Morphosyntactic 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

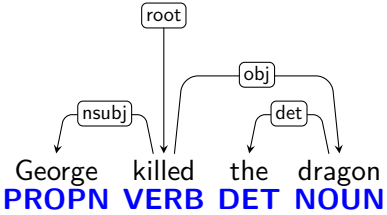
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Universal Dependencies

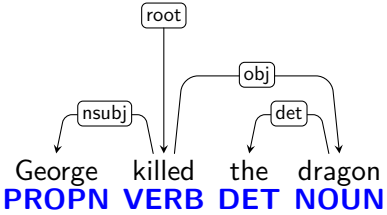
Universal Dependencies

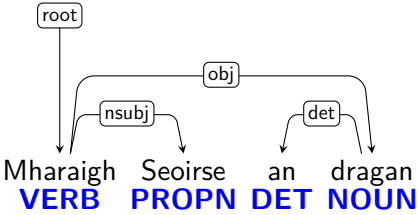
- https://universaldependencies.org/
- Same things annotated same way across languages...
- ... while highlighting different coding strategies

Same Thing Same Way

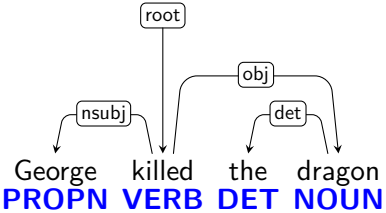


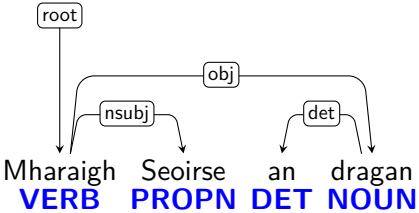
Same Thing Same Way

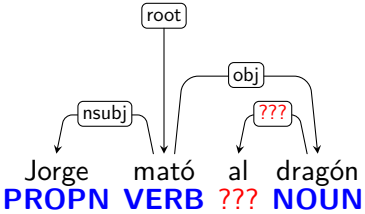




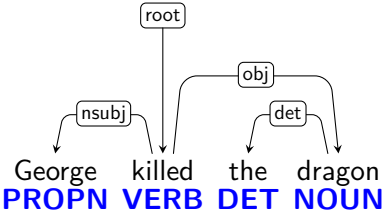
Same Thing Same Way

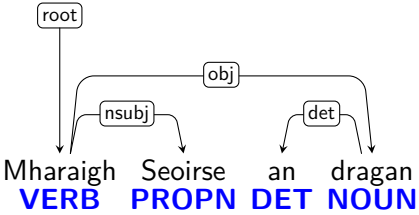


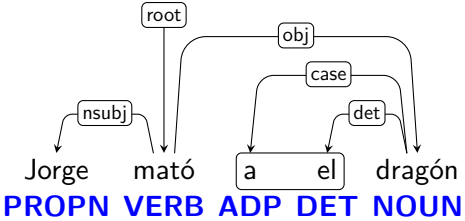




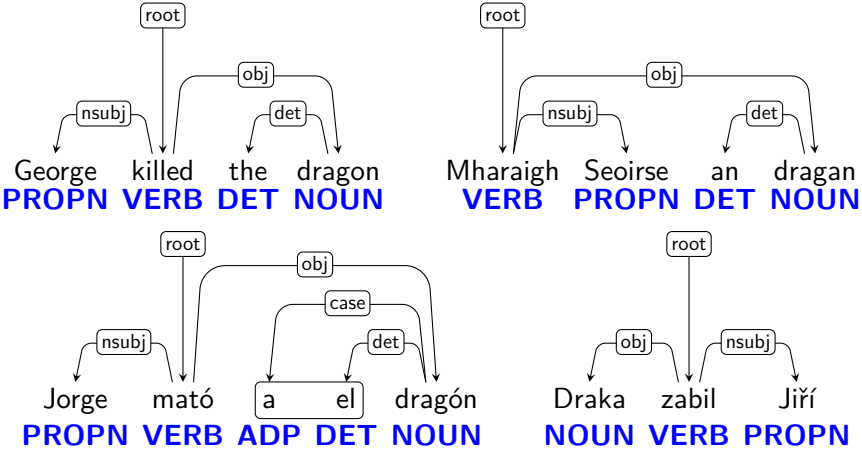
Same Thing Same Way



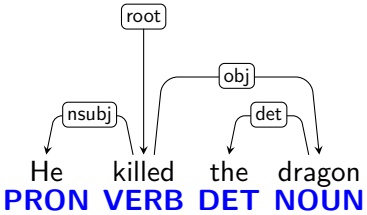




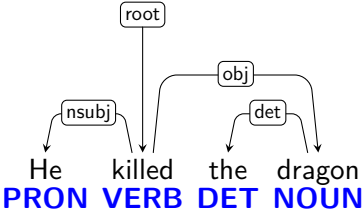
Same Thing Same Way

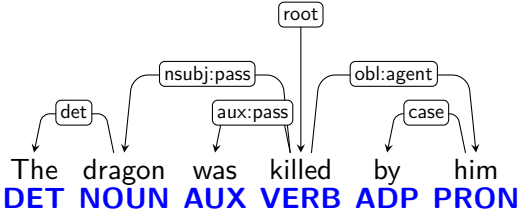


Same Meaning $\neq$ Same Construction!

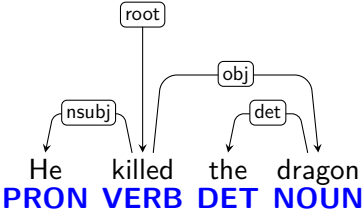


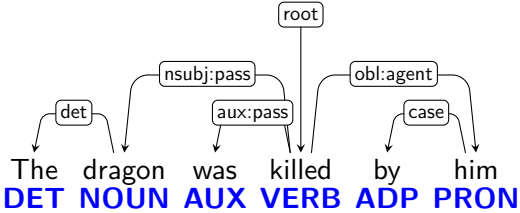
Same Meaning $\neq$ Same Construction!

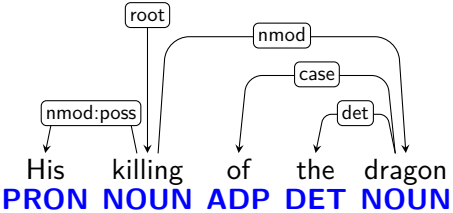




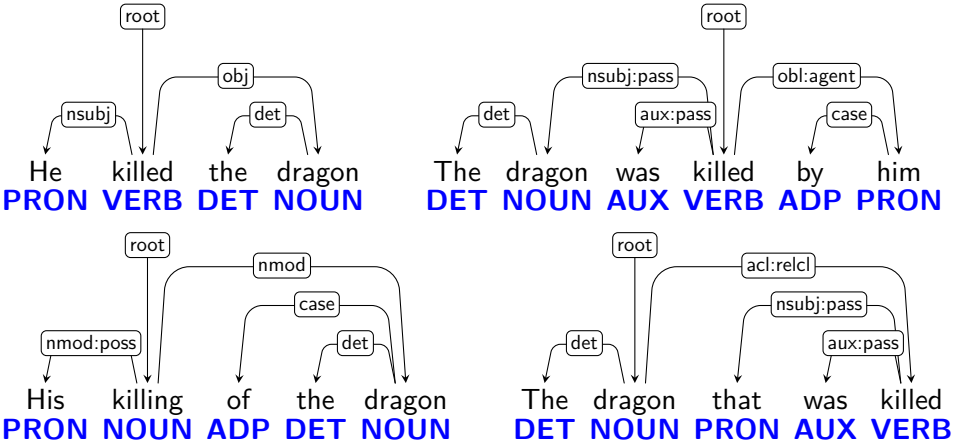
Same Meaning ≠ Same Construction!







Same Meaning ≠ Same Construction!



Morphological Annotation

Le	**chat**	**chasse**	**les**	**chiens**	**.**
le	**chat**	**chasser**	**le**	**chien**	**.**
DET	**NOUN**	**VERB**	**DET**	**NOUN**	**PUNCT**
Definite=Def Gender=Masc Number=Sing	Gender=Masc Number=Sing	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	Definite=Def Gender=Masc Number=Plur	Gender=Masc Number=Plur	

- 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-Azt.:** Nahuatl
- **Mayan:** Kiche
- **Arawakan:** Apurinã
- **Arawan:** Madi
- **Macro-Je:** Xavante
- Bororo
- **Tupian:** Akuntsu, Guajajara, Kaapor, Karo, Makurap, Mundurukú, Nheengatu, Tupinambá, Mbyá, Guaraní, Teko
- **Af.-As.:** Akkadian, ntine, Assyrian, Beja, Coptic, Hebrew

Slavic Languages in UD 2.12

Language	**× 1,000 words**
Belarusian	305
Bulgarian	156
Croatian	199
Czech	2,227
Old Church Slavonic	199
Old East Slavic	333
Polish	499
Pomak	87
Russian	1,832
Serbian	97
Slovak	106
Slovenian	297
Ukrainian	123
Upper Sorbian	11

Czech in UD

Czech UD Treebanks

- PDT (Prague Dependency Treebank)
  - Lidové noviny + Mladá Fronta + ČM Profit + Vesmír, 1993–1994
  - 87K sentences, 1.5M words
- CAC (Czech Academic Corpus / Korpus věcného stylu)
  - non-fiction, 1971–1985
  - 24K sentences, 493K words
- FicTree
  - fiction, from Czech National Corpus, 1991–2007
  - 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
  - ...

Old Czech UD Treebank?

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

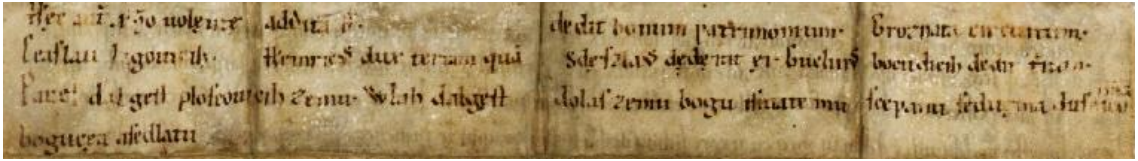
- Bootstrapping:
  - Parse a part using a parser but available models are modern Czech!
  - Manually check and fix
  - Re-train the parser
  - Parse another part
  - Manually check and fix
  - ...

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, přistúpiv)*

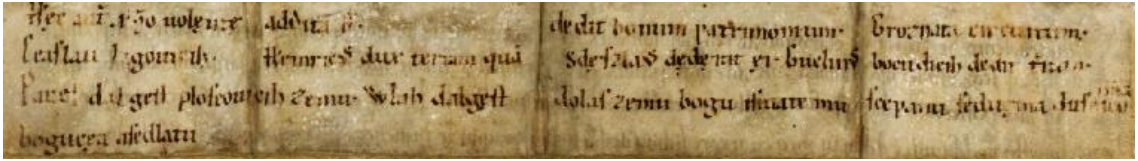
Modern Orthography vs. Modern Language



- *Pauel dal geſt ploſcouicih zemu Wlah dalgeſt dolaſ zemu bogu ıſuıatemu ſcepanu ſeduema duſnıcoma 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."

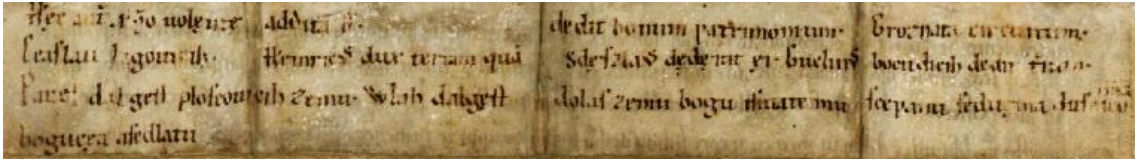
Modern Orthography vs. Modern Language



- *Pauel dal geſt ploſcouicih zemu Wlah dalgeſt dolaſ zemu bogu ıſuıatemu ſcepanu ſeduema duſnıcoma 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."

Modern Orthography vs. Modern Language



- *Pauel dal geſt ploſcouicih zemu Wlah dalgeſt dolaſ zemu bogu ıſuıatemu ſcepanu ſeduema duſnıcoma 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"
  - *Křstitel*
  - *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*, *votsoudici*, *votsoudit*, *votsouditi*, *votsúdici*, *votsúdit*, *votsúditi*
  - "hyperlemma" ⇒ lemma1300 = *otsúditi*

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*, *votsoudici*, *votsoudit*, *votsouditi*, *votsúdici*, *votsúdit*, *votsúditi*
  - "hyperlemma" ⇒ lemma1300 = *otsúditi*
  - (modern) lemma = *odsoudit*

Morphological Features

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."

Morphological Features

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,"

Morphological Features

Dual Number

- Dresden: *... uzřě dva bratry, ... že biešta rybáře.*
  - **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."

Morphological Features

Converbs (= Gerunds = Transgressives)

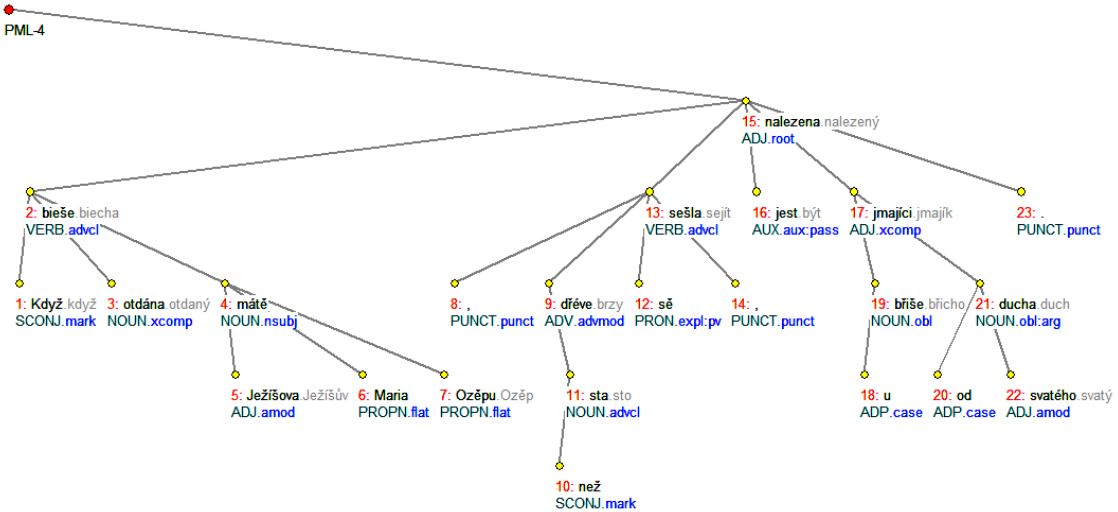
- Dresden: *Tehdy ona ihned ostavše sieti, jidesta po něm.*
  - **NOUN** **VerbForm=Conv Tense=Past Aspect=Perf Voice=Act Number=Dual Polarity=Pos**
- Modern: *Oni hned opustili sítě a následovali ho.*
  - **NOUN** **VerbForm=Part Tense=Past Aspect=Perf Voice=Act Number=Plur Gender=Masc Animacy=Anim 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.

Example Parse (UDPipe 2.0 on UD PDT 2.6)



Example Parse (UDPipe 2.0 on UD PDT 2.6)

ID	FORM	LEMMA	UPOS	XPOS	FEATS
1	Ale	ale	CCONJ	_	_
2	Kristovo	Kristův	ADJ	_	Case=Nom|Gender=Neut|Gender[psor]=Masc|NameType=Sur|Number=Sing|Poss=
3	porozenie	porozenie	NOUN	_	Case=Nom|Gender=Neut|Number=Sing|Polarity=Pos
4	tak	tak	ADV	_	PronType=Dem
5	bieše	biešat	VERB	_	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|Verl
6	.	.	PUNCT	_	_

First Manually Checked Old Czech Sample

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

Tagging Accuracy

UDPipe 2 Model	PDT 2.6	CAC 2.6	CLTT 2.6	FicTree 2.6
(Modern) Lemma	74.96	74.90	74.63	**76.67**
UPOS	**91.29**	90.69	91.03	90.73
Features	**63.00**	62.74	60.38	62.21

(In-domain Tagging Accuracy)

UDPipe 2 Model	PDT 2.6	CAC 2.6	CLTT 2.6	FicTree 2.6
(Modern) Lemma	99.17	98.95	**99.30**	99.21
UPOS	99.30	**99.54**	99.49	98.69
Features	**97.70**	97.07	95.16	96.80

UDPipe 1.2 Models

Test data from the same treebank but UD 2.10

UDPipe 1.2 Model	PDT 2.5	CAC 2.5	CLTT 2.5	FicTree 2.5
(Modern) Lemma	**97.75**	96.53	96.05	96.99
UPOS	**98.32**	98.15	97.50	97.04
Features	90.39	86.08	87.40	**90.69**

UDPipe 1.2 Models

Test data from the same treebank but UD 2.10

UDPipe 1.2 Model	PDT 2.5	CAC 2.5	CLTT 2.5	FicTree 2.5
(Modern) Lemma	**97.75**	96.53	96.05	96.99
UPOS	**98.32**	98.15	97.50	97.04
Features	90.39	86.08	87.40	**90.69**

Test data from PDT UD 2.10

UDPipe 1.2 Model	PDT 2.5	CAC 2.5	CLTT 2.5	FicTree 2.5
(Modern) Lemma		**95.00**	78.73	90.67
UPOS		**95.98**	80.48	90.83
Features		**84.32**	60.83	67.68

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

UDPipe 1.2 Model	PDT 2.5	CAC 2.5	CLTT 2.5	FicTree 2.5
(Modern) Lemma	**69.68**	68.67	51.20	66.97
UPOS	**76.71**	74.00	55.82	70.58
Features	**54.82**	52.71	38.55	48.19

Tagging Chapter 5

UDPipe 1.2 Model	PDT 2.5	CAC 2.5	CLTT 2.5	FicTree 2.5	BDMt1–4
(Modern) Lemma	**69.68**	68.67	51.20	66.97	67.27
UPOS	**76.71**	74.00	55.82	70.58	74.90
Features	**54.82**	52.71	38.55	48.19	58.84

Tagging Chapter 5

UDPipe 1.2 Model	PDT 2.5	FicTree 2.5	BDMt1–4	Fic2.10+BDMt
(Modern) Lemma	69.68	66.97	67.27	**78.41**
UPOS	76.71	70.58	74.90	**85.44**
Features	54.82	48.19	58.84	**64.86**

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