Introduction to Natural Language Processing

a course taught as B4M36NLP at Open Informatics

by members of the Institute of Formal and Applied Linguistics

Today: Week 6, lab
Today’s topic: Universal Dependencies
Today’s teacher: Daniel Zeman

E-mail: zeman@ufal.mff.cuni.cz
WWW: http://ufal.mff.cuni.cz/daniel-zeman
Dependency Treebanks

Dependency Treebank for Slovenian language

A-tree
zone=sl

Označuje
root
Ggnste

Punct
PUNCT

Besednjaku
štiri
Somem

V

dol
Dm

Normalnem
dol
Ppnmem

Se
del
Zp------k

Diskriminacija
ena
Sozei

Privilegij
tri
Somei

Pozitivna
dol
Ppnzei

Kot
vez
Vd
Dependency Treebanks

V normalnem besednjaku se pozitivna diskriminacija označuje kot privilegij
In normal dictionary is positive discrimination referred-to as privilege
Why?

- Linguistic research
  - Corpus query

- Training tools (parsers) for NLP
  - Downstream applications
My daughter bought some bread and cheese.
Universal Dependencies

http://universaldependencies.org/
Universal Dependencies

http://universaldependencies.org/
Universal Dependencies

http://universaldependencies.org/
Universal Dependencies

http://universaldependencies.org/
Universal Dependencies

http://universaldependencies.org/

Stanford Dependencies
CLEAR
Google UD
Stanford UD
Universal Dependencies

http://universaldependencies.org/

Stanford Dependencies
CLEAR
Google UD
Stanford UD
HamleDT
Universal Dependencies

http://universaldependencies.org/
Universal Dependencies

http://universaldependencies.org/
Universal Dependencies

http://universaldependencies.org/
Universal Dependencies

http://universaldependencies.org/

• Milestones:
  – 2014-04: EACL Göteborg, kick-off meeting
  – 2014-10: UD guidelines version 1
  – 2015-01: released 10 treebanks of 10 languages (UD 1.0)
  – 2015-05: released 19 treebanks of 18 languages (UD 1.1)
  – 2015-11: released 37 treebanks of 33 languages (UD 1.2)
  – 2016-05: released 54 treebanks of 40 languages (UD 1.3)
  – 2016-11: UD release 1.4, ~7 new languages
  – 2016 fall: UD guidelines version 2
Goals and Requirements

• Cross-linguistically consistent grammatical annotation
Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de facto standards
Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de facto standards

Caveats:
- Not a new linguistic theory – but linguistically informed and relevant
Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de facto standards

Caveats:
- Not a new linguistic theory – but linguistically informed and relevant
- Not an ideal parsing representation – but useful for comparative evaluation
Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de facto standards

Caveats:
- Not a new linguistic theory – but linguistically informed and relevant
- Not an ideal parsing representation – but useful for comparative evaluation
- Not the ultimate annotation scheme – but a lightweight lingua franca
Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de facto standards

Caveats:
- Not a new linguistic theory – but linguistically informed and relevant
- Not an ideal parsing representation – but useful for comparative evaluation
- Not the ultimate annotation scheme – but a lightweight *lingua franca*

*Not “Universal” in the strictly typological sense!*
Design Principles

- Dependency
  - Widely used in practical NLP systems
  - Available in treebanks for many languages
Design Principles

• Dependency
  – Widely used in practical NLP systems
  – Available in treebanks for many languages

• Lexicalism
  – Basic annotation units are words – syntactic words
  – Words have morphological properties
  – Words enter into syntactic relations
Design Principles

• Dependency
  – Widely used in practical NLP systems
  – Available in treebanks for many languages

• Lexicalism
  – Basic annotation units are words – syntactic words
  – Words have morphological properties
  – Words enter into syntactic relations

• Recoverability
  – Transparent mapping from input text to word segmentation
Golden Rules

- Maximize parallelism
  - Don’t annotate the same thing in different ways
  - Don’t make different things look the same
Golden Rules

- Maximize parallelism
  - Don’t annotate the same thing in different ways
  - Don’t make different things look the same

- But don’t overdo it
  - Don’t annotate things that are not there
  - Balance: is it still the same thing?
  - Allow language-specific extensions
Morphology

Některé dívky si nicméně pochvalovaly zmrzlinu.

29.9.2016, Ljubljana
Morphology

Některé dívky si nicméně pochvalovaly zmrzlinu.

některý dívka se nicméně pochvalovat zmrzlina.

• Lemma representing the semantic content of the word
Některé dívky si nicméně pochvalovaly zmrzlinu.

Pořadová číslo

PronType=Ind
gender=Fem
Number=Plur
Case=Nom

• Lemma representing the semantic content of the word

• Part-of-speech tag representing the abstract lexical category associated with the word
Morphology

Některé dívky si nicméně pochvalovaly zmrzlinu.

některý dívka se nicméně pochvaloval zmrzlina.

- Lemma representing the semantic content of the word
- Part-of-speech tag representing the abstract lexical category associated with the word
- Features representing lexical and grammatical properties associated with the lemma or the particular word form
### Part-of-Speech Tags

<table>
<thead>
<tr>
<th>Open</th>
<th>Closed</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>ADP</td>
<td>PUNCT</td>
</tr>
<tr>
<td>ADV</td>
<td>AUX</td>
<td>SYM</td>
</tr>
<tr>
<td>INTJ</td>
<td>CONJ</td>
<td>X</td>
</tr>
<tr>
<td>NOUN</td>
<td>DET</td>
<td></td>
</tr>
<tr>
<td>PROPN</td>
<td>NUM</td>
<td></td>
</tr>
<tr>
<td>VERB</td>
<td>PART</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRON</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCONJ</td>
<td></td>
</tr>
</tbody>
</table>

- Taxonomy of 17 universal part-of-speech tags, based on the Google Universal Tagset (Petrov et al., 2012)
- All languages use the same inventory, but not all tags have to be used by all languages
Features

<table>
<thead>
<tr>
<th>Lexical</th>
<th>Inflectional / Nominal</th>
<th>Inflectional / Verbal</th>
</tr>
</thead>
<tbody>
<tr>
<td>PronType</td>
<td>Gender</td>
<td>VerbForm</td>
</tr>
<tr>
<td>NumType</td>
<td>Animacy</td>
<td>Mood</td>
</tr>
<tr>
<td>Poss</td>
<td>Number</td>
<td>Tense</td>
</tr>
<tr>
<td>Reflex</td>
<td>Case</td>
<td>Aspect</td>
</tr>
<tr>
<td></td>
<td>Definite</td>
<td>Voice</td>
</tr>
<tr>
<td></td>
<td>Degree</td>
<td>Person</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Negative</td>
</tr>
</tbody>
</table>

- Standardized inventory of morphological features, based on Interset (Zeman, 2008)
- Languages select relevant features and can add language-specific features or values with documentation
The cat could have chased all the dogs down the street.

DET NOUN AUX AUX VERB DET DET NOUN ADP DET NOUN PUNCT
Syntax

- Content words are related by dependency relations

The cat could have chased all the dogs down the street.
Syntax

- Content words are related by dependency relations
- Function words attach to closest content word
Syntax

- Content words are related by dependency relations
- Function words attach to closest content word
- Punctuation attach to head of phrase or clause
Syntax

- Content words are related by dependency relations.
- Function words attach to closest content word.
- Punctuation attaches to head of phrase or clause.

Not "dependency" in the strictly syntactic sense!
The dog was chased by the cat.

Definite=Def, Tense=Past, Definite=Def
Dependency Relations

- Taxonomy of 40 universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
  - Language-specific subtypes may be added
Dependency Relations

- Taxonomy of 40 universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
  - Language-specific subtypes may be added

- Organizing principles
  - Three types of structures: nominals, clauses, modifiers
  - Core arguments vs. other dependents (not arguments vs. adjuncts)
## Dependents of Clausal Predicates

<table>
<thead>
<tr>
<th></th>
<th>Nominal</th>
<th>Clausal</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>nsubj</td>
<td>csubj</td>
<td>advmod</td>
</tr>
<tr>
<td></td>
<td>nsubjpass</td>
<td>csubjpass</td>
<td>neg</td>
</tr>
<tr>
<td></td>
<td>dobj</td>
<td>ccomp</td>
<td>aux</td>
</tr>
<tr>
<td></td>
<td>iobj</td>
<td>xcomp</td>
<td>auxpass</td>
</tr>
<tr>
<td>Non-Core</td>
<td>nmod</td>
<td>advcl</td>
<td>cop</td>
</tr>
<tr>
<td></td>
<td>vocative</td>
<td></td>
<td>mark</td>
</tr>
<tr>
<td></td>
<td>discourse</td>
<td></td>
<td>punct</td>
</tr>
<tr>
<td></td>
<td>expl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dependents of Nominals

<table>
<thead>
<tr>
<th>Nominal</th>
<th>Clausal</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>nmod</td>
<td>acl</td>
<td>amod</td>
</tr>
<tr>
<td>appos</td>
<td></td>
<td>det</td>
</tr>
<tr>
<td>nummod</td>
<td></td>
<td>neg</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case</td>
</tr>
</tbody>
</table>

Ljubljana, the lovely capital of Slovenia
“Stanford-style” Coordination

- Coordinate structures are headed by the first conjunct
  - Subsequent conjuncts depend on it via the conj relation
  - Conjunctions depend on it via the cc relation
  - Punctuation marks depend on it via the punct relation
**Multiword Expressions**

<table>
<thead>
<tr>
<th>Relation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>mwe</td>
<td><em>in spite of</em>, <em>as well as</em>, <em>ad hoc</em></td>
</tr>
<tr>
<td>name</td>
<td><em>Roger Bacon, New York</em></td>
</tr>
<tr>
<td>compound</td>
<td><em>phone book, four thousand, dress up</em></td>
</tr>
<tr>
<td>goeswith</td>
<td><em>not with standing, with out</em></td>
</tr>
</tbody>
</table>

- UD annotation does not permit “words with spaces”
  - Multiword expressions are analyzed using special relations
  - The **mwe**, **name** and **goeswith** relations are always head-initial
  - The **compound** relation reflects the internal structure
## Other Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>parataxis</td>
<td>Loosely linked clauses of same rank</td>
</tr>
<tr>
<td>list</td>
<td>Lists without syntactic structure</td>
</tr>
<tr>
<td>remnant</td>
<td>Orphans in ellipsis linked to parallel elements</td>
</tr>
<tr>
<td>reparandum</td>
<td>Disfluency linked to (speech) repair</td>
</tr>
<tr>
<td>foreign</td>
<td>Elements within opaque stretches of code switching</td>
</tr>
<tr>
<td>dep</td>
<td>Unspecified dependency</td>
</tr>
<tr>
<td>root</td>
<td>Syntactically independent element of clause/phrase</td>
</tr>
</tbody>
</table>
Language-Specific Relations

- Language-specific relations are **subtypes** of universal relations added to capture important phenomena.
- Subtyping permits us to “back off” to universal relations.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>acl:relcl</td>
<td>Relative clause</td>
</tr>
<tr>
<td>compound:prt</td>
<td>Verb particle <em>(dress up)</em></td>
</tr>
<tr>
<td>nmod:poss</td>
<td>Genitive nominal <em>(Mary’s book)</em></td>
</tr>
<tr>
<td>nmod:agent</td>
<td>Agent in passive <em>(saved by the bell)</em></td>
</tr>
<tr>
<td>cc:preconj</td>
<td>Preconjunction <em>(both … and)</em></td>
</tr>
<tr>
<td>det:predet</td>
<td>Predeterminer <em>(all those …)</em></td>
</tr>
</tbody>
</table>
Word Segmentation

- Must be reproducible on new data
- Surface tokens vs. syntactic words
- Chinese, Vietnamese etc.: no clues, non-trivial algorithm
- Arabic, Tamil etc.: part of morphological analysis
- Spanish, German etc.: rather limited cases of contractions
- Others: only punctuation (low-level tokenization)
Word Segmentation

- Fusions
  - $al = a + el$
  - $naň = na + něj$

- Clitics
  - vámonos = vamos + nos
  - изменяться = изменять + ся
  - potrafilibyśmy
    = potrafili + by + jesteśmy
Where Are We Now?
Where Are We Now?

- Two years of UD version 1
- 4 treebank releases (every 6 months)
- 54 (61) treebanks
- 40 (47) languages (over 50% world’s population)
- Over 11M tokens; treebanks range from 1K to 1.5M
- Over 120 contributors
  - language group consistency SIGs
  - version 2 guidelines coming soon
## 47 Languages and Growing

<table>
<thead>
<tr>
<th>Language</th>
<th>Size</th>
<th>Source</th>
<th>License</th>
<th>GitHub</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancient Greek</td>
<td>206K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabic</td>
<td>242K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basque</td>
<td>121K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bulgarian</td>
<td>156K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buryat</td>
<td>5K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalan</td>
<td>530K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>123K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coptic</td>
<td>4K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Croatian</td>
<td>87K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>1,503K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech-CAC</td>
<td>493K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech-CLTT</td>
<td>35K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danish</td>
<td>100K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>209K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch-LassySmall</td>
<td>98K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>254K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-ESL</td>
<td>97K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English-LinES</td>
<td>82K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estonian</td>
<td>234K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faroese</td>
<td>119K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>181K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish-FTB</td>
<td>159K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>390K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galician</td>
<td>138K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>293K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gothic</td>
<td>56K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greek</td>
<td>59K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>115K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>351K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>42K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indonesian</td>
<td>121K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irish</td>
<td>23K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>252K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese-KTC</td>
<td>267K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kazakh</td>
<td>4K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin</td>
<td>47K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin-ITTB</td>
<td>291K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin-PROIEL</td>
<td>165K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latvian</td>
<td>20K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norwegian</td>
<td>311K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old Church Slavonic</td>
<td>87K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persian</td>
<td>151K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td>83K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td>209K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese-BR</td>
<td>298K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romanian</td>
<td>145K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>99K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian-SynTagRus</td>
<td>1,032K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanskrit</td>
<td>1K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenian</td>
<td>140K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenian-SST</td>
<td>29K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>423K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish-AnCora</td>
<td>547K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>96K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish-LinES</td>
<td>79K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tamil</td>
<td>8K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>56K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ukrainian</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urdu</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uyghur</td>
<td>45K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnamese</td>
<td>43K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Where Are We Going?

- UD guidelines version 2 coming soon
- Consistency checking
Common vocabulary is great …

… because we finally understand each other …
... almost

Childs of you be vary acute!
Consistency Checking

- Automatic tests catch only a fraction

- Focus groups on
  - Romance, Germanic, Slavic, Uralic, Turkic languages
# Existing Slavic Treebanks

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>Treebank</th>
<th>Sent</th>
<th>Tok</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>[bg]</td>
<td>BulTreeBank</td>
<td>13,221</td>
<td>196K</td>
</tr>
<tr>
<td>Church Slavonic</td>
<td>[cu]</td>
<td>PROIEL</td>
<td>7,818</td>
<td>72K</td>
</tr>
<tr>
<td>Croatian</td>
<td>[hr]</td>
<td>SETimes.HR</td>
<td>3,736</td>
<td>84K</td>
</tr>
<tr>
<td>Croatian</td>
<td>[hr]</td>
<td>HOBS</td>
<td>4,626</td>
<td>117K</td>
</tr>
<tr>
<td>Czech</td>
<td>[cs]</td>
<td>PDT</td>
<td>87,913</td>
<td>1504K</td>
</tr>
<tr>
<td>Czech</td>
<td>[cs]</td>
<td>CAC</td>
<td>24,709</td>
<td>494K</td>
</tr>
<tr>
<td>Polish</td>
<td>[pl]</td>
<td>IPI PAN</td>
<td>8,227</td>
<td>84K</td>
</tr>
<tr>
<td>Russian</td>
<td>[ru]</td>
<td>SynTagRus</td>
<td>59,130</td>
<td>1033K</td>
</tr>
<tr>
<td>Russian</td>
<td>[ru]</td>
<td>Google</td>
<td>5,030</td>
<td>99K</td>
</tr>
<tr>
<td>Slovak</td>
<td>[sk]</td>
<td>SNK</td>
<td>63,238</td>
<td>994K</td>
</tr>
<tr>
<td>Slovenian</td>
<td>[sl]</td>
<td>SSJ500K</td>
<td>27,829</td>
<td>500K</td>
</tr>
<tr>
<td>Slovenian</td>
<td>[sl]</td>
<td>SST</td>
<td>3,188</td>
<td>29K</td>
</tr>
</tbody>
</table>
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions
Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective (this, which, every)
Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective (*this, which, every*)
- We don’t have this category! (Traditionally → PRON.)
Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective \((this, which, every)\)
- We don’t have this category! (Traditionally → PRON.)
- Some authors do recognize determiners in Slavic!
Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective (*this, which, every*)
- We don’t have this category! (Traditionally → PRON.)
- We have the words (except for articles).
Pronouns and Determiners

• English + Romance languages: DET = article or pronominal adjective (this, which, every)
• We don’t have this category! (Traditionally → PRON.)
• We have the words (except for articles).
• Currently functional borderline (but ellipsis?)
  This.DET car is expensive.
  This.PRON is expensive.
• Less strict in UD v2.
Pronouns Only

- Personal pronouns (including reflexives, but not possessives)
- Interrogative *who, what*
- Indefinite and negative derivatives
- Relative [cs] *jenž*
  - cs: já, ty, on, my, vy, oni, se, kdo, co, někdo, něco, nikdo, nic
  - sk: ja, ty, on, my, vy, oni, sa, kto, čo, nieko, niečo, nikto, nič
  - pl: ja, ty, on, my, wy, oni, się, kto, co, ktoś, coś, nikt, nic
  - ru: я, ты, он, мы, вы, они, ся, кто, что, кто-нибудь, что-нибудь, никто, ничто
  - sl: jaz, ti, on, mi, vi, oni, se, kdo, kaj, nekdo, nekaj, nihče, nič
  - hr: ja, ti, on, mi, vi, oni, se, tko, što, neki, nešto, nitko, ništa
  - bg: аз, ти, ние, вие, се, кой, кое, някой, нещо, никой, нищо
  - cu: азъ, тъ, мъ, въ, и, съ, къто, чьто
Possessives: Determiners

- If they occur without a noun … ellipsis

*Můj otec je starší. Tvůj má ale více zkušeností.*
My father is older. But *yours* is more experienced.

- sl: *moj, tvoj, njegov, njen, najin, vajin, njun, naš, vaš, njihov, svoj*
- bg: *мой, твой, негов, неин, наш, ваш, техен, свой*
- cs: *můj, tvůj, jeho, její, náš, váš, jejich, svůj*
- sk: *môj, tvoj, jeho, jej, náš, váš, ich, svoj*
- cu: *мои, твои, нашь, вашь, свои / его, ем, ею, ихь*
Both Possible?

- **Demonstratives**
  - cs: *ten, to, tento, tenhle, tamten*, ...
  - sl: *ta, to, tisti, oni, takšen*, ...

- **Adjectival interrogatives/relatives, indefinites, negatives**
  - *jaký, který, čí, nějaký, některý, něčí, každý, žádný*
  - *všechnen, všichni, všechno*

- **Relative pronouns cannot be explained by ellipsis!**
  - *Muž, kterého* *muže jsem vám představil.*
  - *The man, which* *man I introduced to you.*
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions
Quantified Noun Phrase

nummod  nsubj  punct

Jedno  kotě  spalo  One  kitten  slept

29.9.2016, Ljubljana
Quantified Noun Phrase

Pět  kot’at  spalo
Five  kittens  slept

nummod  nsubj  punct
Quantified Noun Phrase

Pět
Five

kot’at
kittens

spalo
slept

Genitive!
Quantified Noun Phrase

Skupina Group

kot’at of-kittens

spala slept
Quantified Noun Phrase

Pěť  kot’at  spalo
Five  kittens  slept

nummod:gov  nsubj  punct
Pronominal Quantifiers

Kolik
How-many

kot’at
kittens

spalo
slept

?  ?
## Language-Specific Labels

<table>
<thead>
<tr>
<th>Noun governs</th>
<th>Numeric</th>
<th>Pronominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeral governs</td>
<td>nummod:gov</td>
<td>det:numgov</td>
</tr>
</tbody>
</table>
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions
Verb Forms

- Conflicting terminologies in traditional grammars
- Participle … verb or adjective?
- Converb … verb or adverb?

- Tags and features apply to *individual words!*
Verb Forms

- POS tags and features apply to individual words!
- *A ko so se leta 1942 *vračali*, …
  - past tense
- … *da ne bi* v Atene *prišli* …
  - conditional mood
- … *v prihodnje ne* *bodo vozili* zgolj les …
  - future tense
Verb Forms

• POS tags and features apply to individual words!

• A ko so se leta 1942 vračali, …
  – past tense

• … da ne bi v Atene prišli …
  – conditional mood

• … v prihodnje ne bodo vozili zgolj les …
  – future tense
Verb Forms

- POS tags and features apply to individual words!
- *A ko so se leta 1942 *vračali*, ...
  - past tense
- *... da ne bi v Atene prišli* ...
  - conditional mood
- *... v prihodnje ne bodo vozili zgolj les* ...
  - future tense
Verb Forms

- **vračali, prišli, vozili**

- [cs] “active participle” / “past tense”
- [ru] “past tense” / “finite!”
  - Active participle is something else: нарушивший
- [bg] “participle + past (aorist) / imperfect” (two subtypes)
- [cu] “participle + resultative aspect” (lang-spec)
- “l-participle”
  - But that would be a language-specific verb form.
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions
Core Arguments

- Easier cross-linguistically than argument-adjunct?
- Subject of intransitive verb
- Agent of transitive verb
- Patient (direct object) of transitive verb

- Indirect object? Dative only?
Core vs. Oblique Dependents

• **Core arguments**: what exactly is it?

• English:
  - *He gave John the book.* (iobj)
  - *He gave the book to John.* (nmod)

• Spanish:
  - *Dio el libro a John.* (iobj)

• Czech:
  - Every Obj is translated to dobj, regardless the case and the presence of preposition
dobj / iobj

• Not as easy as accusative vs. dative.
• Default: dobj
• Heuristics for iobj
  – *Cením si vaší pomoci*. (Gen)
    I appreciate your help.
  – *Čelíme velkým problémům*. (Dat)
    We are facing big problems.
  – *Nedisponuje takovým rozpočtem*. (Ins)
    He does not have such budget.
  – *Učí mou dceru fyziku*. (2 × Acc)
    He teaches my daughter physics.
All Slavic Treebanks Have Non-Accusative “Direct” Objects

- podrobit se testu; odpovídají smlouvě; jednat s někým
- mówi o niej; używa wielkich słów
- от которых зависит; относится к программам
- potřebuje informacij; slediti evropskim smernicam; ukvarjal se bom orožjem
- odriče se imuniteta; priključiti se naporima
- се характеризира с развитие; моля за внимание
Reflexive Pronouns

- Direct or indirect object (dobj, iobj):
  Řízl se do prstu / Řízl ho do prstu.
  - Including reciprocal usage:
    Políbili se. / They kissed each other.

- Inherently reflexive verbs: smát se, bát se / laugh, fear
  - expl:pv (pronominal verb; previously compound)

- Reflexive passive:
  To se snadněji řekne než udělá. / That is easier said than done.
  - expl:pass (previously auxpass:reflex)

- Impersonal construction (~ passive?):
  Zde se mluví německy. / German is spoken here.
  - expl:impers
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions
Modal Auxiliary in English

- nsubj
- aux
- aux
- aux
- advmod

I should have been waiting there
Modal Verb in Czech

- byla
- bych
- tam
- měla
- čekat
- was
- I-would
- there
- should
- wait

aux
advmod
xcomp
Modal Adverb in Russian

Мне надо выпить воды.
Mne надо выпить воды.
To-me necessary to-drink water.

пункт

iobj
xcomp
dobj

Modal / Control Verb in English

I need to drink some water.
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions
Comparative Constructions

Му́й, Мой, Мо́й, My Му́й, Мой, Мо́й, My
otec, otec, otec, father
je, старше, starše, is
starší, сваршье, старше, older
než, твоего, твоего, than
tvůj, твоего, yours
punct
nsubj
cop
nmod
case
Comparative Constructions

Ubytovali nás v dražším hotelu, než jsme čekali.

Oni вселили нас в более дорогой отель, чем мы ожидали.

They put us in more expensive hotel, than we had expected.
Wrapping Up
Wrapping Up

• UD has had a great start
Wrapping Up

- UD has had a great start
- Still a long way to go. Consistency matters!
Wrapping Up

- UD has had a great start

- Still a long way to go. Consistency matters!

- Get involved. It’s fun!
Thank you!
Questions?

Děkuji!
Otázky?

Благодаря!
Въпроси?

Dziękuję!
Pytania?

Благодаря!
Въпроси?

Thank you!
Questions?

Dəkũjedy!
Pytania?

Благодаря!
Въпроси?

Hvala!
Pitanja?

Hvala!
Pitanja?

Hvala!
Pitanja?