Zpracování staré češtiny s novočeskými modely

Daniel Zeman

18.10.2022

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

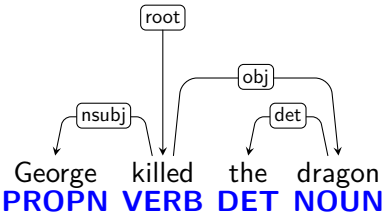
Universal Dependencies

Outline

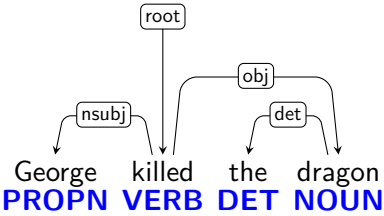
Universal Dependencies

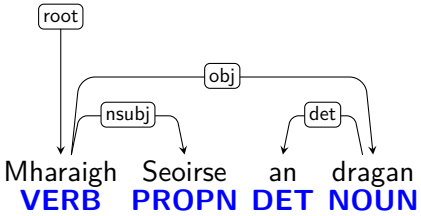
- https://universaldependencies.org/
- Same things annotated same way across languages...
- ... while highlighting different coding strategies

Same Thing Same Way

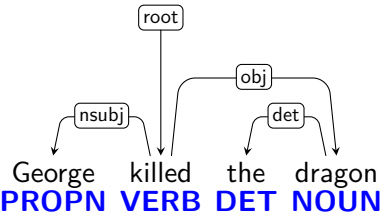


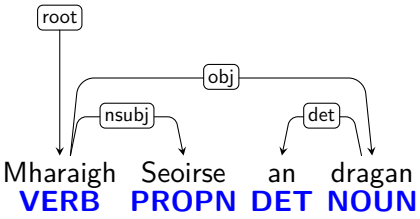
Same Thing Same Way

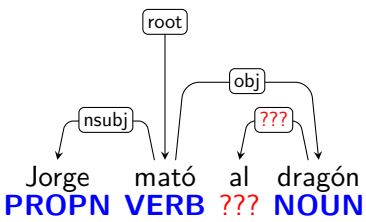




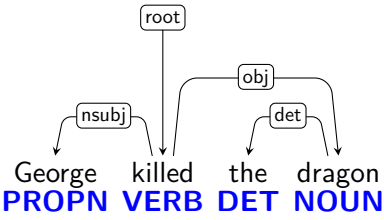
Same Thing Same Way

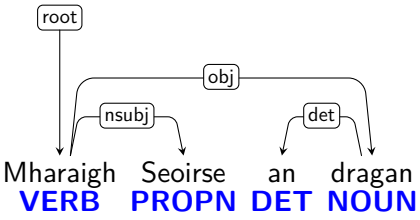


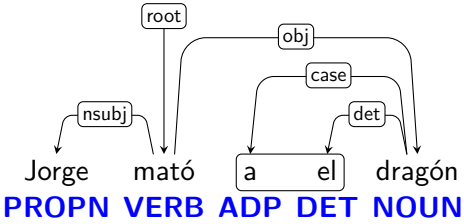




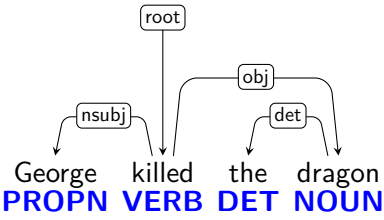
Same Thing Same Way

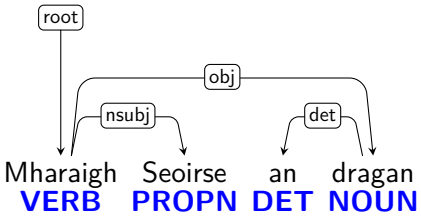


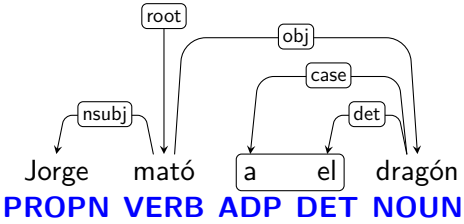


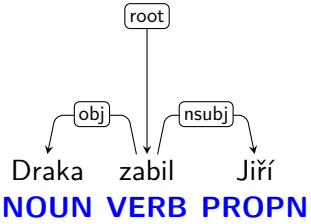


Same Thing Same Way

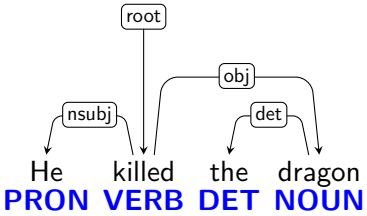




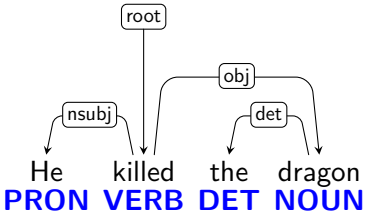


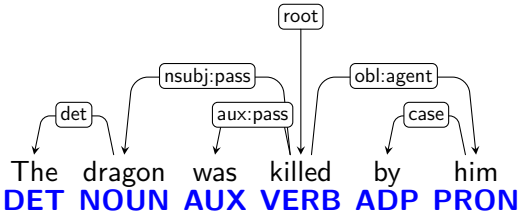


Same Meaning $\neq$ Same Construction!

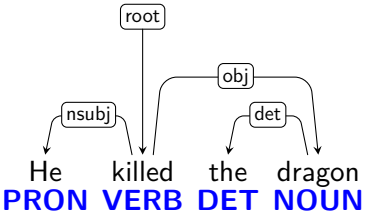


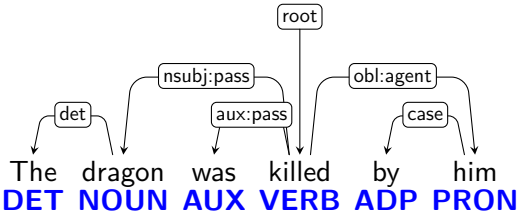
Same Meaning ≠ Same Construction!

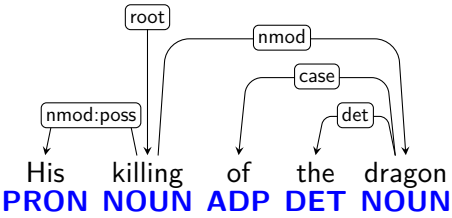




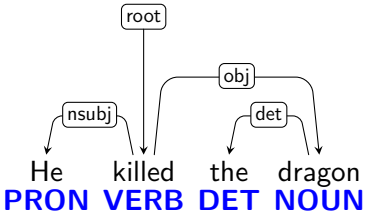
Same Meaning ≠ Same Construction!

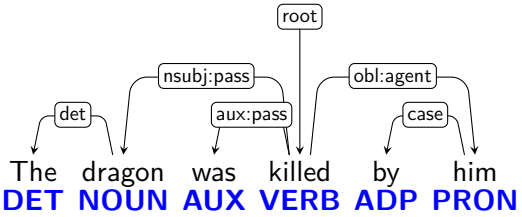


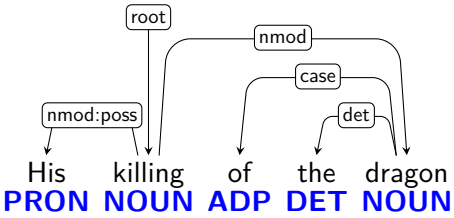


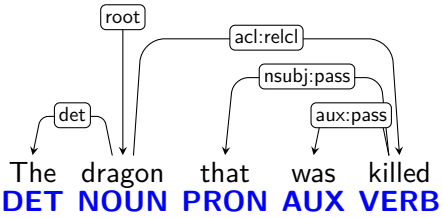


Same Meaning ≠ Same Construction!









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	

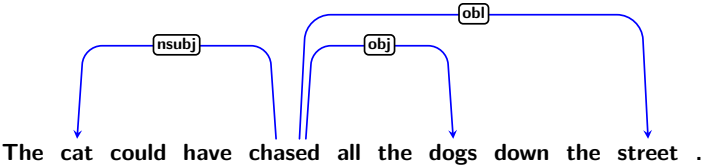
- Lemma representing the semantic content of a word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form

Syntactic Annotation

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

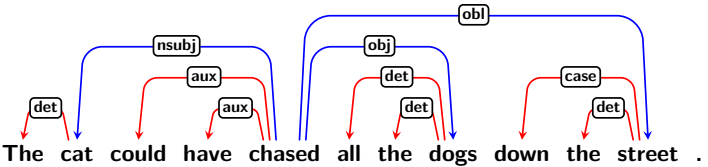
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



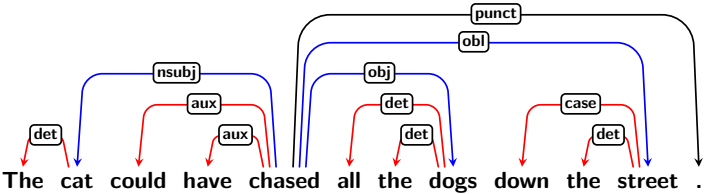
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntactic Annotation



- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

CoNLL-U Format

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	_	_	2	det	_	_
2	chat	chat	NOUN	_	_	3	nsubj	_	_
3	boit	boire	VERB	_	_	0	root	_	_
4-5	du	_	_	_	_	_	_	_	_
4	de	de	ADP	_	_	6	case	_	_
5	le	le	DET	_	_	6	det	_	_
6	lait	lait	NOUN	_	_	3	obj	_	SpaceAfter=No
7	.	.	PUNCT	_	_	3	punct	_	_

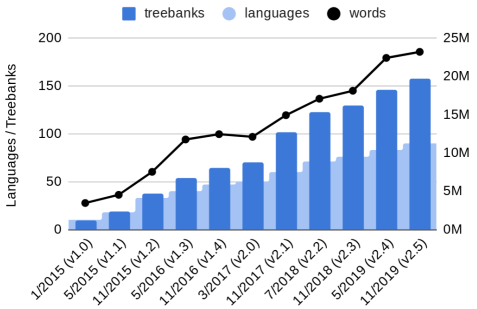
- Revised and extended version of CoNLL-X format
- Two-level segmentation and enhanced dependencies
- CHYBA: V TOMTO PRIKLADU JE “DU” PARTITIV, NIKOLI PREDLOZKA + CLEN, NEMELO BY TO TEDY BYT ROZDELENO!

Basic Universal Dependencies: 130 (128) Languages and Growing

▪ **I.-E.:** Armenian (+West), Greek (+Ancient), Albanian, Hittite, Breton, Irish, 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 ▪ **Dravidian:** Tamil, Telugu ▪ **Uralic:** Erzya, Estonian, Finnish, Hungarian, Karelian, Livvi, Komi Permyak+Zyrian, Moksha, Sámi North+Skolt ▪ **Turkic:** Kazakh, 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 ▪ **Mayan:** Kiche ▪ **Arawakan:** Apurinã ▪ **Arawan:** Madi ▪ **Tupian:** Akuntsu, Guajajara, Kaapor, Karo, Makurap, Mundurukú, Tupinambá, Mbyá, Guaraní, Teko ▪ **Af.-As.:** Akkadian, Amharic, Arabic Standard+Levantine, Assyrian, Beja, Coptic, Hebrew (+Ancient), Maltese ▪ **Niger-Congo:** Yoruba ▪ **Other:** Basque, Sw. Sign, Naija

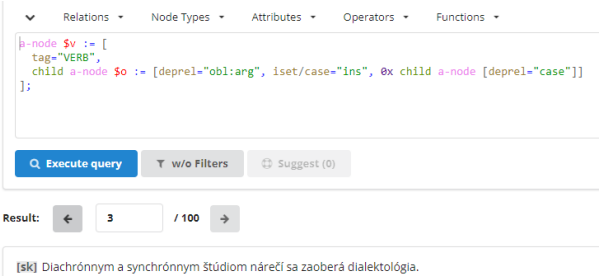
Where are we today?

- Brief history of UD:
  - First guidelines launched in October 2014
  - Treebank releases (roughly) every six months
  - Version 2 guidelines/treebanks in 2016–2017
  - New: guideline amendments since May 2022
  - Extensions: MWEs, PropBanks, Coreference
- UD in numbers:
  - 130 languages
  - 228 treebanks
  - 502 contributors
  - 150,000+ downloads
- Past and current UD events:
  - 4 CoNLL and IWPT shared tasks on UD parsing
  - UD workshops: next in Washington 2023
  - COST action: UniDive (since 2022)
  - Next release in November 2022 (v2.11)



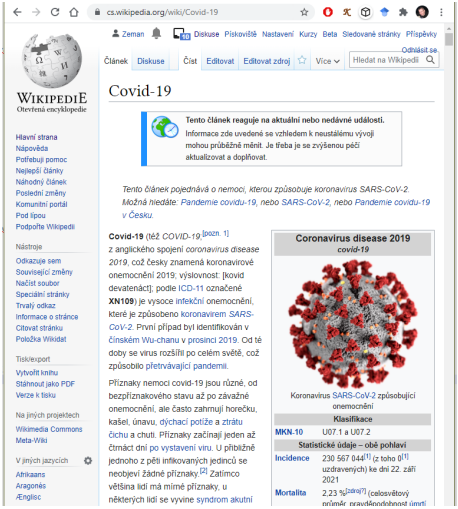
Linguists Can Search Treebanks

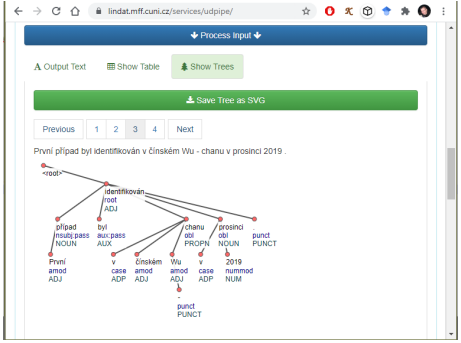
https://lindat.mff.cuni.cz/services/pmltq/



Linguists Can Parse and Search New Data

https://lindat.mff.cuni.cz/services/udpipe/



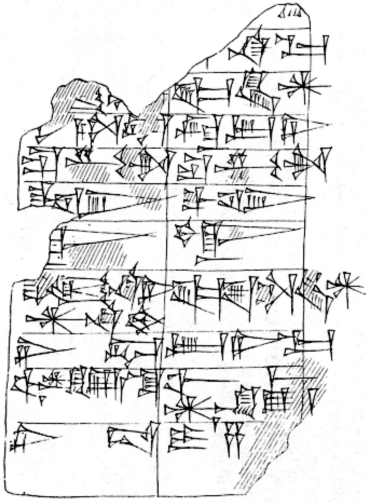


Language Learning

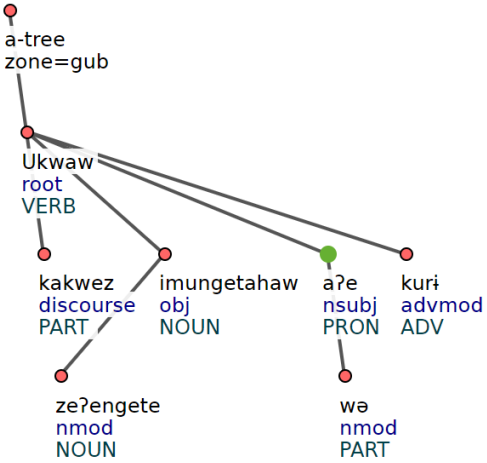
- Check grammar usage in the corpus
- Learner corpora

Historical Linguistics, Classical Languages

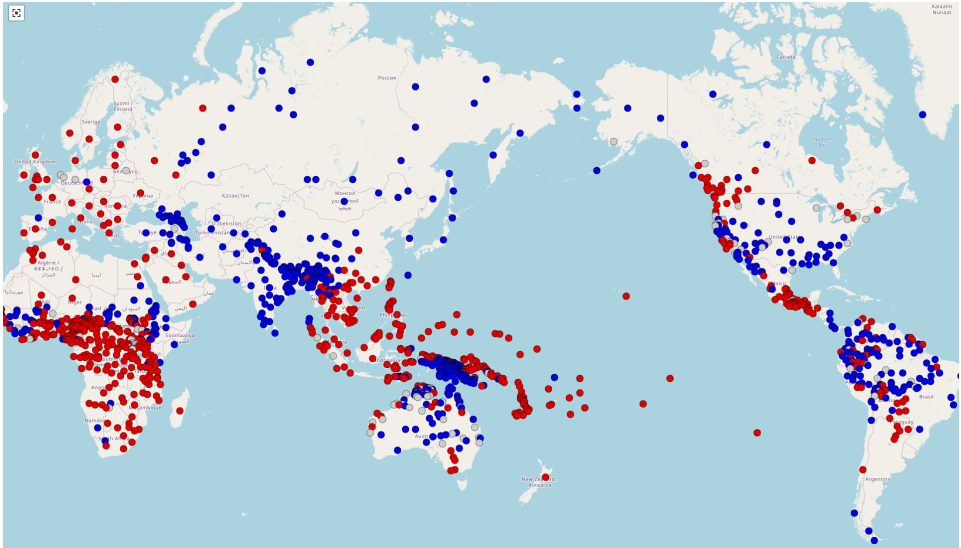
- Old Turkish
- Classical Chinese
- Sanskrit
- Hittite
- Akkadian
- Coptic
- Ancient Hebrew
- Ancient Greek
- Latin
- Old French
- Gothic
- Old Church Slavonic
- Old East Slavic



Documentation of Endangered Languages



Linguistic Typology



Linguistic Typology

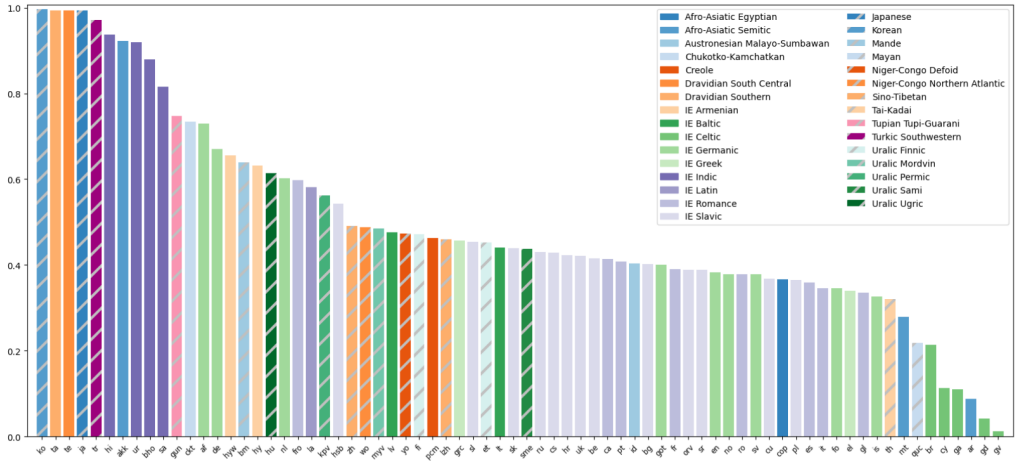


Figure 7 Percentage of head-final dependencies. Each bar is one language.

Czech in UD, Parsing

Outline

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)
  - 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 (with colleagues from MU, Brno, and ÚJČ, Prague)
- Dresden Bible (around 1360)
- Olomouc Bible (1417)
- Gospel of Matthew (from both versions)
  - 2K sentences, 44K words

Old Czech UD Treebank?

- Pilot study (with colleagues from MU, Brno, and ÚJČ, Prague)
- 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?

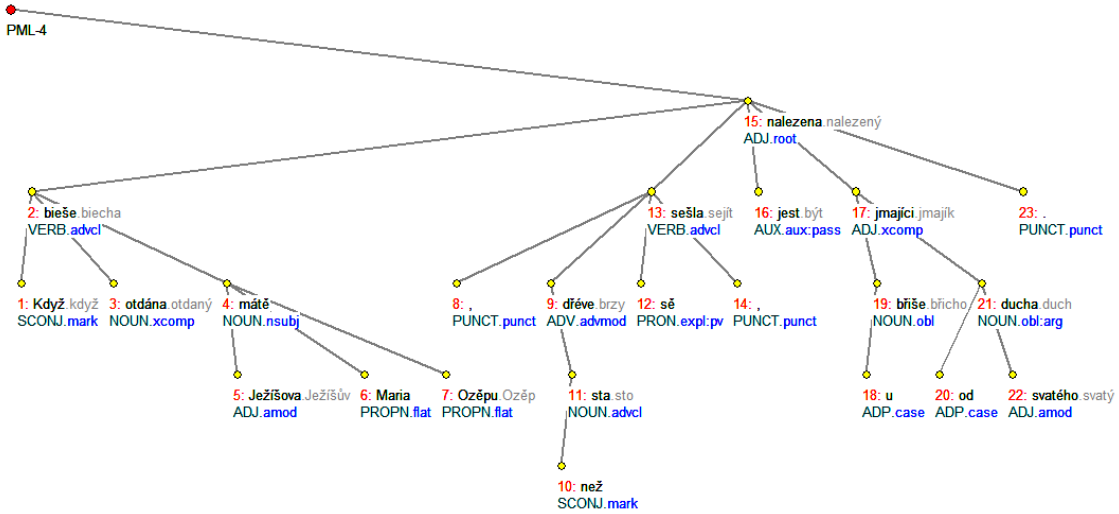
- Pilot study (with colleagues from MU, Brno, and ÚJČ, Prague)
- Dresden Bible (around 1360)
- Olomouc Bible (1417)
- Gospel of Matthew (from both versions)
  - 2K sentences, 44K words

- 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)*

Example Parse (UDPipe 2.0 on UD PDT 2.6)



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**

Děkuji za pozornost!

https://universaldependencies.org/