The Universe of Dependencies: Past and Future

Daniel Zeman
zeman@ufal.mff.cuni.cz
http://universaldependencies.org/
My daughter bought some bread and cheese

Min datter købte nogle brød og ost

Min dotter köpte några bröd och ost
Universal Dependencies

http://universaldependencies.org/

Milestones:

- 2008-05 Interset (morphological features)
- 2012-05 Google Universal POS tags
- 2012-05 HamleDT (harmonized Prague-style treebanks)
- 2013-08 Google Universal Dependency Treebank
- 2014-05 Universal Stanford Dependencies
- 2014-04 EACL Göteborg, kick-off meeting of UD
- 2014-10 UD guidelines version 1
- 2015-01 released first 10 treebanks
- every ~6 months new release
- 2016-12 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*

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

Některé dívky si nicméně pochvalovaly zmrzlinu.

Some girls nevertheless praised ice-cream.
Některé dívky si nicméně pochvalovaly zmrzlinu.

Some girls nevertheless praised ice-cream.

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

Some girls nevertheless praised ice-cream.

- **Lemma** representing the semantic content of the word
- **Part-of-speech** tag representing the abstract lexical category associated with the word
Některé dívky si nicméně pochvalovaly zmrzlinu.

Some girls nevertheless praised ice-cream.

- 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>CCONJ</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>PRON</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 v1

<table>
<thead>
<tr>
<th><strong>Lexical</strong></th>
<th><strong>Inflectional (Nominal)</strong></th>
<th><strong>Inflectional (Verbal)</strong></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>Reflect</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>
</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
<table>
<thead>
<tr>
<th><strong>Lexical</strong></th>
<th><strong>Inflectional (Nominal)</strong></th>
<th><strong>Inflectional (Verbal)</strong></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>Reflect</td>
<td>Case</td>
<td>Aspect</td>
</tr>
<tr>
<td><strong>Foreign</strong></td>
<td><strong>Definite</strong></td>
<td><strong>Voice</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Degree</strong></td>
<td><strong>Evident</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Person</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Polite</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Polarity</strong></td>
</tr>
<tr>
<td><strong>Abbr</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The cat could have chased all the dogs down the street.

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

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

- Content words are related by dependency relations
Syntax

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

- Function words attach to closest content words
- Punctuation attach to head of phrase or clause
Syntax

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

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

Kučeto se presledvaše ot kotkata.
The dog was chased by the cat.

Kučeto se presledvaše ot kotkata.
The dog was chased by the cat.

Kučeto beše presledvano ot kotkata.
The dog was chased by the cat.

Pes byl honěn kočkou.
Dependency Relations

- Taxonomy of 38 (40) universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
  - Language-specific subtypes may be added
Dependency Relations

- Taxonomy of 38 (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)
<table>
<thead>
<tr>
<th>Core</th>
<th>Nominal</th>
<th>Clausal</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nsubj</td>
<td>csubj</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(nsubj:pass)</td>
<td>(csubj:pass)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(d)obj</td>
<td>ccomp</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iobj</td>
<td>xcomp</td>
<td></td>
</tr>
<tr>
<td>Non-Core</td>
<td>obl (nmod)</td>
<td>advcl</td>
<td>advmod</td>
</tr>
<tr>
<td></td>
<td>vocative</td>
<td></td>
<td>(neg)</td>
</tr>
<tr>
<td></td>
<td>discourse</td>
<td></td>
<td>aux</td>
</tr>
<tr>
<td></td>
<td>expl</td>
<td></td>
<td>(aux:pass)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cop</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>mark</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>punct</td>
</tr>
</tbody>
</table>
Dependents of Nominals

**Nominal**
- nmod
- appos
- nummod
- clf (classif?)

**Clausal**
- acl

**Other**
- amod
- det
- (neg)
- case

Ljubljana, the lovely capital of Slovenia

Daniel Zeman (ÚFAL MFF UK)

Universal Dependencies

Praha, 28.11.2016
“Stanford-style” Coordination in V1 ...

- 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
Coordinate structures are headed by the first conjunct

- Subsequent conjuncts depend on it via the `conj` relation
- Conjunctions depend on the next conjunct via the `cc` relation
- Punctuation marks depend on the next conjunct via the `punct` relation
### Multiword Expressions

<table>
<thead>
<tr>
<th><strong>Relation</strong></th>
<th><strong>Examples</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed (mwe)</td>
<td>in spite of, as well as, ad hoc</td>
</tr>
<tr>
<td>flat (name)</td>
<td>president Havel, New York, four thousand</td>
</tr>
<tr>
<td>compound</td>
<td>phone book, dress up</td>
</tr>
<tr>
<td>goeswith</td>
<td>notwith standing, with out</td>
</tr>
</tbody>
</table>

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

- Words with spaces may be allowed in v2:
  - Vietnamese (spaces delimit syllables, not words)
  - Numbers (“1 000 000”)
  - Possibly other approved cases, e.g. multi-word abbreviations
<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>orphan</td>
<td>Orphans in ellipsis linked together</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.

### Language-Specific 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 (dress <em>up</em>)</td>
</tr>
<tr>
<td>nmod:poss</td>
<td>Possessive nominal (Mary ‘s book)</td>
</tr>
<tr>
<td>obl:agent</td>
<td>Agent in passive (saved <em>by the bell</em>)</td>
</tr>
<tr>
<td>cc:preconj</td>
<td>Preconjunction (<em>both</em> ... and)</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**

\[
\begin{align*}
\text{Vamos} & \quad \text{nos} & \quad a & \quad \text{el} & \quad \text{mar} & . & \\
\text{VERB} & \quad \text{PRON} & \quad \text{ADP} & \quad \text{DET} & \quad \text{NOUN} & \quad \text{PUNCT} \\
\text{Vámonos} & \quad \text{al} & \quad \text{mar} & . & \\
\text{VERB}+\text{PRON} & \quad \text{ADP}+\text{DET} & \quad \text{NOUN} & \quad \text{PUNCT}
\end{align*}
\]
Word Segmentation

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

- **Clitics**
  - vámonos = vamos + nos
  - izmenjat’sja = izmenjat’ + sja
  - potrafilibyśmy = potrafili + by + jesteśmy
UD Version 2

- Words with spaces
- POS tags (CONJ → CCONJ, redefined AUX, PART, PRON/DET)
- Morphological features (new features, new values, some renamed)
- Predicate dependents: dobj/obj, nmod/obl
  - nsubj:pass, csubj:pass, aux:pass made subtypes
- Nonverbal predicates (copula)
- Coordination and ellipsis
- Functional relations
  - New relation clf (classifier)
  - aux not only for verbs
  - Removed relation neg (use features instead)
- Multiword expressions
  - mwe renamed fixed, name removed, flat added instead (covers more)
  - compound extended to complex predicates (light verbs, serial verbs)
Nonverbal Predicate and Copula

- Some languages use a copula verb:

  Ivan is the best dancer .

- Some languages use a copula **pronoun**:

  Ivan – to najlepszy tancerz .
Nonverbal Predicate and Copula

- Some languages use a copula verb:

  Ivan is the best dancer.

- Some languages omit the copula:

  Ivan lučšij tancor.
Nonverbal Predicate and Copula

- Some languages use a copula verb:

  Ivan was the best dancer.

- Some languages use it only in some tenses:

  Ivan byl lučším tancorom.
Copula Verbs: We Are Restrictive!

- *To be* is copula:

```
Ivan is the best dancer .
```

- *To become* is not copula:

```
Ivan became the best dancer .
```
Once Copula, Always Copula!

- This is parallel with Russian:

```
Ivan is the best dancer.
```

- This is also parallel with Russian:

```
Ivan is today in Moscow.
```
Well, Almost…

- This is parallel with Russian:

```
Ivan is today in Moscow .
```

- But not with this in English:

```
There is a dancer in Moscow .
```
Clauses and Copula

- A clause can be the subject:

```
The problem is that he is missing.
```

- But it cannot be annotated as the nonverbal predicate:

```
The problem is that he is missing.
```
UD Version 2

- Words with spaces
- POS tags (CONJ → CCONJ, redefined AUX, PART, PRON/DET)
- Morphological features (new features, new values, some renamed)
- Predicate dependents: dobj/obj, nmod/obl
  - nsubj:pass, csubj:pass, aux:pass made subtypes
- Nonverbal predicates (copula)
- Coordination and ellipsis
- Functional relations
  - New relation clf (classifier)
  - aux not only for verbs
  - Removed relation neg (use features instead)
- Multiword expressions
  - mwe renamed fixed, name removed, flat added instead (covers more)
  - compound extended to complex predicates (light verbs, serial verbs)
V2 Coordination Is Still Left-Headed

- Coordinate structures are headed by the first conjunct
  - Subsequent conjuncts depend on it via the \textit{conj} relation
  - Conjunctions depend on the next conjunct via the \textit{cc} relation
  - Punctuation marks depend on the next conjunct via the \textit{punct} relation
But Some Languages Might Prefer the Opposite

Coordinate structures would be headed by the last conjunct
- Preceding conjuncts would depend on it via the conj relation
- Conjunctions would depend on the preceding conjunct
- Punctuation marks would depend on the preceding conjunct
Some treebanks would use an empty node to represent the second *went*.

UD enhanced representation now allows empty nodes!

But the basic representation sticks with the overt words.
Kate went to Florida and Jane (went) to Europe
PDT: The ExD Relation

Kate went to Florida and Jane (went) to Europe
Kate went to Florida and Jane (went) to Europe
UD V2: The orphan Relation

Kate went to Florida and Jane (went) to Europe
Where Are We Now?

- Two years of UD version 1
- 5 treebank releases (every 6 months)
- 64 treebanks
- 47 languages (over 50% world’s population)
- Over 12M tokens; treebanks range from <1K to 1.5M
- Over 140 contributors
  - language group consistency SIGs
- Version 2 guidelines to be launched this week
- CoNLL Shared Task in parsing UD coming next year
# 47 Languages and Growing

<table>
<thead>
<tr>
<th>Language</th>
<th>Size</th>
<th>Tag</th>
<th>License</th>
<th>Dependency</th>
<th>Language</th>
<th>Size</th>
<th>Tag</th>
<th>License</th>
<th>Dependency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancient Greek-PROIEL</td>
<td>206K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Irish</td>
<td>23K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Arabic</td>
<td>242K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Italian</td>
<td>252K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Basque</td>
<td>121K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Japanese-KTC</td>
<td>267K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Bulgarian</td>
<td>156K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Kazakh</td>
<td>4K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Buryat</td>
<td>5K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Korean</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Catalan</td>
<td>530K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Latin</td>
<td>47K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Chinese</td>
<td>123K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Latin-ITTB</td>
<td>291K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Coptic</td>
<td>4K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Latin-PROIEL</td>
<td>165K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Croatian</td>
<td>87K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Latvian</td>
<td>20K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Czech</td>
<td>1,503K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Norwegian</td>
<td>311K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Czech-CAC</td>
<td>493K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Old Church Slavonic</td>
<td>57K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Czech-CLTT</td>
<td>35K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Persian</td>
<td>151K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Danish</td>
<td>100K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Polish</td>
<td>83K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Dutch</td>
<td>209K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Portuguese</td>
<td>209K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Dutch-LassySmall</td>
<td>98K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Portuguese-BR</td>
<td>298K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>English</td>
<td>254K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Romanian</td>
<td>145K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>English-ESL</td>
<td>97K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Russian</td>
<td>99K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>English-LinES</td>
<td>82K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Russian-Syntagrus</td>
<td>1,032K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Estonian</td>
<td>234K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Sanskrit</td>
<td>1K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Faroese</td>
<td>119K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Slovenian</td>
<td>140K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Finnish</td>
<td>181K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Slovenian-SST</td>
<td>29K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Finnish-FTB</td>
<td>159K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Spanish</td>
<td>423K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>French</td>
<td>390K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Spanish-Ancora</td>
<td>547K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Galician</td>
<td>138K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Swedish</td>
<td>96K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>German</td>
<td>293K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Swedish-LinES</td>
<td>79K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Gothic</td>
<td>56K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Tamil</td>
<td>8K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Greek</td>
<td>59K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Turkish</td>
<td>56K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Hebrew</td>
<td>115K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Ukrainian</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hindi</td>
<td>351K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Urdu</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hungarian</td>
<td>42K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Uyghur</td>
<td>45K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
<tr>
<td>Indonesian</td>
<td>121K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
<td>Vietnamese</td>
<td>43K</td>
<td><img src="flag.png" alt="Flag" /></td>
<td><img src="license.png" alt="License" /></td>
<td><img src="dependency.png" alt="Dependency" /></td>
</tr>
</tbody>
</table>
Where Are We Going?

- Launching UD version 2
- Treebanks v2 release in February/March (CoNLL Shared Task!)
- CoNLL Shared Task in parsing UD coming next year
- Consistency checking
• Common vocabulary is great …
• 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>6,289</td>
<td>139K</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>Czech</td>
<td>[cs]</td>
<td>CLTT</td>
<td>1,125</td>
<td>35K</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>60,551</td>
<td>1068K</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>
<tr>
<td>Ukrainian</td>
<td>[uk]</td>
<td>UDT</td>
<td>255</td>
<td>2K</td>
</tr>
</tbody>
</table>
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.)
- 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).

- 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).
- Some authors do recognize determiners in Slavic!
- UD v1: functional borderline (but ellipsis?)
  - *This.DET car is expensive.*
  - *This.PRON is expensive.*
- Less strict in UD v2.
Quantified Noun Phrase

Jedno kotě spalo.
One kitten slept.

nummod nsubj punct
Jedno kotě spalo .
One kitten slept .
Quantified Noun Phrase

Five kittens slept.

Pět koťat spalo.

Five kittens slept.
Quantified Noun Phrase

Pět koťat spalo.
Five kittens slept.

Case=Gen
Quantified Noun Phrase

Skupina kotat spala.

Group kittens slept.

Case=Gen
Quantified Noun Phrase

Pět koťat spalo.
Five kittens slept.
Quantified Noun Phrase

Kolik koťat spalo?

How-many kittens slept?
## Language-Specific Labels

<table>
<thead>
<tr>
<th></th>
<th>Numeric</th>
<th>Pronominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun governs</td>
<td>nummod</td>
<td>det:nummod</td>
</tr>
<tr>
<td>Numeral governs</td>
<td>nummod:gov</td>
<td>det:numgov</td>
</tr>
</tbody>
</table>
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
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
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: *narušivšij*

  *bg* “participle + past (aorist) / imperfect” (two subtypes)

  *cu* “participle + resultative aspect” (lang-spec)
  
  - “l-participle”
    
    - But that would be a language-specific verb form.
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.* (obl)
- **Spanish**:
  - *Dio el libro a John.* (iobj)
- **Czech**:
  - Currently every Obj is translated to (d)obj, regardless the case and the presence of preposition
Direct and Indirect Object

- Not as easy as accusative vs. dative.
- Default: obj
- 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 (D)Objects

- podrobit se testu; odpovídají smlouvě; jednat s někým
- mówi o niej; używa wielkich słów
- ot kotorix zavisit; otnositsja k programmam
- potrebuje informacij; slediti evropskim smernicam; ukvarjal se bom orožjem
- odriče se imuniteta; priključiti se naporima
- se karakterizira s razvitie; molja za vnimanie
Reflexive Pronouns

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

- Direct or indirect object (obj, 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 Pronouns

- Direct or indirect object (obj, 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)
Reflexive Pronouns

- Direct or indirect object (obj, 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
Modal Auxiliary in English

I should have been waiting there
Modal Verb in Czech

byla bych tam měla čekat
was I-would there should wait
Modal Predicative in Russian

Mne 
nado  vypit’  vody
To-me necessary to-drink water
Modal / Control Verb in English

I need to drink some water

```
I  nsubj
   \-
  need  xcomp
     \-
  to  mark
    \-
drink  obj
       \-
det
```

Daniel Zeman (ÚFAL MFF UK)
Comparative Constructions

Můj otec je starší než tvůj.
Moj otec starše tvoego.
My father is older than yours.
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?