Morphosyntactictic Annotation in Universal Dependencies for Old Czech

Daniel Zeman, Pavel Kosek, Martin Březina, Jiří Pergler
Charles University & Masaryk University & Czech Language Institute

20.10.2023
Universal Dependencies
Universal Dependencies

- [https://universaldependencies.org/](https://universaldependencies.org/)
- Same things annotated same way across languages...
- ... while highlighting different coding strategies
Same Thing Same Way

George killed the dragon

Morphosyntactic Annotation in Universal Dependencies for Old Czech
George killed the dragon

Mharaigh Seoirse an dragan
Same Thing Same Way

George killed the dragon

Mharaigh Seoirse an dragan

Jorge mató al dragón

Morphosyntactic Annotation in Universal Dependencies for Old Czech
George killed the dragon

Mharaigh Seoirse an dragan

Jorge mató a el dragón
George killed the dragon

Mharaigh Seoirse an dragan

Jorge mató a el dragón

Draka zabil Jiří
Same Meaning ≠ Same Construction!

He killed the dragon

PRON VERB DET NOUN
Same Meaning ≠ Same Construction!

He \_ killed the dragon

The dragon was killed by him

Morphosyntactic Annotation in Universal Dependencies for Old Czech
Same Meaning ≠ Same Construction!

- **He** killed the **dragon**
  - `He` (PRON)
  - `killed` (VERB)
  - `the` (DET)
  - `dragon` (NOUN)

- **The dragon** was killed **by him**
  - `The` (DET)
  - `dragon` (NOUN)
  - `was killed` (AUX VERB ADP PRON)
  - `by` (ADP)
  - `him` (PRON)

- **His killing of the dragon**
  - `His` (PRON)
  - `killing` (NOUN)
  - `of` (ADP)
  - `the` (DET)
  - `dragon` (NOUN)

**Morphosyntactic Annotation in Universal Dependencies for Old Czech**
Same Meaning ≠ Same Construction!

He killed the dragon.

The dragon was killed by him.

His killing of the dragon.

The dragon that was killed.
Morphological Annotation

- Lemma representing the semantic content of a word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form
Basic Universal Dependencies: 141 Languages and Growing

- I.-E.: Armenian (+West), Greek (+Ancient), Albanian, Hittite, Breton, Irish (+Old), Manx, Scottish, Welsh, Afrikaans, Danish, Dutch, English, Faroese, Frisian, German, Gothic, Icelandic, Low Saxon, Norwegian, Swedish, Swiss German, Catalan, French, Galician, Italian, Latin, Ligurian, Neapolitan, Old French, Portuguese, Romanian, Spanish, Umbrian, Belarusian, Bulgarian, Church Slavonic, Croatian, Czech, Old Russian, Polish, Pomak, Russian, Serbian, Slovak, Slovenian, Ukrainian, Upper Sorbian, Latvian, Lithuanian, Kurmanji, Persian, Khunsari, Nayini, Soi, Urdu, Hindi, Kangri, Bhojpuri, Bengali, Marathi, Sanskrit, Sinhala
- Dravidian: Tamil, Malayalam, Telugu
- Uralic: Erzya, Estonian, Finnish, Hungarian, Karelian, Livvi, Komi Permyak+Zyrian, Moksha, Sámi North+Skolt
- Turkic: Kazakh, Kyrgyz, Old Turkish, Tatar, Turkish, Uyghur, Yakut, Buryat, Xibe
- Korean, Japanese
- Sino-T.: Cantonese, Classical Chinese, Chinese
- Tai-Kadai: Thai
- Aus.-As.: Vietnamese
- Austron.: Indonesian, Javanese, Tagalog, Cebuano
- Pama-Nyu.: Warlpiri
- Chu.-Kam.: Chukchi
- Esk.-Al.: Yupik
- Uto-Aztec.: Nahuatl, Mayan: Kiche
- Arawakan: Apurinã, Madi
- Macro-Je: Xavante, Bororo, Tupian: Akuntsu, Guajajara, Kaapor, Karo, Makurap, Mundurukú, Nheengatu, Tupinambá, Mbyá, Guaraní, Teko
- Af.-As.: Akkadian, Assyrian, Beja, Coptic, Hebrew

Morphosyntactic Annotation in Universal Dependencies for Old Czech
### Slavic Languages in UD 2.12

<table>
<thead>
<tr>
<th>Language</th>
<th>× 1,000 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belarusian</td>
<td>305</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>156</td>
</tr>
<tr>
<td>Croatian</td>
<td>199</td>
</tr>
<tr>
<td>Czech</td>
<td>2,227</td>
</tr>
<tr>
<td>Old Church Slavonic</td>
<td>199</td>
</tr>
<tr>
<td>Old East Slavic</td>
<td>333</td>
</tr>
<tr>
<td>Polish</td>
<td>499</td>
</tr>
<tr>
<td>Pomak</td>
<td>87</td>
</tr>
<tr>
<td>Russian</td>
<td>1,832</td>
</tr>
<tr>
<td>Serbian</td>
<td>97</td>
</tr>
<tr>
<td>Slovak</td>
<td>106</td>
</tr>
<tr>
<td>Slovenian</td>
<td>297</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>123</td>
</tr>
<tr>
<td>Upper Sorbian</td>
<td>11</td>
</tr>
</tbody>
</table>
Czech in UD
Czech UD Treebanks

- **PDT** (Prague Dependency Treebank)
  - 87K sentences, 1.5M words
- **CAC** (Czech Academic Corpus / Korpus věcného stylu)
  - non-fiction, 1971–1985
  - 24K sentences, 493K words
- **FicTree**
  - 12K sentences, 166K words
- **CLTT** (Czech Legal Text Treebank)
  - The Accounting Act (Zákon o účetnictví), 1991–2016
  - 1K sentences, 36K words
- **PUD** (Parallel Universal Dependencies)
  - online news + Wikipedia, translated from en/de/fr/it/es, around 2016
  - 1K sentences, 18K words
Old Czech UD Treebank?

- Pilot study
- Dresden Bible (around 1360)
- Olomouc Bible (1417)
- Gospel of Matthew (from both versions)
  - 2K sentences, 44K words
Old Czech UD Treebank?

- Pilot study
- Dresden Bible (around 1360)
- Olomouc Bible (1417)
- Gospel of Matthew (from both versions)
  - 2K sentences, 44K words

- Bootstrapping:
  - Parse a part using a parser
  - Manually check and fix
  - Re-train the parser
  - Parse another part
  - Manually check and fix
  - ...

Universal Dependencies
Old Czech UD Treebank?

- Pilot study
- Dresden Bible (around 1360)
- Olomouc Bible (1417)
- Gospel of Matthew (from both versions)
  - 2K sentences, 44K words

- Bootstrapping:
  - Parse a part using a parser but available models are modern Czech!
  - Manually check and fix
  - Re-train the parser
  - Parse another part
  - Manually check and fix
  - ...

Morphosyntactic Annotation in Universal Dependencies for Old Czech | Universal Dependencies
PDT Model vs. Old Czech Data

- Genre, vocabulary: news vs. Bible
- Old vocabulary
- Orthography
  - Cleaned, transcribed, unified
  - But still not modern forms: sě, viece
- Grammar:
  - Dual number
  - Simple past (imperfect, aorist) (bieše, vecě, jide)
  - Converbs (přechodníky) (řka, přístůpiv)
• Pavel dal geft ploſcouicích zemu Wlah dalgeft dolaf zemu bogu iſuiatemu Šcepanu ſeduema dušnícoma bogucea aſedlatu

• Pavel dal jest Ploskovicích zem’u, Vlach dal jest Dolas zem’u bogu i sv’atému Ščepánu se dvěma dušníkoma, Bogučeja a Sedlatu.

• Pavel dal v Ploskovicích zemi, Vlach dal v Dolanech zemi bohu i svatému Štěpánovi se dvěma dušníky, Bogučejem a Sedlatou.

• “Pavel gave land in Ploskovice, Vlach gave land in Dolas to God and St. Stephen with two villeins, Bogučej and Sedlata.”
• Pavel dal geſt ploſcouicih zemu Wlah dalgeſt dolaſ zemu bogu ſfuiaatemu ſcepanu ſeduema duſnicoma bogucea aſedlatu

• Pavel dal jest Ploskovicích zem’u, Vlach dal jest Dolas zem’u bogu i sv’atému Ščepánu se dvěma dušníkoma, Bogučeja a Sedlatu.

• Pavel dal v Ploskovicích zemi, Vlach dal v Dolanech zemi bohu i svatému Štěpánovi se dvěma dušníky, Bogučejem a Sedlatou.

• “Pavel gave land in Ploskovice, Vlach gave land in Dolas to God and St. Stephen with two villeins, Bogučej and Sedlata.”
Pavel dal geſť ploſcouicích zemu Wlah dalgeſť dolaf zemu bogu ľuatiemu Šcepanu ľeduema duſnícoma bogucea aſedlatu

Pavel dal jest Ploskovicích zem’u, Vlach dal jest Dolas zem’u bogu i sv’atému Ščepánu se dvěma dušníkoma, Bogučeja a Sedlatu.

Pavel dal v Ploskovicích zemi, Vlach dal v Dolanech zemi bohu i svatému Štěpánovi se dvěma dušníky, Bogučejem a Sedlatou.

“Pavel gave land in Ploskovice, Vlach gave land in Dolas to God and St. Stephen with two villeins, Bogučej and Sedlata.”
Lemmatization

• Not only picking base form in the paradigm...
• ... but also normalization among alternatives...
• ... even after modernizing orthography!
  • Křtitel “Baptist”
  • Křstítel
  • Krstitel
Lemmatization

• Not only picking base form in the paradigm...
• ... but also normalization among alternatives...
• ... even after modernizing orthography!
  • Křtitel “Baptist”
  • Křstitel
  • Krstitel

• Modern lemma vs. old lemma:
  • forms = otsúdí / otsúdie “she / they will condemn”
  • lemma candidates = odsúziti, otsoudici, otsoudit, otsouditi, otsúdici, otsúdit, otsúditi, otsoudici, otsoudit, otsouditi, otsúdici, otsúdit, otsúditi, otsoudic, otsouditi, otsúdici, otsúdit, otsúditi
  • “hyperlemma” ⇒ lemma1300 = otsúditi
Lemmatization

- Not only picking base form in the paradigm...
- ... but also normalization among alternatives...
- ... even after modernizing orthography!
  - \textit{Křtitel} “Baptist”
  - \textit{Křstitel}
  - \textit{Krstitel}

- Modern lemma vs. old lemma:
  - \textit{forms = otsúdí / otsúdie} “she / they will condemn”
  - \textit{lemma candidates = odsúziti, otsoudici, otsoudit, otsouditi, otsúdici, otsúdit, otsúditi, votsoudici, votsoudit, votsouditi, votsúdici, votsúdit, votsúditi}
  - “hyperlemma” \Rightarrow \textit{lemma1300 = otsúditi}
  - (modern) \textit{lemma = odsoudit}
Simple Past Tense: Imperfect

- Dresden: *Ale Kristovo porozenie tak bieše.*
  - **VERB** VerbForm=Fin Mood=Ind Tense=Imp Aspect=Imp Voice=Act Number=Sing Person=3 Polarity=Pos
- Modern: *S narozením Ježíše Krista to bylo takto:*
  - **VERB** VerbForm=Part Tense=Past Aspect=Imp Voice=Act Number=Sing Gender=Neut Polarity=Pos
- English: “Now the birth of Jesus Christ took place in this way.”
Simple Past Tense: Imperfect

- Dresden: *Ale Kristovo porozenie tak bieše.*
  - **VERB** VerbForm=Fin Mood=Ind Tense=Imp Aspect=Imp Voice=Act Number=Sing Person=3 Polarity=Pos
- Modern: *S narozením Ježíše Krista to bylo takto:*
  - **VERB** VerbForm=Part Tense=Past Aspect=Imp Voice=Act Number=Sing Gender=Neut Polarity=Pos
- English: “Now the birth of Jesus Christ took place in this way.”

Simple Past Tense: Aorist

- Dresden: *Tehdy oni pověděchu jemu:*
  - **VERB** VerbForm=Fin Mood=Ind Tense=Past Aspect=Perf Voice=Act Number=Plur Person=3 Polarity=Pos Variant=Long
- Modern: *Oni mu řekli:*
  - **VERB** VerbForm=Part Tense=Past Aspect=Perf Voice=Act Number=Plur Gender=Masc Animacy=Anim Polarity=Pos
- English: “They told him,”
Dual Number

- Dresden: ... uzřě dva bratry, ... že biešta rybářě.
  - **NOUN** Gender=Masc Animacy=Anim Number=Dual Case=Nom Polarity=Pos
- Modern: ... uviděl dva bratry, ... byli totiž rybáři.
  - **NOUN** Gender=Masc Animacy=Anim Number=Plur Case=Nom Polarity=Pos
- English: “... he saw two brothers, ... for they were fishermen.”
Morphological Features

Dual Number

• Dresden: … uzřě dva bratry, … že biešta rybářě.
  • **NOUN** Gender=Masc Animacy=Anim **Number=**Dual Case=Nom Polarity=Pos
• Modern: … uviděl dva bratry, … byli totiž rybáři.
  • **NOUN** Gender=Masc Animacy=Anim **Number=**Plur Case=Nom Polarity=Pos
• English: “… he saw two brothers, … for they were fishermen.”

Animacy

• Not significant grammatically as in modern Czech
• But tentatively annotated anyway, to be consistent with modern Czech data
Morphological Features

Converbs (＝ Gerunds ＝ Transgressives)

- Dresden: *Tehdy ona ihned ostavše sieti, jidesta po něm.*
  - **NOUN** VerbForm=Conv Tense=Past Aspect=Perf Voice=Act Number=Dual Polarity=Pos
- Modern: *Oni hned opustili sítě a následovali ho.*
  - **NOUN** VerbForm=Part Tense=Past Aspect=Perf Voice=Act Number=Plur Gender=Masc Animacy=Anim Polarity=Pos
- English: “Immediately they left their nets and followed him.”

Accusative Converbs?

- Dresden: … někteří … neuzřie syna člověčieho, přijdúce v svém království.
  - **VERB** VerbForm=Conv Tense=Pres Aspect=Perf Voice=Act Number=Plur Polarity=Pos Case=Acc
- Modern: … někteří … nespatří Syna člověka přicházejícího ve své královské moci.
  - **ADJ** VerbForm=Part Tense=Pres Aspect=Imp Voice=Act Number=Sing Gender=Masc Animacy=Anim Polarity=Pos Case=Acc
- English: “… some … (will not) see the Son of Man coming in his kingdom.”

Decision: No Case feature with converbs.
Morphological Features

Converbs (= Gerunds = Transgressives)

- Dresden: *Tehdy ona ihned ostavše sieti, jidesta po něm.*
  - NOUN VerbForm=Conv Tense=Past Aspect=Perf Voice=Act Number=Dual Polarity=Pos
- Modern: *Oni hned opustili sítě a následovali ho.*
  - NOUN VerbForm=Part Tense=Past Aspect=Perf Voice=Act Number=Plur Gender=Masc Animacy=Anim Polarity=Pos
- English: “Immediately they left their nets and followed him.”

Accusative Converbs?

- Dresden: … někteří … neuzřie syna člověčieho, přijdúce v svém království.
  - VERB VerbForm=Conv Tense=Pres Aspect=Perf Voice=Act Number=Plur Polarity=Pos Case=Acc
- Modern: … někteří … nespatří Syna člověka přicházejícího ve své královské moci.
  - ADJ VerbForm=Part Tense=Pres Aspect=Imp Voice=Act Number=Sing Gender=Masc Animacy=Anim Polarity=Pos Case=Acc
- English: “… some … (will not) see the Son of Man coming in his kingdom.”

Decision: No Case feature with converbs.
<table>
<thead>
<tr>
<th>ID</th>
<th>FORM</th>
<th>LEMMA</th>
<th>UPOS</th>
<th>XPOS</th>
<th>FEATS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ale</td>
<td>ale</td>
<td>CCONJ</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>Kristovo</td>
<td>Kristův</td>
<td>ADJ</td>
<td>_</td>
<td>Case=Nom</td>
</tr>
<tr>
<td>3</td>
<td>porozenie</td>
<td>porozenie</td>
<td>NOUN</td>
<td>_</td>
<td>Case=Nom</td>
</tr>
<tr>
<td>4</td>
<td>tak</td>
<td>tak</td>
<td>ADV</td>
<td>_</td>
<td>PronType=Dem</td>
</tr>
<tr>
<td>5</td>
<td>biešat</td>
<td>biešat</td>
<td>VERB</td>
<td>_</td>
<td>Aspect=Imp</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>PUNCT</td>
<td>_</td>
<td>_</td>
</tr>
</tbody>
</table>
First Manually Checked Old Czech Sample

- Dresden Bible, Matthew chapters 1–5
- 148 sentences, 2665 words
<table>
<thead>
<tr>
<th>UDPipe 2 Model</th>
<th>PDT 2.6</th>
<th>CAC 2.6</th>
<th>CLTT 2.6</th>
<th>FicTree 2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Modern) Lemma</td>
<td>74.96</td>
<td>74.90</td>
<td>74.63</td>
<td>76.67</td>
</tr>
<tr>
<td>UPOS</td>
<td>91.29</td>
<td>90.69</td>
<td>91.03</td>
<td>90.73</td>
</tr>
<tr>
<td>Features</td>
<td>63.00</td>
<td>62.74</td>
<td>60.38</td>
<td>62.21</td>
</tr>
</tbody>
</table>
### In-domain Tagging Accuracy

<table>
<thead>
<tr>
<th>UDPipe 2 Model</th>
<th>PDT 2.6</th>
<th>CAC 2.6</th>
<th>CLTT 2.6</th>
<th>FicTree 2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Modern) Lemma</td>
<td>99.17</td>
<td>98.95</td>
<td><strong>99.30</strong></td>
<td>99.21</td>
</tr>
<tr>
<td>UPOS</td>
<td>99.30</td>
<td><strong>99.54</strong></td>
<td>99.49</td>
<td>98.69</td>
</tr>
<tr>
<td>Features</td>
<td><strong>97.70</strong></td>
<td>97.07</td>
<td>95.16</td>
<td>96.80</td>
</tr>
</tbody>
</table>
## UDPipe 1.2 Models

<table>
<thead>
<tr>
<th>UDPipe 1.2 Model</th>
<th>PDT 2.5</th>
<th>CAC 2.5</th>
<th>CLTT 2.5</th>
<th>FicTree 2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Modern) Lemma</td>
<td>97.75</td>
<td>96.53</td>
<td>96.05</td>
<td>96.99</td>
</tr>
<tr>
<td>UPOS</td>
<td>98.32</td>
<td>98.15</td>
<td>97.50</td>
<td>97.04</td>
</tr>
<tr>
<td>Features</td>
<td>90.39</td>
<td>86.08</td>
<td>87.40</td>
<td><strong>90.69</strong></td>
</tr>
</tbody>
</table>

Test data from the same treebank but UD 2.10
### UDpipe 1.2 Models

Test data from the same treebank but UD 2.10

<table>
<thead>
<tr>
<th>UDpipe 1.2 Model</th>
<th>PDT 2.5</th>
<th>CAC 2.5</th>
<th>CLTT 2.5</th>
<th>FicTree 2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Modern) Lemma</td>
<td>97.75</td>
<td>96.53</td>
<td>96.05</td>
<td>96.99</td>
</tr>
<tr>
<td>UPOS</td>
<td>98.32</td>
<td>98.15</td>
<td>97.50</td>
<td>97.04</td>
</tr>
<tr>
<td>Features</td>
<td>90.39</td>
<td>86.08</td>
<td>87.40</td>
<td>90.69</td>
</tr>
</tbody>
</table>

Test data from PDT UD 2.10

<table>
<thead>
<tr>
<th>UDpipe 1.2 Model</th>
<th>PDT 2.5</th>
<th>CAC 2.5</th>
<th>CLTT 2.5</th>
<th>FicTree 2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Modern) Lemma</td>
<td>95.00</td>
<td>78.73</td>
<td></td>
<td>90.67</td>
</tr>
<tr>
<td>UPOS</td>
<td>95.98</td>
<td>80.48</td>
<td></td>
<td>90.83</td>
</tr>
<tr>
<td>Features</td>
<td>84.32</td>
<td>60.83</td>
<td></td>
<td>67.68</td>
</tr>
</tbody>
</table>
Split the Manually Checked Sample

- Dresden Bible, Matthew chapters 1–5
  - 148 sentences, 2665 words

- Chapters 1–4 for training
  - 86 sentences, 1669 words

- Chapter 5 for testing
  - 62 sentences, 996 words
### Tagging Chapter 5: UDPipe 1.2 Trained on UD 2.5

<table>
<thead>
<tr>
<th>UDPipe 1.2 Model</th>
<th>PDT 2.5</th>
<th>CAC 2.5</th>
<th>CLTT 2.5</th>
<th>FicTree 2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Modern) Lemma</td>
<td>69.68</td>
<td>68.67</td>
<td>51.20</td>
<td>66.97</td>
</tr>
<tr>
<td>UPOS</td>
<td>76.71</td>
<td>74.00</td>
<td>55.82</td>
<td>70.58</td>
</tr>
<tr>
<td>Features</td>
<td>54.82</td>
<td>52.71</td>
<td>38.55</td>
<td>48.19</td>
</tr>
</tbody>
</table>
Tagging Chapter 5

<table>
<thead>
<tr>
<th>UDPipe 1.2 Model</th>
<th>PDT 2.5</th>
<th>CAC 2.5</th>
<th>CLTT 2.5</th>
<th>FicTree 2.5</th>
<th>BDMt1–4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Modern) Lemma</td>
<td>69.68</td>
<td>68.67</td>
<td>51.20</td>
<td>66.97</td>
<td>67.27</td>
</tr>
<tr>
<td>UPOS</td>
<td>76.71</td>
<td>74.00</td>
<td>55.82</td>
<td>70.58</td>
<td>74.90</td>
</tr>
<tr>
<td>Features</td>
<td>54.82</td>
<td>52.71</td>
<td>38.55</td>
<td>48.19</td>
<td>58.84</td>
</tr>
</tbody>
</table>
## Tagging Chapter 5

<table>
<thead>
<tr>
<th>UDPipe 1.2 Model</th>
<th>PDT 2.5</th>
<th>FicTree 2.5</th>
<th>BDMt1–4</th>
<th>Fic2.10+BDMt</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Modern) Lemma</td>
<td>69.68</td>
<td>66.97</td>
<td>67.27</td>
<td>78.41</td>
</tr>
<tr>
<td>UPOS</td>
<td>76.71</td>
<td>70.58</td>
<td>74.90</td>
<td>85.44</td>
</tr>
<tr>
<td>Features</td>
<td>54.82</td>
<td>48.19</td>
<td>58.84</td>
<td>64.86</td>
</tr>
</tbody>
</table>
Thanks!

Ďakujem!

https://universaldependencies.org/

We are grateful for useful input provided by our colleagues Klára Osolsobě, Olga Navrátilová, Kateřina Granátová, Martina Ježová, Linda Rudenka, Radek Čech, Jana Zdeňková and Ondřej Svoboda