MORPHOSYNTACTIC ANNOTATION IN UNIVERSAL DEPENDENCIES FOR OLD CZECH

DANIEL ZEMAN[1] – PAVEL KOSEK[2]
– MARTIN BŘEZINA[2] – JIŘÍ PERGLER[3]

[1] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague

[2] Department of Czech Language, Faculty of Arts, Masaryk University, Brno

[3] Czech Language Institute, Czech Academy of Sciences, Prague

Abstract: We describe the first steps in preparation of a treebank of 14th-century Czech in the framework of Universal Dependencies. The Dresden and Olomouc versions of the Gospel of Matthew have been selected for this pilot study, which also involves modification of the annotation guidelines for phenomena that occur in Old Czech but not in Modern Czech. We describe some of these modifications in the paper. In addition, we provide some interesting observations about applicability of a Modern Czech parser to the Old Czech data.

Keywords: morphology, dependency syntax, universal dependencies, old czech language

1 INTRODUCTION

Universal Dependencies (UD)[1] (de Marneffe et al. 2021) and (Zeman 2015) is a project aiming at developing morphosyntactic annotation guidelines applicable to all human languages, and creating treebanks annotated according to the guidelines. In the course of the last nine years, UD grew from just 10 languages in release 1.0 to 245 treebanks of 141 languages from 30 families in release 2.12 (May 2023), and is now a de facto standard for morphological and syntactic annotation. With well over 500 contributors, UD has also become a worldwide research community.

Besides covering many modern languages, dialects and genres, UD also contains data for a considerable number of classical languages, such as Akkadian, Sanskrit or Latin, and for some historical varieties, such as Old French and Old East Slavic.

The Slavic genus is particularly well represented, with most languages having a decent UD treebank, and some of them having more than one treebank (Tab. 1). This includes two historical languages that have their own ISO 639 code: Old Church Slavonic and Old East Slavic.

1 https://universaldependencies.org/

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

Tab. 1. Slavic languages in UD 2.12 with treebank sizes (× 1,000 words).

Treebank	**Train**	**Dev**	**Test**
CAC	473	11	11
CLTT	14	11	11
FicTree	134	17	17
PDT	1,173	159	174
PUD			19
Total	**1,794**	**198**	**232**

Tab. 2. Czech treebanks in UD 2.12 with data splits and sizes (× 1,000 words).

In the present paper, we describe the first steps towards creating a UD treebank of the oldest preserved stage of Czech. With five treebanks of various genres and more than two million words (Tab. 2), Czech is the best represented Slavic language in UD, yet all the available data represents the modern language, spanning roughly the period 1971–2016. Morphosyntactic diversity in the Czech treebanks can be largely attributed to different genres; diachronic variation is primarily lexical. In contrast, the oldest Czech texts from around 1300 contain

grammatical features that later vanished from the language. It would be interesting and useful to be able to compare them both with modern Czech and with other languages within a unified annotation framework.

Note that unlike some other historical languages, there is no ISO 639 code for Old Czech,[2] so technically it must be treated as one of the varieties of the single Czech language. This can be also seen as an advantage, as in the future, the continuous development of the language from its earliest stages to modern times can be studied and documented under one set of language-specific guidelines.

A large body of digitized Old Czech texts is available in the diachronic part of the Czech National Corpus[3] and in Old Czech Text Bank (Staročeská textová banka, STB);[4] in its current version, the latter contains almost 7 million tokens. The texts have been transcribed into modernized orthography while preserving their linguistic features. This has the double advantage of making the contents more easily accessible and standardizing the spelling. The transcription was done manually, as it often involves disambiguation where the unsettled orthography did not capture pronunciation unequivocally. For example, what is sometimes considered the oldest preserved Czech sentence,[5] originally spelled as (1), is transcribed as (2). A possible translation to modern Czech is shown in (3):[6]

(1) *Pauel dal geſt ploſcouıcıh zemu Wlah dalgeſt dolaſ zemu bogu ıſuıatemu ſcepanu ſeduema duſnıcoma bogucea aſedlatu*

(2) *Pavel dal jest Ploskovicích zem'u, Vlach dal jest Dolas zem'u bogu i sv'atému Ščepánu se dvěma dušníkoma, Bogučeja a Sedlatu.*

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

(4) 'Pavel gave land in Ploskovice, Vlach gave land in Dolas to God and St. Stephen with two villeins, Bogučej and Sedlata.'

2 Also unlike Old Church Slavonic, which is older (it was used on the Czech territory since the 9th century) but distinct from Czech.

3 https://www.korpus.cz/kontext/query?corpname=diakorp_v6

4 https://korpus.vokabular.ujc.cas.cz/

5 A note from early 13th century on the (much older and written in Latin) charter of the Litoměřice chapter. Státní oblastní archiv v Litoměřicích, fond litoměřické kapituly (Litoměřice, Czechia), sign. R2, 1v. Editor Černá, Alena M. The status as the oldest Czech sentence has been challenged by Dittmann (2012).

6 Precise wording is debatable; the word *dušník* is not used in Modern Czech unless as a historical term. Our goal is to illustrate morphosyntactic and phonological changes.

The texts in the above sources are not annotated any further. Although lemmas, part-of-speech tags and morphological features are available for part of the Old Czech vocabulary, context-based disambiguation of individual tokens is still pending.

2 DATA SELECTION AND PREPROCESSING

For the pilot UD annotation, we chose two versions of the Gospel of Matthew from the oldest Czech translations of the Bible: the "Dresden Bible" (dated ca. 1365) and the "Olomouc Bible" (1417). Besides having two parallel Old Czech translations of the same Latin source, we can also compare the annotated corpus to other treebanks in UD. As many as 16 languages in UD 2.12 contain some biblical material. Out of these, at least[7] 6 contain fragments of the Gospel of Matthew: Ancient Greek, Gothic, Latin, Old Church Slavonic, Romanian, and Yoruba.

The source text is segmented to chapters and verses but not to sentences. A verse often corresponds to a sentence, but sometimes the relation is 1:N or N:1. We used UDPipe 2[8] (Straka 2018), trained on UD Czech PDT 2.6, to obtain an initial tokenization and sentence segmentation. The segmentation was then manually corrected and frozen for any subsequent processing.

The whole corpus (Dresden and Olomouc versions of the Gospel) comprises 2,447 sentences and 44,574 tokens.

Besides tokenization and segmentation, UDPipe also provided initial annotation on the morphological and syntactic layers. Its accuracy is, naturally, relatively low, given the significant differences between the data UDPipe was trained on (news from early 1990s) and the target text. The intended workflow is bootstrapping: After manual correction of the initial segment of the data, the corrected part will be used to re-train the parser, which will then pre-annotate the next segment, hopefully with fewer errors, thus sparing more effort of the human annotators. The process will be repeated until the whole text is annotated and verified by humans. The existing UD annotation guidelines for Czech will be gradually adapted to the specifics of Old Czech during the process. The main adaptation steps, described below,

7 As identified by "Ref=MATT" in the data; this annotation is optional in UD, some other treebanks may thus have verses from Matthew without being counted here.

8 https://ufal.mff.cuni.cz/udpipe/2

have already been done during the first round, nevertheless, further adjustments during the later stages cannot be excluded.

3 LEMMATIZATION

Lemmatization of the old language is a complex issue, as the language was not regulated in any way. Therefore, lemmatization does not only involve picking a canonical form of a lexeme, such as the infinitive of a verb, but sometimes also normalization of several variants of a morpheme (including the stem), e.g., *Křtitel – Křstitel – Krstitel* 'Baptist'.

Moreover, it would be beneficial to be able to match old and modern forms of the same lexeme despite the changes they underwent over the centuries; this can be achieved if the old forms are annotated with the corresponding modern lemma. Not only can a human user take advantage of this, it is also the lemma that a parser with a Modern-Czech model knows. For example, the infinitive in Old Czech typically ends in *-ti* and while this form may occasionally occur in current texts, it is considered archaic, whereas the canonical form (and lemma of the verb) ends in *-t*. A Modern Czech parser may stand a chance of guessing the modernized lemma of Old Czech verbs, but it will never predict the archaic infinitive.

However, there is a downside. When studying Old Czech without comparing it to newer stages, the modernized lemma seems inappropriate and may not be preferred by the users. There are also theoretical questions about which changes count as variants of one lexeme (e.g., if the conjugation class changes, should we say it is a different lexeme, hence a separate lemma?) Some words have fallen out of use and their modern form is not attested, even if we can estimate how it would look like, following phonological evolution of the language. For all these reasons, we maintain two lemmas for each word: the modernized one, and a canonical form as expected around the year 1300. Though in UDPipe experiments, we only evaluate the modern lemma.

4 MORPHOLOGICAL FEATURES

While in Modern Czech the past tense is formed periphrastically, using the l-participle and a finite auxiliary *být* 'be', Old Czech had finite

simple past forms called *imperfect* and *aorist*. These were also attested in Old Church Slavonic and they have survived in a few modern Slavic languages, such as Bulgarian or Upper Sorbian. We thus use morphological features that are used in UD for these languages. For imperfect, we use `Tense=Imp` (we cannot use the `Aspect` feature because it would clash with the lexical aspect that has developed in Slavic languages), for aorist we use `Tense=Past` (together with `VerbForm=Fin`, meaning the finite verb, while the l-participle is tagged `Tense=Past` and `VerbForm=Part`). Sigmatic and asigmatic forms of the aorist are distinguished using the language-specific feature `Variant`.

Other than that, the morphological features already defined for Modern Czech were sufficient, although some of them occur in new combinations. The dual number (`Number=Dual`) is in Modern Czech used only for certain forms, mostly adjectives and nouns related to paired body organs, while in Old Czech it can occur almost everywhere where singular or plural can.

Animacy of masculine nouns in the 14th century did not yet play the role it plays in the grammar today, yet we tentatively annotate it to stay consistent with Modern Czech datasets.

Present (simultaneous) and past (anterior) converbs (`VerbForm=Conv`; also called gerunds) still exist in modern data, although they are very rare and archaic; in Old Czech their frequency is much higher but the annotation is analogous. However, the neuter singular form, which nowadays concurs with feminines, was identical to masculines in Old Czech. We also briefly considered adding the `Case` feature, as there are claims (Gebauer 1898, p. 83, § 35) that some of the converb forms correspond to the accusative, but we abandoned the idea both for consistency with modern data and for inability to reliably assign case values to some of the forms.

5 TAGGING RESULTS

In this section we report on the accuracy of the UDPipe models in the initial stages of the project. First of all, it is important to note that there are currently two versions of the UDPipe tool available: 1.2 and 2.0. Version 2.0 is known to perform significantly better; however, it is only available as a web service with pre-trained models. If we want to train a model on our own data, we have to downgrade to UDPipe 1.2, which is

available as a standalone, trainable tool (but there are pre-trained models for it as well).

Currently available pre-trained models for UDPipe 1.2 are based on UD release 2.5; for UDPipe 2.0, one can choose between UD releases 2.6 and 2.10.[9] The differences between the UD releases should not be large and the size of the training data should be stable; nevertheless, there may be corrections of annotation errors, meaning that the resulting models are not the same. UDPipe did not have access to any dedicated morphological dictionary in our experiments, only to the training data.

Pre-trained models correspond to the four Czech UD treebanks that have designated training data (see Tab. 2). All four are automatic conversions from the PDT annotation style. Besides varying sizes of the treebanks (bigger is better), the genre also plays a role. None of the Czech treebanks contains biblical texts (which would be the best match, even if in a modern variety of the language). The largest treebank, PDT (Hajič et al. 2020), is based on newspapers and journals from early 1990s. CAC (Vidová Hladká et al. 2008) is non-fiction from 1971–1985. CLTT (Kríž and Hladká 2018) is small and focused on the legal domain, containing the Accounting Act. Finally, FicTree (Jelínek 2017) contains fiction from 1991–2007; genre-wise, this treebank is probably most similar to our target data.

There are various sources of divergence between the trained models and the Old Czech data. First, the vocabulary is quite different in the news (or any other training genre) vs. in the Bible. Even parsing a Modern Czech Bible translation would be difficult because of this. Second, as mentioned above, some of the old words have fallen out of use and the parser cannot know them. Third, for words that survived to modern days, even though the old orthography is cleaned, transcribed and unified, many word forms still differ from their modern counterparts because their old pronunciation differs: *sě* vs. *se* ‘oneself’ (the reflexive marker), *viece* vs. *více* ‘more’ etc. Fourth, the morphological differences described above mean that some old forms do not exist any more (imperfect such as *bieše* ‘was’; aorist such as *vece* ‘said’, *jide* ‘went’; most forms of the dual) or are much less frequent than in Old Czech (converbs such as *řka* ‘saying’, *přistúpiv* ‘having approached’).

9 Only models on UD 2.6 were available at the time when we ran UDPipe 2.0.

We have manually verified the first five chapters of the Dresden version of Matthew, which amounts[10] to 148 sentences and 2,665 words. This data can be used to evaluate the accuracy of the initial model, trained only on Modern Czech. All five chapters can be used when training a model to preprocess the next segment of the corpus. However, if we want to evaluate a re-trained model on gold-standard Old Czech data, we need to split the dataset into training and test part. In such experiments, we reserve chapters 1–4 (86 sentences, 1,669 words) for training and chapter 5 (62 sentences, 996 words) for testing.

Since the present annotation is manually checked only at the morphological layer, we report accuracy of lemmatization and tagging (separately the main part-of-speech category and the remaining morphological features); we do not study the accuracy of dependency relations yet.

UDPipe 2 Model	**PDT**	**CAC**	**CLTT**	**FicTree**
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

Tab. 3. For comparison, we show in-domain accuracy of UD 2.6 pre-trained models. Each model is evaluated on the test data from the corresponding treebank.

UDPipe 1.2 Model	**PDT**	**CAC**	**CLTT**	**FicTree**
(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

Tab. 4. Tagging Dresden Matthew chapter 5 (that is, all results are on the same test set). UDPipe 1.2 models pre-trained on UD 2.5.

10 Chapter 1 is not complete, the genealogy in the beginning was omitted.

UDPipe 1.2 Model	PDT	FicTree	DMt1–4	Fic+Mt
(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**

Tab. 5. Tagging Dresden Matthew chapter 5 (that is, all results are on the same test set). First two columns are UDPipe 1.2 models pre-trained on UD 2.5, repeated from Tab. 4. The third column is a model trained on Dresden Matthew chapters 1–4. The last column is a combined model trained on chapters 1–4 and on FicTree from UD 2.10.

The lemmatization and tagging scores on chapter 5 of the Gospel of Matthew from the Dresden Bible are shown in Tab. 4 and 5. For comparison, Tab. 3 shows what scores one can expect when the pre-trained models are applied to Modern Czech data. Note however that these numbers are not directly comparable even among themselves, as each model was evaluated on different test set (namely on the test set from the treebank from which the model's training data were taken).

The DMt1–4 column in Tab. 5 demonstrates how important it is to train on data from the same domain and same stage of the language: Despite the fact that the training data is ridiculously small (less than 2K words), the results are not much worse (and for Features they are even better) than the Modern Czech model trained on over 1M words from PDT.

When combined with a larger Modern Czech corpus, the four chapters of Matthew provide for a model that is much better than any of the other models in isolation (a jump of 9–10 percent points). The results also proved that FicTree is, out of the Modern Czech treebanks, the best fit for our biblical data; when we combined Matthew with PDT, which is ten times larger than FicTree, the negative effect of the out-of-domain data prevailed and the scores were worse than with FicTree.

6 CONCLUSION

We have described the initial steps in order to create a UD-style annotated treebank of Old Czech biblical texts. Some peculiarities of the historical language were discussed, a bootstrapping approach with a Modern Czech parser was proposed and first experiments evaluated. In the next phase we will address the syntactic layer. As soon as the

syntactic annotation is ready, we intend to publish the treebank in a future release of Universal Dependencies.

ACKNOWLEDGMENTS

This work was supported by the grants 20-16819X (LUSyD) of the Czech Science Foundation (GAČR), and The Grammar and Lexicon of Czech III – 2023 (MUNI/A/1249/2022).

The work uses data and tools provided by the research infrastructure LINDAT/CLARIAH-CZ (https://lindat.cz/), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project no. LM2023062).

The authors are grateful for input provided by the colleagues from Masaryk University and the Czech Language Institute, in particular Klára Osolsobě, Olga Navrátilová, Kateřina Granátová, Martina Ježová, Linda Rudenka, Radek Čech, Jana Zdeňková and Ondřej Svoboda.

References

de Marneffe, M.-C., Manning, C., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), pages 255–308.

Dittmann, R. (2012). Problém tzv. nejstarší české věty. Bohemica Olomucensia, 4(1), pages 26–36. Available at: https://bohemica.actavia.cz/pdfs/boh/2012/01/03.pdf.

Gebauer, J. (1898). Historická mluvnice jazyka českého. Díl III., II. Tvarosloví. Časování. Wien: F. Tempský. 508 pages. Available at: https://kramerius5.nkp.cz/view/uuid:223066b0-6f7b-11eb-9f97-005056827e51?page=uuid:2bd40f78-a469-43e9-bc4b-2b579e00865c.

Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánek, J., and Štěpánková, B. (2020). Prague Dependency Treebank – Consolidated 1.0. In: Proceedings of LREC., pages 5208–5218.

Jelínek, T. (2017). FicTree: a Manually Annotated Treebank of Czech Fiction. In: Proceedings of ITAT, pages 181–185.

Kríž, V. and Hladká, B. (2018). Czech Legal Text Treebank 2.0. In: Proceedings of LREC, pages 4501–4505.

Staročeská textová banka [online]. (2006–2023). Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. Verze dat 1.1.22 [cit. 26. 3. 2023]. Available at: https://vokabular.ujc.cas.cz/banka.aspx?idz=STB.

Straka, M. (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. Bruxelles: Association for Computational Linguistics, pages 197–207.

Vidová Hladká, B., Hajič, J., Hana, J., Hlaváčová, J., Mírovský, J., and Raab, J. (2008). The Czech Academic Corpus 2.0 Guide. In: PBML 89, pages 41–96.

Vokabulář webový: webové hnízdo pramenů k poznání historické češtiny [online]. (2006–2023) Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. Verze dat 1.1.22 [cit. 26. 3. 2023]. Available at: https://vokabular.ujc.cas.cz/.

Zeman, D. (2015). Slavic Languages in Universal Dependencies. In Natural Language Processing, Corpus Linguistics, E-learning, pages 151–163, Lüdenscheid: RAM-Verlag.