Universal dependencies and non-native Czech

Jirka Hana, Barbora Hladká

Charles University, Malostranské nám. 25, 118 00 Prague 1, Czech Republic

{hana,hladka}@ufal.mff.cuni.cz

ABSTRACT

CzeSL is a learner corpus of texts produced by non-native speakers of Czech. Such corpora are a great source of information about specific features of learners' language, helping language teachers and researchers in the area of second language acquisition. In our project, we have focused on syntactic annotation of the non-native text within the framework of Universal Dependencies. As far as we know, this is a first project annotating a richly inflectional non-native language. Our ideal goal has been to annotate according to the non-native grammar in the mind of the author, not according to the standard grammar. However, this brings many challenges. First, we do not have enough data to get reliable insights into the grammar of each author. Second, many phenomena are far more complicated than they are in native languages. We believe that the most important result of this project is not the actual annotation, but the guidelines and principles that can be used as a basis for other non-native languages.

KEYWORDS: learner corpus, second language, syntax annotation, universal dependencies, second language acquisition.

1 Introduction

Universal Dependencies (UD) is a unified approach to grammatical annotation that is consistent across languages.[1] It facilitates both linguistic and NLP research. However, the absolute majority of these treebanks are based on corpora of standard language. In this paper, we describe a project of creating a syntactically annotated corpus of learner Czech, the CzeSL corpus. The choice of Universal Dependencies as the annotation standard was relatively straightforward. It is an established framework used for more than 100 treebanks in 60 languages (including two other learner corpora). The common guidelines make the data easily accessible to a large audience of researchers and comparable across languages. Also, following the UD schema and format makes it easier to train and test NLP tools on the basis of our annotation.

CzeSL (Hana et al., 2010), (Rosen et al., 2014) is a learner corpus of texts produced by non-native speakers of Czech.[2] Such corpora are a great source of information about specific features of learners' language, helping language teachers and researchers in the area of second language acquisition. Each sentence in the CzeSL corpus has an error annotation and a target hypothesis with its morphological and syntactic annotation. However, there is no linguistic annotation of the original text. This means we can see what grammatical constructions the authors should have used but not what they actually used. And we can analyze their grammar only indirectly via the error annotation. Therefore we have focused on syntactic annotation of the non-native text within the framework of UD. Figure 1 shows a UD tree structure for (1) selected from CzeSL.

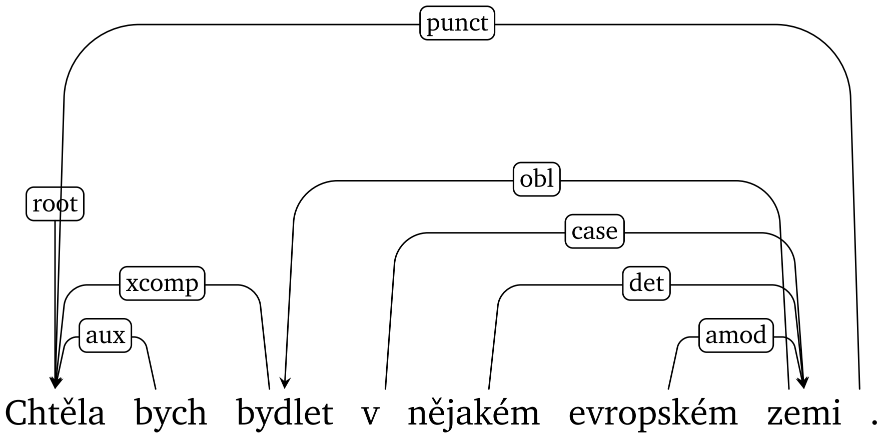

Figure 1: A sample UD tree

(1) Chtěla bych bydlet v nějakém evropském zemi .
I-liked would to-live in some$_{masc}$ European$_{masc}$ country$_{fem}$.
'I would like to live in some European country.'

The remainder of this paper is divided into five sections. In Section 2, we provide an overview of works related to the UD annotation of learner corpora. Section 3 gives a general description of the CzeSL corpus that we annotate in the UD framework. Section 4 presents a core part of the paper. We formulate our annotation principles and describe the challenges that we meet while applying the existing guidelines. The annotation procedure itself is presented in Section 5. In Setion 6, we provide the conclusions.

[1] http://universaldependencies.org

[2] http://utkl.ff.cuni.cz/learncorp/

Corpus	Language	Size (annotated)	
TLE (Berzak et al., 2016)	English	5,124 sentences	97,681 words
REALEC (Kuzmenko and Kutuzov, 2014)	English	373 sentences	7,196 words
Tweebank (Liu et al., 2018)	English	3,550 tweets	45,607 words
CFL (Lee et al., 2017)	Chinese	451 sentences	7,256 words

Table 1: Relevant UD-annotated Corpora

2 Related work

The great majority of currently available UD treebanks were converted from already existing treebanks annotated using a different annotation scheme, Moreover, these corpora contain texts written completely by native speakers. The UD annotation of learner corpora has been initiated later on.

Table 1 summarizes UD-annotated corpora relevant for our task. The Treebank of Learner English (TLE) contains manually annotated POS tags and UD trees for sentences selected from the Cambridge First Certificate in English learner corpus (Yannakoudakis et al., 2011). The REALEC corpus is a collection of English texts written by Russian-speaking students. Unlike TLE, the REALEC sample was first automatically annotated by the UDPipe pipeline (Straka and Straková, 2017) and then manually corrected. (Lyashevskaya and Panteleeva, 2018) analyzed the errors made by the parser that originate from differences between English and Russian, typologically different languages. (Lee et al., 2017) have adapted existing UD guidelines for Mandarin Chinese to annotate learner Chinese texts. As an annotation workbench, they used essays written by Chinese language learners representing 46 different mother tongue languages (Lee et al., 2016). As far as we know, there is no similar project for a richly inflected language. We include Tweebank, a twitter corpus, because even though it is not a corpus of non-native language, it brings similar challenges. The wordings and language style used in tweets are often far from the straightforward and well researched syntactic constructions used by the news corpora.

Among UD languages, Czech has an exceptional status because of the greatest number of Czech sentences annotated in UD, namely 127 thousand sentences included in 5 treebanks. Most of the Czech UD treebanks were originally annotated according to the Prague Dependency Treebank annotation scheme[3] and then transformed into UD. The only treebank annotated from scratch is the Czech part[4] of the Parallel Universal Dependencies treebanks created for the CoNLL 2017 shared task (Zeman et al., 2017). Czech holds its exceptional status among the UD treebanks of Slavic languages as well, see (Lasota, Brielen Madureira, 2018). (Zeman, 2015) focuses on a few morphological and syntactic phenomena that occur in Slavic languages and their treatment

[3] https://ufal.mff.cuni.cz/prague-dependency-treebank

[4] https://github.com/UniversalDependencies/UD_Czech-PUD/blob/master/cs_pud-ud-test.conllu

in UD.

The task of corpus annotation deals with a fundamental issue of consistent annotation of the same phenomena within and across corpora. (de Marneffe et al., 2017) assessed the consistency within the Universal Dependency Corpora of English, French, and Finnish by checking dependency labels of words occurring in the same context. (Ragheb and Dickinson, 2013) reports on a study of inter-annotator agreement for a dependency annotation scheme designed for learner English.

3 The CzeSL corpus

The whole CzeSL corpus contains about 1.1 million tokens in 8,600 documents and is compiled from texts written by students of Czech as a second or foreign language at all levels of proficiency. CzeSL-MAN is a subset of CzeSL, manually annotated for errors.[5] It consists of 128 thousand tokens in 645 documents written by native speakers of 32 different languages. In the rest of this paper, when we refer to CzeSL, we refer to CzeSL-MAN. Each CzeSL document is accompanied with:

- metadata – information about the native language of the author, length of study, type of task, etc.
- error annotation (see below)
- linguistic annotation of the target hypothesis

The CzeSL error annotation consists of three tiers:

- Tier 0 (T0): an anonymized transcript of the hand-written original with some properties of the manuscript preserved (variants, illegible strings),
- Tier 1 (T1): forms that are incorrect in isolation are fixed. The result is a string consisting of correct Czech forms, even though the sentence may not be correct as a whole
- Tier 2 (T2): the remaining error types (valency, agreement, word order, etc.), i.e. this is the target hypothesis.

Links between the tiers allow capturing errors in word order and complex discontinuous expressions. Errors are not only corrected, but also classified according to a taxonomy. As an example consider (2) – showing the original text (T0) and the target hypothesis (T2). The full error analysis, including error tags is in Figure 2

(2)	T0:	Myslím		že	kdy	by	byl	se	svím	dítem	…
	T2:	Myslím	,	že	kdybych		byl	se	svým	dítětem	…
		think$_{1sg}$	,	that	if-would$_{1sg}$		was$_{masc}$	with	my$_{neut.sg.inst}$	child$_{neut.sg.inst}$	…

’I think that if I were with my child …’

[5] https://bitbucket.org/czesl/czesl-man/

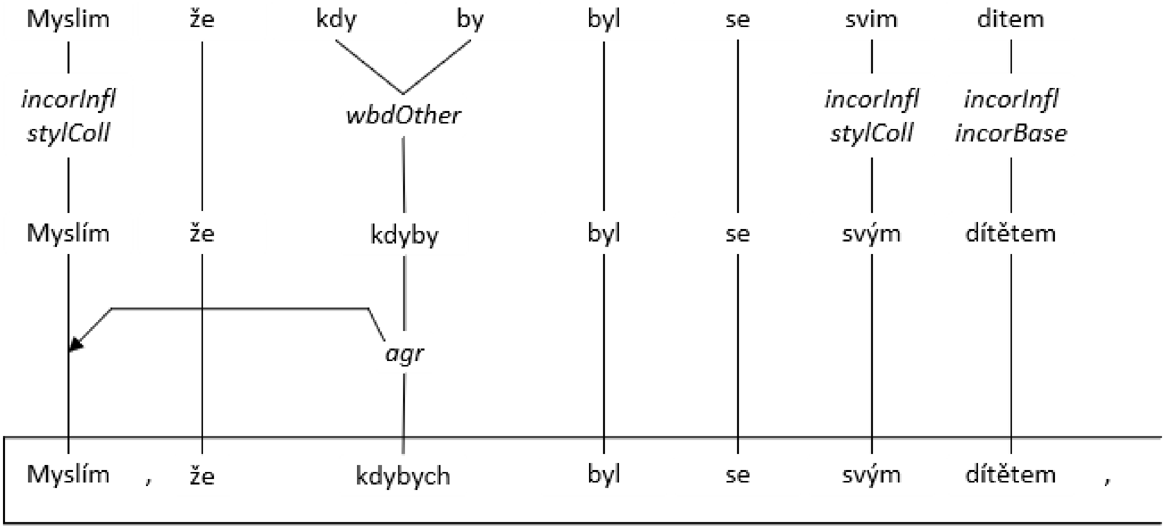

Figure 2: Error annotation of a sample sentence in (2)

Annotation of this kind is supplemented by a formal classification, e.g. an error in morphology can also be specified as being manifested by a missing diacritic or a wrong consonant change. The annotation scheme was tested in two rounds, each time on a doubly-annotated sample – first on a pilot annotation of approx. 10,000 words and later on nearly half of all the data, both with fair inter-annotator agreement results. Error annotation of this kind is a challenging task, even more so for a language such as Czech, with its rich inflection, derivation, agreement, and a largely information structure-driven constituent order.

In addition to error annotation, the target hypothesis is annotated linguistically: for morphology and syntax. However, as mentioned above, there is no linguistic annotation of the original text, a gap we are in a process of filling.

4 Approach

Similarly as the projects above, we follow the basic annotation principle of the SALLE project (Dickinson and Ragheb, 2013), and attempt to annotate literally: we annotate the sentences as they are written, not as they should be. In other words, our ideal goal is to annotate according to the non-native grammar in the mind of the author (i.e. the grammar of their interlanguage), not according to the standard grammar.

However, this brings several challenges. First, in many cases, we do not have enough data to get reliable insights into the grammar of each author. Second, many phenomena are far more complicated than they are in native languages. Our annotation principles include:

- When form and function clash, form is considered less important. For example, if a word functions as an adjective, we annotate it as an adjective even if it has a verbal ending.

- When lacking information, we make conservative statements.

- We focus on syntactic structure and the most important grammatical functions, annotating unclear functions with an underspecified label.

4.1 Tokenization

There is an established tokenization used by Czech UD corpora that builds on the general UD tokenization rules. However, we used the original CzeSL tokenization to make the UD structures compatibles with its error annotation. The differences affect mostly alternatives offered by the author due to their uncertainty (e.g. *b(y/i)l* is considered one token), hyphenated words and certain numerical expressions.

4.2 Part-of-speech and Morphology

Czech, as other Slavic languages, is richly inflected. It has 7 cases, 4 genders, colloquial variants, etc. Therefore, corpora of standard Czech are usually annotated with detailed morphological tags (for example, the tagset used for the Prague Dependency Treebank has 4000+ tags, distinguishing roughly 12 different categories). We have decided not to perform such annotation. There are several reasons, for this decision, mainly:

- many endings are homonymous; therefore it is not obvious which form was used if we wanted to annotated according to the form. For example, the ending -a has more than 10 different morphological functions depending on the paradigm.

- these complications do not always correlate with understandability. Some texts are easy to understand yet, they use wrong or non-existing suffixes, mix morphological paradigms etc.

- the corpus can be still searched for pedagogical reasons: the intended morphological tag can be derived from the corresponding target hypothesis, the error annotation marks mistakes in inflection and the original forms can be matched existing standard forms

Instead, we have limited ourselves to the Universal POS Tagset (Petrov et al., 2011). When form and function clash, form is considered less important. For example, if a word functions as an adjective, we annotate it as an adjective even if it has a verbal ending.

One of the common deviating characteristics of learner Czech was the neutralization between adjectives and adverbs. In (3), the adjective *rychlé* 'quick' is used instead of the correct adverb *rychle* 'quickly'.

(3)	T0:	Kvalita	života	by	se	zlepšila	moc	rychlé.
	T2:	Kvalita	života	by	se	zlepšila	moc	rychle.
		Quality	of-life	would	refl	improve	too	quick(ly)

'Life quality would improve too quickly.'

This is similar to German or colloquial English. Unfortunately, UPOS force us to choose between adjectives and adverbs even for speakers who clearly use the same word for both. We annotate such words as adjectives with an additional note.

4.3 Lemmata

Ideally, we would use lemmata from the author's interlanguage. For example, in (4), we would use the lemma *Praga* (correctly *Praha*). The situation is clear, because the word is in the lemma form already (nominative singular). Often knowing the native language of the author helps – for example, in (5) the lemma of *krasivaja* is *krasivyj*, based on Russian.

(4) T0: Praga je hezké město. → lemma: Praga
T2: Praha je hezké město. → lemma: Praha
Prague is nice city
'Prague is a nice city.'

(5) T0: Praga je krasivaja. → lemma: krasivyj
T2: Praha je krásná. → lemma: krásný
Prague is beautiful
'Prague is beautiful.'

Sometimes we can see that the author declines a word using a paradigm of another word. For example, for *večeřem* 'dinner$_{inst}$ in (6) we can hypothesize the masculine lemma *večeř*, formed in analogy with the word *oběd* – *obědem* 'lunch'. The correct forms are feminine *večeře* – *večeří*.

(6) T0: Začíname večeřem. → lemma: večeř
T2: Začíname večeří. → lemma: večeře
we-start with-dinner$_{inst}$
'We start with dinner.'

However in many cases, the situation is much more complicated and it is not clear whether a certain deviation is due to a spelling error, incorrect case (Czech has 7 cases + prepositions), wrong paradigm (Czech has 14+ basic noun paradigms) or simply a random error. Sometimes, we can see particular patterns in the whole document, e.g. the author uses only certain cases, or certain spelling convention (Russian speakers sometimes use '*g*' instead of Czech '*h*'), not distinguishing between adjectives and adverbs, etc. These patterns can help us to deduce lemmas in concrete cases. Unfortunately, in some cases we simply do not have enough data to reliably deduce the correct lemma. In that case, we are trying to be as conservative as possible and assume as little as possible: we use the form of the word as its lemma and mark it as unclear in the note field.

The alternative is to use the correct lemma (*Praha* in (4) and *večeře* in (6)). This would obviously make the situation clearer and the annotation more reliable. However, the benefit would be minimal: error annotation already provides us with the correct forms so we can easily derive their lemmas using available approaches for standard native language.

4.4 Syntactic Structure

In annotating syntactic structure, we again follow the rule of annotation the structure of interlanguage. For example, if the learner uses the phrase (7), the word *místnost* 'room' is annotated as a direct object (OBJ), even though a native speaker would use an adverbial (OBL) *do místnosti* 'into room' as in (8).

(7) vstoupit místnost$_{OBJ}$
enter room
intended: ‘enter a/the room’

(8) vstoupit do místnost$_{OBL}$
enter into room
‘enter a/the room’

5 Annotation procedure

For a pilot annotation, we have randomly selected 100 sentences shorter than 15 tokens. The average sentence length is 6.8. Technically, we use the TrEd editor with the *ud* extension to display and edit Universal Dependency trees and labels.[6]

An annotator with a philological background and a secondary-school student annotated the sample. They did not annotate the sentences from scratch, but corrected the output of UDPipe (Straka and Straková, 2017). They did not undergo any special training prior to the annotation, but instead relied on a secondary-school grammar training and the guidelines for Czech available at the UD project site.[7] When they were not sure with a particular construction, they referred to existing Czech and English UD corpora, compiling shared guide and a cheat sheet[8] in the process.

UPOS	LABEL	REL
0.934	0.89	0.927

Table 2: Inter-annotator agreement on the sample of CzeSL measured using Cohen’s *kappa* on UPOS labels, syntactic labels and unlabelled heads respectively

6 Conclusion

We are in the process of creating a syntactically annotated corpus of learner Czech. So far, we have annotated around 2,000 sentences. The goal is to annotate all of the approximately 11 thousand sentences in CzeSL. To the best of our knowledge this is a first such corpus of any inflectional language. We are also planning to have a significant portion of the corpus annotated by two annotators. Currently, we have only around 100 sentences doubly annotated with a good but not perfect inter-annotator agreement. We believe that the most important result of this project is not the actual annotation, but the guidelines that can be used as a basis for other non-native languages. The high-level annotation principles of ours include: (1) When form and function clash, form is considered less important. (2) When lacking information, we make conservative statements. (3) We focus on syntactic structure and the most important grammatical functions, annotating unclear functions with an underspecified label.

Acknowledgments

We thank Filip Hana and Jana Vitoušová Alferyová for annotating the data and commenting the annotation guidelines. We gratefully acknowledge support from the Grant Agency of the Czech Republic, grant No. ID 16-10185S.

[6] `https://ufal.mff.cuni.cz/tred`
[7] `http://universaldependencies.org/guidelines.html`
[8] `http://bit.ly/UDCheat`

References

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., and Katz, B. (2016). Universal dependencies for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746. Association for Computational Linguistics.

de Marneffe, M.-C., Grioni, M., Kanerva, J., and Ginter, F. (2017). Assessing the annotation consistency of the universal dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (DepLing)*, pages 108–115, Pisa, Italy.

Dickinson, M. and Ragheb, M. (2013). Annotation for learner english guidelines, v. 0.1 (june 2013).

Hana, J., Rosen, A., Škodová, S., and Štindlová, B. (2010). Error-tagged Learner Corpus of Czech. In *Proceedings of The Fourth Linguistic Annotation Workshop (LAW IV)*, Uppsala.

Kuzmenko, E. and Kutuzov, A. (2014). Russian error-annotated learner english corpus: a tool for computer-assisted language learning. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, number 107, page 87–97. Linköping University Electronic Press, Linköpings universitet.

Lasota, Brielen Madureira (2018). Slavic Languages and the Universal Dependencies Project: a seminar. `http://www.coli.uni-saarland.de/~andreeva/Courses/SS2018/SlavSpr/presentation_25062018).pdf`. 25 June 2018.

Lee, J., Leung, H., and Li, K. (2017). Towards universal dependencies for learner chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71. Association for Computational Linguistics.

Lee, L.-H., Chang, L.-P., and Tseng, Y.-H. (2016). Developing learner corpus annotation for chinese grammatical errors. *2016 International Conference on Asian Language Processing (IALP)*, pages 254–257.

Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and Smith, N. A. (2018). Parsing tweets into universal dependencies. *CoRR*, abs/1804.08228.

Lyashevskaya, O. and Panteleeva, I. (2018). REALEC learner treebank: annotation principles and evaluation of automatic parsing. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 80–87, Prague, Czech Republic.

Petrov, S., Das, D., and McDonald, R. T. (2011). A universal part-of-speech tagset. *CoRR*, abs/1104.2086.

Ragheb, M. and Dickinson, M. (2013). Inter-annotator agreement for dependency annotation of learner language. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Atlanta, Georgia. Association for Computational Linguistics.

Rosen, A., Hana, J., Štindlová, B., and Feldman, A. (2014). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaliation*, 48(1):65–92.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeman, D. (2015). Slavic languages in universal dependencies. In Gajdošová, K. and Žáková, A., editors, *Natural Language Processing, Corpus Linguistics, E-learning*, pages 151–163, Lüdenscheid, Germany. Slovenská akadémia vied, RAM-Verlag.

Zeman, D., Popel, M., Straka, M., Hajič, J., and Nivre, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Stroudsburg, PA, USA. Charles University, Association for Computational Linguistics.