Morphosyntactic Parser for Old Czech

Lenka Krippnerová and **Daniel Zeman**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics (ÚFAL)
Prague, Czechia
lenka.krippnerova@volny.cz, zeman@ufal.mff.cuni.cz

Relevant UniDive working groups: WG1, WG3

1 Introduction

Old Czech is the oldest phase of Czech documented in written records. It is a language from the mid-12th century to the 16th century. Until the end of the 13th century, however, written records of Old Czech are only fragmentary; we only have glosses and notes. During the Old Czech period, the language underwent many phonological and morphological changes. Old Czech had grammatical features that gradually disappeared from the language during later stages. Morphosyntactic annotations of Old Czech will help to examine the language from a diachronic perspective, especially its diversity over time.

Despite the fact that Modern Czech is very well represented in UD, Old Czech can be considered a low-resource language from the NLP perspective. Searching for examples is hampered by the lack of annotation. Manual morphological annotation in the UD style (de Marneffe et al., 2021) is now being added to selected old texts at the Institute of the Czech Language (ÚJČ).

No syntactic annotation was available so far. We present an adaptation of the UDPipe 2 parser (Straka, 2018) to Old Czech with the help of a data sample that we annotated manually. Furthermore, we compare Modern and Old Czech syntactic constructions and extract distinct patterns with the help of the adapted parser and the treebank analysis tool STARK (Krsnik and Dobrovoljc, 2025).

2 Data

The Old Czech texts we use are transcribed, meaning that they are adapted to the Modern Czech spelling system while respecting the specific features of the historical language. Punctuation is added according to Modern Czech rules. Furthermore, we take advantage of manual morphological annotation (lemma, UPOS, features) carried out recently by ÚJČ. Lemmatization maps old forms to modern lemmata.

3 Manual Syntactic Annotation

We selected two texts from the 14th century: a section of the Gospel of Matthew from the Dresden Bible and the Satires from the Hradec Manuscript. These two texts are very different in genre. While the former is a biblical text, the latter is a work in verse intended for bourgeois audience, criticizing human greed and dishonesty (Opelík et al., 2000). We annotated these texts manually, obtaining a total of 329 sentences (5,437 tokens) with manual morphological and syntactic annotation.

For the manual annotation, we followed the UD guidelines and relied on the annotation practice of the UD treebanks of Modern Czech. However, this was not always possible, as Old Czech differs from Modern Czech in many syntactic constructions.

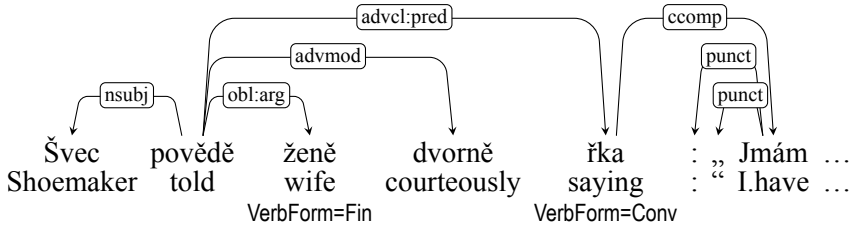
3.1 Old Czech-Specific Constructions

One of the most significant differences is the declension of verbal arguments. In Modern Czech, there are many more objects in the accusative case than there were in Old Czech (Gebauer, 2007, 348–350).

Another significant difference is the introduction of direct speech. In Old Czech, formalized reporting verbs were used very often. In some cases, two verbs of speech occur in one reporting clause—one finite and one converb. We decided to attach the head of direct speech to this formalized verb in the converb form (Figure 1), mainly because the finite verb in the reporting clause is not always the speech verb, and in some cases it would not make sense to make the direct speech its complement.

4 Training the Model

UDPipe 2 (Straka, 2018) is a trainable pipeline for tagging, lemmatization, and syntactic analysis. It uses pre-trained language representations under the hood, specifically for Czech it uses RobeCzech (Straka et al., 2021). We trained UDPipe 2 for Old Czech data. Because our manually annotated corpus contains only 329 sentences,



‘The shoemaker told his wife courteously, saying: “I have …’

Figure 1: Example of syntactic annotation of a sentence with two speech verbs in one reporting clause. Modern Czech equivalent would be *Švec řekl ženě dvorně: „Mám …*

	UPOS	UFeats	Lemmas	UAS	LAS
czech-pdtc-ud-2.17-251125	86.52	65.77	70.99	70.15	61.19
czech-fictree-ud-2.17-251125	90.70	64.21	72.54	72.26	64.03
K-fold cross-validation (Old Czech only)	96.01	78.27	82.07	82.56	76.70
K-fold cross-validation (Old Czech + FicTree)	**97.15**	**84.09**	**87.59**	**86.62**	**82.39**

Table 1: Preliminary parsing results on Old Czech data. The first two rows correspond to existing models trained on Modern Czech: *czech-pdtc-ud-2.17-251125*, trained on 2.7M tokens of news, reviews, nonfiction, academic texts, spoken data, and social media; and *czech-fictree-ud-2.17-251125*, trained on 133K tokens of fiction. The third row shows a model trained only on Old Czech data, while the last row shows a model trained on both Old Czech data and Modern Czech fiction.

we perfomed K-fold cross-validation. For this process, the two manually annotated documents were joined together, and the whole combined text was split just as it was into 10 approximately equally sized parts to be evaluated over 10 iterations. We also finetuned RobeCzech, on 8M tokens of unannotated Old and Middle Czech texts. Preliminary results are shown in Table 1.

5 Periphrastic Verb Forms

Previous work in UniDive WG3 introduced phrase-level features for periphrastic verb forms in UD treebanks, derived automatically from the UD annotation (Krippnerová and Zeman, 2025). The tool currently covers all 15 Slavic languages and 5 Romance languages in UD. We enhanced the tool to Old Czech and applied it to our manual Old Czech annotations. As it relies on common Slavic patterns with little dependence on specifics of individual languages, most of the periphrastic verb forms were annotated correctly without modifying the code. We only had to add a new rule for constructions involving the finite form of the auxiliary verb ‘to be’ and the converb (1) (Večerka, 2017), which had not been encountered in other Slavic languages.

(1)
kdež *budeš* *na* *věky* *vězě*
where you.will.be for ages being.stuck
VerbForm=Fin VerbForm=Conv
‘where you will be stuck forever’

6 Syntactic Development over Time

We compare syntactic structures found in Old Czech parsed with the model we trained with Modern Czech, using STARK (Krsnik and Dobrovoljc, 2025). We have sorted the identified structures in descending order by log ratio, which STARK calculates as the binary logarithm of the ratio of normalized frequencies. The most prominent differences in syntactic structures are shown in Table 2.

Thanks to the enriched annotation of periphrastic verb forms described in Section 5, we were able to compare the development of verb forms as well. The most common differences are as follows. Old Czech uses simple past tenses and converb constructions, or the auxiliary ‘to be’ in the past tense in the third person, which is omitted in Modern Czech.

7 Conclusion

Our contribution is the adaptation of the UD guidelines to Old Czech texts from 14th and 15th centuries, manual annotation of a small sample ac-

Structure	Log ratio
DET >discourse PART	68.54
ADJ >discourse PART	67.37
ADV <mark VERB	66.69
VERB >xcomp PROPN	66.69
PRON <obl:agent VERB	65.96

Table 2: The most prominent differences in two-node syntactic structures between Old Czech texts (the Gospel of Matthew and the Satires from the Hradec Manuscript) and Modern Czech (the UD FicTree treebank). The structures shown are frequent in Old Czech and infrequent in Modern Czech.

cording to the modified guidelines, and demonstrating that UDPipe trained on large Modern Czech data with addition of a bit of Old Czech data performs significantly better than pretrained Modern Czech models. The parser is then combined with rules for periphrastic verb forms and applied to larger quantities of Old Czech data. Finally the STARK tool is used to contrast changing syntactic structures in different stages of the language evolution. We thus put a new language variety on the UD map, but also combine two other lines of research previously developed within UniDive.

Acknowledgements

This work was supported by the COST Action 21167 'UniDive'. The authors also wish to express their gratitude to the Institute of the Czech Language for providing access to their data.

References

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jan Gebauer. 2007. *Historická mluvnice jazyka českého. Díl IV. Skladba (Historical Grammar of the Czech Language. Volume IV. Syntax)*, 2 edition. Academia, Praha.

Lenka Krippnerová and Daniel Zeman. 2025. Periphrastic verb forms in Universal Dependencies. In *Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025)*, pages 140–149, Ljubljana, Slovenia. Association for Computational Linguistics.

Luka Krsnik and Kaja Dobrovoljc. 2025. STARK: A toolkit for dependency (sub)tree extraction and analysis. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 44–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Jiří Opelík, Vladimír Forst, and Luboš Merhaut. 2000. *Lexikon české literatury: osobnosti, díla, instituce. Díl 3, část 2. M-Ř. P-Ř. (Encyclopedia of Czech Literature: Personalities, Works, Institutions. Volume 3, Part 2. M-Ř. P-Ř.)*, 1 edition. Academia, Praha.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

Radoslav Večerka. 2017. PŘECHODNÍK VE STARŠÍ ČEŠTINĚ (Converb in Older Czech). In Petr Karlík, Marek Nekula, and Jana Pleskalová, editors, *CzechEncy - Nový encyklopedický slovník češtiny (New Encyclopedic Dictionary of Czech)*. NLN, Nakladatelství Lidové noviny.