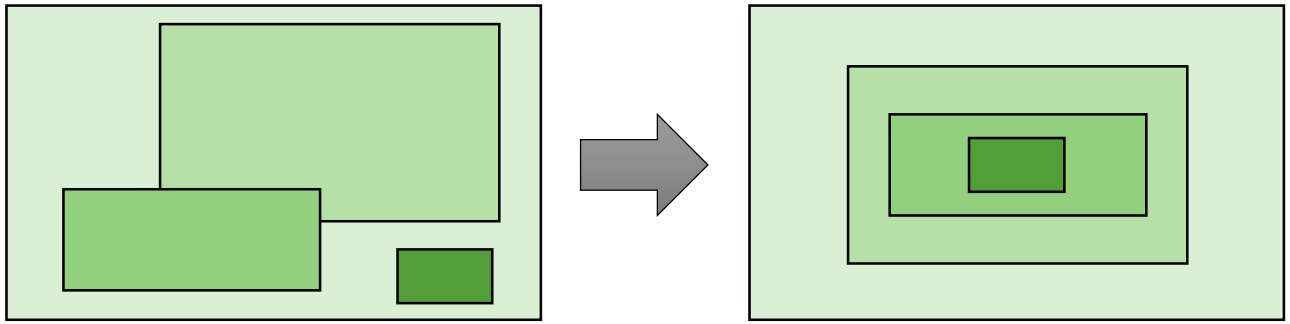
2. lexicon lookup for lemmatization
   - looking up training data lexicon (baseline) vs. looking up Sloleks/hrLex

   → lemmatization + parsing

- gold segmentation
- CoNLL 2018 evaluation script

# Data split

- *babushka-bench*: a benchmarking platform for South Slavic languages
  - https://github.com/clarinsi/babushka-bench
- a universal split for variously-sized subsets of the same dataset
- no spillage between train, dev or test for different annotation layers



- different to official UD splits (but comparable)

# Data split

| | sl-UD | ssj500k | hr-UD | hr500k |
|---|---|---|---|---|
| train | 110,711 | 474,322 | 165,989 | 415,328 |
| dev | 16,589 | 62,967 | 14,184 | 39,765 |
| test | 13,370 | 48,959 | 16,855 | 51,364 |
| **Total** | **140,670** | **586,248** | **197,028** | **506,457** |

# RESULTS

# 1. Larger training sets for UD morphology

| | sl-UD | ssj500k | hr-UD | hr500k |
|---|---|---|---|---|
| LEMMA | 95.88 | **97.44** | 95.30 | **96.21** |
| UPOS | 98.45 | **98.69** | 97.91 | **98.05** |
| XPOS | 95.65 | **97.00** | 94.60 | **95.12** |
| FEATS | 95.95 | **97.23** | 95.13 | **95.66** |
| UAS | 93.40 | **93.72** | 90.22 | **90.76** |
| LAS | 91.62 | **92.28** | 85.30 | **86.00** |
| MLAS | 84.24 | **86.22** | 75.54 | **76.88** |
| BLEX | 84.04 | **86.96** | 76.45 | **78.56** |

# 1. Larger training sets for UD morphology

| | sl-UD | ssj500k | | hr-UD | hr500k | |
|---|---|---|---|---|---|---|
| LEMMA | 95.88 | **97.44** | +1.56 | 95.30 | **96.21** | +0.91 |
| UPOS | 98.45 | **98.69** | | 97.91 | **98.05** | |
| XPOS | 95.65 | **97.00** | | 94.60 | **95.12** | |
| FEATS | 95.95 | **97.23** | | 95.13 | **95.66** | |
| UAS | 93.40 | **93.72** | | 90.22 | **90.76** | |
| LAS | 91.62 | **92.28** | | 85.30 | **86.00** | |
| MLAS | 84.24 | **86.22** | | 75.54 | **76.88** | |
| BLEX | 84.04 | **86.96** | | 76.45 | **78.56** | |

# 1. Larger training sets for UD morphology

| | sl-UD | ssj500k | | hr-UD | hr500k | |
|---|---|---|---|---|---|---|
| LEMMA | 95.88 | **97.44** | | 95.30 | **96.21** | |
| UPOS | 98.45 | **98.69** | | 97.91 | **98.05** | |
| XPOS | 95.65 | **97.00** | +1.35 | 94.60 | **95.12** | +0.53 |
| FEATS | 95.95 | **97.23** | +1.28 | 95.13 | **95.66** | +0.52 |
| UAS | 93.40 | **93.72** | | 90.22 | **90.76** | |
| LAS | 91.62 | **92.28** | | 85.30 | **86.00** | |
| MLAS | 84.24 | **86.22** | | 75.54 | **76.88** | |
| BLEX | 84.04 | **86.96** | | 76.45 | **78.56** | |

# 1. Larger training sets for UD morphology

| | sl-UD | ssj500k | | hr-UD | hr500k | |
|---|---|---|---|---|---|---|
| LEMMA | 95.88 | **97.44** | | 95.30 | **96.21** | |
| UPOS | 98.45 | **98.69** | +0.24 | 97.91 | **98.05** | +0.14 |
| XPOS | 95.65 | **97.00** | | 94.60 | **95.12** | |
| FEATS | 95.95 | **97.23** | | 95.13 | **95.66** | |
| UAS | 93.40 | **93.72** | | 90.22 | **90.76** | |
| LAS | 91.62 | **92.28** | | 85.30 | **86.00** | |
| MLAS | 84.24 | **86.22** | | 75.54 | **76.88** | |
| BLEX | 84.04 | **86.96** | | 76.45 | **78.56** | |

# 1. Larger training sets for UD morphology

| | sl-UD | ssj500k | | hr-UD | hr500k | |
|---|---|---|---|---|---|---|
| LEMMA | 95.88 | **97.44** | | 95.30 | **96.21** | |
| UPOS | 98.45 | **98.69** | | 97.91 | **98.05** | |
| XPOS | 95.65 | **97.00** | | 94.60 | **95.12** | |
| FEATS | 95.95 | **97.23** | | 95.13 | **95.66** | |
| UAS | 93.40 | **93.72** | +0.32 | 90.22 | **90.76** | +0.54 |
| LAS | 91.62 | **92.28** | +0.66 | 85.30 | **86.00** | +0.70 |
| MLAS | 84.24 | **86.22** | +1.98 | 75.54 | **76.88** | +1.34 |
| BLEX | 84.04 | **86.96** | +2.92 | 76.45 | **78.56** | +2.11 |

# 2. Larger lexicons of inflected forms

| | sl-UD | +Sloleks | ssj500k | +Sloleks |
|---|---|---|---|---|
| LEMMA | 95.88 | **98.48** | 97.44 | **98.89** |
| UAS | 93.40 | **93.43** | 93.72 | 93.72 |
| LAS | 91.62 | **91.75** | 92.28 | 92.27 |
| MLAS | 84.24 | **84.34** | 86.22 | 86.05 |
| BLEX | 84.04 | **88.00** | 86.96 | **89.01** |

| | hr-UD | +hrLex | hr500k | +hrLex |
|---|---|---|---|---|
| LEMMA | 95.30 | **97.24** | 96.21 | **97.29** |
| UAS | 90.22 | **90.53** | 90.76 | 90.44 |
| LAS | 85.30 | **85.81** | 86.00 | 85.85 |
| MLAS | 75.54 | **76.16** | 76.88 | 76.83 |
| BLEX | 76.45 | **79.60** | 78.56 | **80.04** |

# 2. Larger lexicons of inflected forms

| | sl-UD | +Sloleks | | ssj500k | +Sloleks |
|---|---|---|---|---|---|
| LEMMA | 95.88 | **98.48** | +2.6 | 97.44 | **98.89** |
| UAS | 93.40 | **93.43** | | 93.72 | 93.72 |
| LAS | 91.62 | **91.75** | | 92.28 | 92.27 |
| MLAS | 84.24 | **84.34** | | 86.22 | 86.05 |
| BLEX | 84.04 | **88.00** | | 86.96 | **89.01** |

| | hr-UD | +hrLex | | hr500k | +hrLex |
|---|---|---|---|---|---|
| LEMMA | 95.30 | **97.24** | +1.94 | 96.21 | **97.29** |
| UAS | 90.22 | **90.53** | | 90.76 | 90.44 |
| LAS | 85.30 | **85.81** | | 86.00 | 85.85 |
| MLAS | 75.54 | **76.16** | | 76.88 | 76.83 |
| BLEX | 76.45 | **79.60** | | 78.56 | **80.04** |

# 2. Larger lexicons of inflected forms

| | sl-UD | +Sloleks | ssj500k | +Sloleks | |
|---|---|---|---|---|---|
| LEMMA | 95.88 | **98.48** | 97.44 | **98.89** | +1.45 |
| UAS | 93.40 | **93.43** | 93.72 | 93.72 | |
| LAS | 91.62 | **91.75** | 92.28 | 92.27 | |
| MLAS | 84.24 | **84.34** | 86.22 | 86.05 | |
| BLEX | 84.04 | **88.00** | 86.96 | **89.01** | |

| | hr-UD | +hrLex | hr500k | +hrLex | |
|---|---|---|---|---|---|
| LEMMA | 95.30 | **97.24** | 96.21 | **97.29** | +1.08 |
| UAS | 90.22 | **90.53** | 90.76 | 90.44 | |
| LAS | 85.30 | **85.81** | 86.00 | 85.85 | |
| MLAS | 75.54 | **76.16** | 76.88 | 76.83 | |
| BLEX | 76.45 | **79.60** | 78.56 | **80.04** | |

# 2. Larger lexicons of inflected forms

| | sl-UD | +Sloleks | | ssj500k | +Sloleks |
|---|---|---|---|---|---|
| LEMMA | 95.88 | **98.48** | | 97.44 | **98.89** |
| UAS | 93.40 | **93.43** | +0.03 | 93.72 | 93.72 |
| LAS | 91.62 | **91.75** | +0.13 | 92.28 | 92.27 |
| MLAS | 84.24 | **84.34** | +0.10 | 86.22 | 86.05 |
| BLEX | 84.04 | **88.00** | +3.96 | 86.96 | **89.01** |

| | hr-UD | +hrLex | | hr500k | +hrLex |
|---|---|---|---|---|---|
| LEMMA | 95.30 | **97.24** | | 96.21 | **97.29** |
| UAS | 90.22 | **90.53** | +0.29 | 90.76 | 90.44 |
| LAS | 85.30 | **85.81** | +0.51 | 86.00 | 85.85 |
| MLAS | 75.54 | **76.16** | +0.72 | 76.88 | 76.83 |
| BLEX | 76.45 | **79.60** | +3.15 | 78.56 | **80.04** |

# 2. Larger lexicons of inflected forms

| | sl-UD | +Sloleks | ssj500k | +Sloleks | |
|---|---|---|---|---|---|
| LEMMA | 95.88 | **98.48** | 97.44 | **98.89** | |
| UAS | 93.40 | **93.43** | 93.72 | 93.72 | +0.0 |
| LAS | 91.62 | **91.75** | 92.28 | 92.27 | -0.01 |
| MLAS | 84.24 | **84.34** | 86.22 | 86.05 | -0.17 |
| BLEX | 84.04 | **88.00** | 86.96 | **89.01** | +2.05 |

| | hr-UD | +hrLex | hr500k | +hrLex | |
|---|---|---|---|---|---|
| LEMMA | 95.30 | **97.24** | 96.21 | **97.29** | |
| UAS | 90.22 | **90.53** | 90.76 | 90.44 | -0.32 |
| LAS | 85.30 | **85.81** | 86.00 | 85.85 | -0.15 |
| MLAS | 75.54 | **76.16** | 76.88 | 76.83 | -0.05 |
| BLEX | 76.45 | **79.60** | 78.56 | **80.04** | +1.45 |

# Conclusions

- four new open-source language resources conformant with UD morphological specifications for Croatian and Slovenian

# Conclusions

- four new open-source language resources conformant with UD morphological specifications for Croatian and Slovenian
- improvements of baseline StanfordNLP performance obtained on official released UD treebanks

# Conclusions

- four new open-source language resources conformant with UD morphological specifications for Croatian and Slovenian
- improvements of baseline StanfordNLP performance obtained on official released UD treebanks
  - esp. for lemmatization and morphological feature prediction
  - retrained models already in use

# Conclusions

- four new open-source language resources conformant with UD morphological specifications for Croatian and Slovenian
- improvements of baseline StanfordNLP performance obtained on official released UD treebanks
  - esp. for lemmatization and morphological feature prediction
  - retrained models already in use
- such satellite UD resources are likely to exist or emerge in many other languages
  - a need for a standardized format and distribution?
  - system architectures that enable lexicon plug-ins

*Kaja Dobrovoljc, Tomaž Erjavec, Nikola Ljubešić*

Improving UD processing via satellite resources for morphology

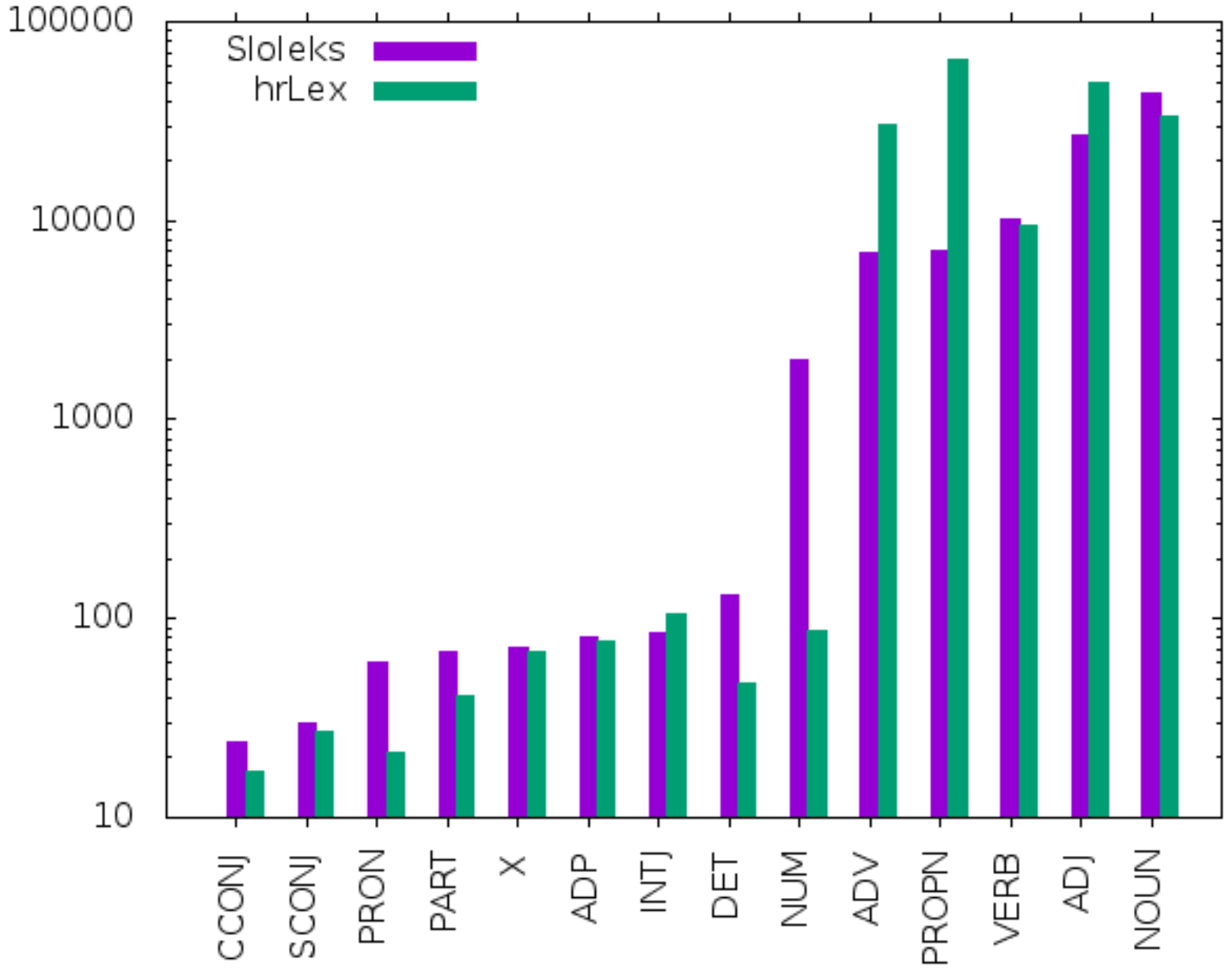
# THANK YOU

code available at
https://github.com/clarinsi/jos2ud
https://github.com/clarinsi/babushka-bench

# Overview

| | Tokens | Types | Wforms | Lemmas | Tags | Ambiguity |
|---|---|---|---|---|---|---|
| ssj500k | 586,248 | 98,641 | 78,707 | 38,818 | 1,304 | 1.25 |
| hr500k | 506,457 | 84,789 | 66,797 | 34,321 | 768 | 1.27 |

| | Entries | Wforms | Lemmas | Tags | Ambiguity |
|---|---|---|---|---|---|
| Sloleks | 2,792,003 | 921,869 | 96,593 | 1,900 | 3.03 |
| hrLex | 6,427,709 | 1,697,943 | 164,206 | 900 | 3.79 |


100000
10000
1000
100
10
Sloleks
hrLex
CCONJ
SCONJ
PRON
PART
X
ADP
INTJ
DET
NUM
ADV
PROPN
VERB
ADJ
NOUN