Parsing Writings of Non-Native Czech

Jirka Hana and **Barbora Hladká**
Charles University
Malostranské nám. 25
118 00 Prague 1
Czech Republic
{hana, hladka}@ufal.mff.cuni.cz

Abstract

We present a pilot study on parsing non-native texts written by learners of Czech. We performed experiments that have shown that at least high-level syntactic functions, like subject, predicate, and object, can be assigned based on a parser trained on standard native language.

1 Introduction

Texts written by non-native speakers pose a challenge for natural language processing. In this paper, we focus on parsing texts written by learners of Czech. There is no syntactically annotated corpus of non-native Czech. Therefore, we are exploring a question whether it is possible to use the parser trained a traditional newspaper corpus.

In our experiments we use three main components: the *Prague Dependency Treebank*, the *CzeSL* corpus, and the *maximum-spanning tree parser*.

The *Prague Dependency Treebank* (PDT) [1] is a corpus of newspaper texts with rich linguistic annotation. As an illustration, consider the sentence in (1) and the corresponding labeled dependency tree in Figure 1:

(1) Ráno půjdu se svým kamarádem na houby.
in-the-morning I-will-go with my friend mushrooming.
'I will go mushrooming with my friend in the morning.'

The *CzeSL* corpus includes essays written by non-native speakers of Czech (Rosen et al., 2013). Finally, the *maximum-spanning tree (MST) parser* is a non-projective dependency parser that searches for maximum spanning trees over directed graphs (McDonald et al., 2005).

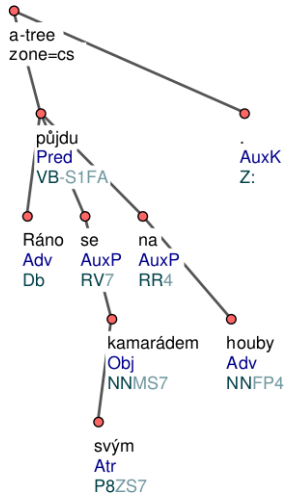

Figure 1: A sample of PDT tree

Given these data and tool components we specify the following initial steps to address our research question:

1. Create a testing corpus, by annotating CzeSL according to the PDT annotation guidelines
2. Parse CzeSL by the MST parser trained on PDT or its subset and evaluate its performance

2 Related work

Research on parsing has mostly concentrated on parsing the traditional treebanks. Therefore most parsers have statistical models that are optimized for the syntactic annotations in these treebanks and more generally for their language. This means that such parsers will show a degradation in performance when used for parsing data from another domain. Thus research has started on adapting parsers to new domains. One of the first venues

[1] https://ufal.mff.cuni.cz/pdt3.0

Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications, pages 12–16,
Taipei, Taiwan, December 1, 2017

at which domain adaptation was targeted was the 2007 CoNLL shared task on dependency parsing, see (Nivre et al., 2007).

One of the challenges in domain adaptation for parsing is the lack of annotated data in the target domain that could be used for evaluation. Focusing on the domain of learner texts and their parsing, the great majority of works concern texts of English learners. We support this fact with a list of learner corpora in Table 1 where their basic characteristics are provided.

Dickinson and Ragheb (2015) consider very carefully the SALLE annotation scheme for syntactically annotating learner English.[2] Napoles et al. (2016) studied the effect of grammatical errors on the dependency parse. As the source of the data, they used the NUCLE corpus. Berzak et al. (2016) benchmarked POS tagging and dependency parsing performance on the TLE dataset and measure the effect of grammatical errors on parsing accuracy. Cahill et al. (2014) used self-training parsing technique with both native and non-native training texts. They found that both training sets performed at about the same level, but that both significantly outperformed the baseline parser trained on traditional labeled data.

3 Syntactic annotation of CzeSL

The *CzeSL* corpus includes transcriptions of essays written by non-native speakers of Czech. It is focused on native speakers of three main language groups: Slavic, other Indo-European, and non-Indo-European. The hand-written texts cover all language levels, from real beginners to advanced learners.

In this paper, the CzeSL corpus refers to the *CzeSL-man* corpus that consists of 645 texts written by 262 different authors who are native speakers of 32 different languages. As shown in Table 2, the texts belong mostly to A2-B2 CEFR.[3]

Its annotation scheme consists of three interconnected layers:

- the T0 layer contains anonymised transcripts of the originals,
- the T1 layer corrects non-existing word forms ignoring context,
- the T2 layer corrects all other types of errors, including syntactic errors.

Level	Documents	Level	Documents
A1	57	B1	176
A1+	3	B2	124
A2	111	C1	12
A2+	145	Unknown	17

Table 2: Composition of CzeSL according to CEFR levels.

Jmenujese Adam. Ja jsem Mongolska. Mongolska ma 21 kraji. Moje rodina je hezka ještě velka. Mongolska je 3000 million lidi. Ma tradični píseňka, taneční. Mongolska tradični písenka je hezka. Ještě ma "Morin khuur". Morin Khuur to je muzika. Ten hezka tradični pohádka, píseň. Mongolska má mnoho tradiční svátík. Třiba Naadam, Tsagaarsur. Ješte mnoho Velbloud, Kůn, Kravá, Koza, Ovce. Mongolsky lidi dobrý. Mongolsko ma mnoho horý a nemam ocean. Mongolska hlavní naměsto. Ulaanbaatar.
ADAM, 18 Let
Bydlim v Čechagh už 6 měsíc.
1. AHOJ

Figure 2: An essay written by a 16+ male student of the non Indo-European language group staying in the Czech Republic less than a year. The essay is on My family.

In our experiment, we focus on learner language, therefore we use only the T0 layer, disregarding any corrections made on the T1 and T2 layers.

Learner texts typically differ from newspaper corpora, i.e., highly edited texts written by native speakers, in two aspects: first, they contain errors in spelling, grammar, vocabulary, and collocations; second, they have a different distribution of vocabulary and syntactic constructions.

For illustration, consider a sample essay in Table 2. The text is perfectly understandable, yet it contains errors practically in every sentence and about every other word. Some of these deviations from native language make annotation with traditional grammatical categories quite complicated. For example, consider the second sentence: *Ja jsem Mongolska* meaning 'I am Mongolian' or 'I

[2] http://cl.indiana.edu/~salle/

[3] The Common European Framework of Reference for Languages; see https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages.

	TLE	NUCLE	SALLE	CzeSL
	Berzak et al. (2016)	Dahlmeier et al. (2013)	Dickinson and Ragheb (2013)	Rosen et al. (2013)
L2	English	English	English	Czech
entity	corpus	corpus	framework	corpus
volume	5,124 sentences	1,414 texts	NA	645 texts
error annotation	•	•	NA	•
POS tags	•	○	NA	○
Universal Dependencies	•	○	NA	○

Table 1: A number of learner corpora (• annotation present, ○ annotation not-present, Not Applicable).

am from Mongolia'. The non-existent word *Mongolska* can be interpreted in at least the following three ways:

1. it is an adjective (*mongolská* or *mongolský*) and thus syntactically a predicative nominal;
2. it is a name of an inhabitant (*Mongol*), a noun, syntactically a predicative nominal;
3. a place (*z Mongolska*), a noun, syntactically an adverbial (adjunct).

It is not clear, whether the language of the speaker actually distinguishes all of these categories.

Learner language is challenging not just for NLP tools, but for human annotation as well. We decided to start partial syntactic analysis. Instead of building a complete dependency tree and labelling each node, we opted to perform a linear annotation of subjects, objects, predicates and predicative nominals.

Two high-level annotation instructions were formulated:

1. Use the PDT guidelines[4] to mark subjects, objects, predicates and nominals with the corresponding PDT syntactic functions `Sb`, `Obj`, `Pred`, `Pnom`, resp.
2. Annotate the language of the learner, not the target hypothesis (a standard Czech expression with the same meaning). For example, if the learner uses the phrase (2), the word *místnost* 'room' is annotated as an object, even though a native speaker would use an adverbial *do místnosti* 'into room'.

(2) vstoupit místnost .
enter room .
'intended: enter a room.'

(3) vstoupit do místnosti .
enter into room .
'enter a room.'

In this, we are consistent with other annotation projects, for example the SALLE project: *Try to assume as little as possible about the intended meaning of the learner.* (Dickinson and Ragheb, 2013)

Unlike a traditional treebanking project, which is a very expensive activity, the CzeSL corpus was annotated in three months by one annotator with a philological education. Instead of intensive training, the annotator annotated the data and studied the guidelines in parallel. When she was in doubt, she consulted the problem with an experienced linguist who annotated the Prague Dependency Treebank. The annotator used the Brat editor.[5]

4 Experiments

We experimented with two different parsers: (i) a traditional parser trained on PDT (ii) a parser trained on a simpler subset of Czech. In both cases, we used the MST parser.

4.1 STYX – Training on simpler language

On average, sentences in newspapers have a more complicated structure than sentences found in a typical non-native text. This motivated us to experiment with a parser that would be trained on a corpus using simpler syntax than that of PDT.

Hladká and Kučera (2008) present the STYX, an electronic corpus-based exercise book of Czech grammar.[6] The STYX corpus is based on PDT, but contains only "simple" sentences.

[4] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf

[5] http://brat.nlplab.org/

[6] http://hdl.handle.net/11234/1-2391

syntactic function	P	R	F1	# tokens in the test data
1. parser: MST_{PDT}			test data: CzeSL	
`Pred`	0.87	0.80	0.83	13,762
`Sb`	0.60	0.37	0.46	8,119
`Obj`	0.44	0.61	0.51	12,615
`Pnom`	0.38	0.50	0.43	2,870
avg / total	0.63	0.62	0.61	37,366
2. parser: MST_{Styx}			test data: CzeSL	
`Pred`	0.70	0.83	0.76	13,762
`Sb`	0.41	0.31	0.35	8,119
`Obj`	0.51	0.50	0.51	12,615
`Pnom`	0.40	0.23	0.29	2,870
avg / total	0.55	0.56	0.55	37,366
3. parser: MST_{PDT}			test data: Styx_{etest}	
`Pred`	0.99	0.98	0.98	1,220
`Sb`	0.58	0.34	0.42	1,059
`Obj`	0.34	0.60	0.43	1,066
`Pnom`	0.37	0.52	0.43	198
avg / total	0.64	0.65	0.62	3,543
4. parser: MST_{Styx}			test data: Styx_{etest}	
`Pred`	0.95	0.96	0.95	1,220
`Sb`	0.62	0.37	0.47	1,059
`Obj`	0.58	0.49	0.53	1,066
`Pnom`	0.49	0.34	0.40	198
avg / total	0.71	0.61	0.65	3,543

Table 3: We measure the performance of the MST parser using the following performance measures: Precision, Recall, and F1 measure.

4.2 Results

We have evaluated the two parsers against the manual annotation of CzeSL. For comparison, we have also evaluated them on Styx, i.e. a corpus of native Czech. The results are summarized in Table 3. The subscript indicates which corpus was used for training (MST_{PDT} vs MST_{Styx}). The results are surprising in two ways:

1. the performance on the learner language is nearly comparable to the performance on native Czech.

2. training the parser on a simpler language not only does not help, but actually hurts the performance

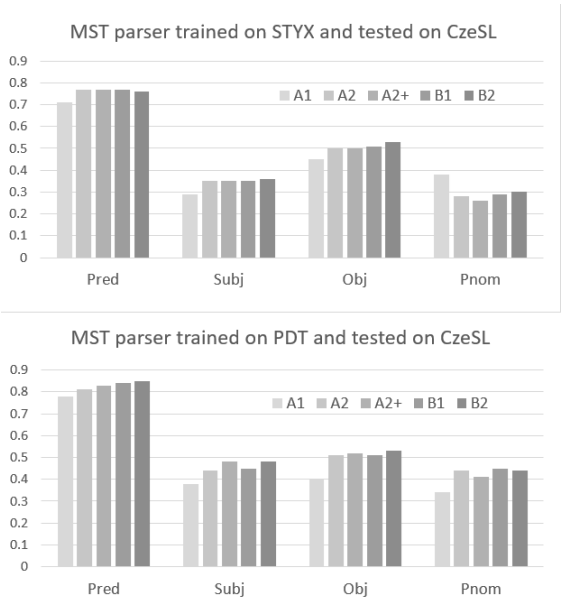

Figure 3: Performance of the parsers on CzeSL: F1-measure CEFR levels. (A1+ and C1 are omitted due to low number of documents)

Figure 3 shows performance of the parsers on the CzeSL corpus by syntactic category and CEFR level. For the PDT parser, the performance is worst for A1 and better on more advanced levels, as expected. However, starting with A2+ level, the performance does not improve with CEFR levels. One of the explanations might be that on the one hand those advanced texts contain less "low-level" errors (errors in spelling and morphology), but on the other hand the sentences get longer and get a more complicated syntax structure.

5 Conclusion and future work

Our experiments have shown that at least high-level syntactic functions, non-native text can be assigned based on a parser trained on standard native language. It has also shown, that training the parser on a subset of standard language limited to simpler construction provided no benefit. Currently, we focus on two main tasks:

- Repeating the annotation of a part of the CzeSL corpus with a second annotator, to be able to calculate inter-annotator agreement

- Evaluating the possibility of annotating additional syntactic functions and possibly a limited structure

Acknowledgments

We gratefully acknowledge support from the Grant Agency of the Czech Republic, grant No. ID 16-10185S.

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for Learner English. *CoRR*, abs/1605.04278.

Aoife Cahill, Binod Gyawali, and James Bruno. 2014. Self-training for parsing learner text. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 66–73, Dublin, Ireland. Dublin City University.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

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

Markus Dickinson and Marwa Ragheb. 2015. On Grammaticality in the Syntactic Annotation of Learner Language. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 158–167, Denver, CO.

Barbora Hladká and Ondřej Kučera. 2008. An annotated corpus outside its original context: A corpus-based exercise book. In *ACL 2008: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 36–43, Columbus, OH, USA. Association for Computational Linguistics (ACL).

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

Courtney Napoles, Aoife Cahill, and Nitin Madnani. 2016. The effect of multiple grammatical errors on processing non-native writing. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, San Diego, CA. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic. Association for Computational Linguistics.

Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2013. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, pages 1–28.