2018 International Conference on Bilingual Learning and Teaching

25- 27 October 2018, OUHK Jockey Club Campus

Title: Syntactic annotation of a second-language learner corpus

Name: Jirka Hana / Barbora Hladka

Affiliation: Charles University

Email: hana@ufal.mff.cuni.cz / hladka@ufal.mff.cuni.cz

Bio data

Jirka Hana and Barbora Hladka are senior researchers at the Institute of Formal and Applied Linguistics, at Charles University in Prague. Jirka Hana's research focuses on learner corpora, non-standard language and computational morphology. Barbora Hladka specializes in Corpora, Non-native English and Czech and Machine learning.

Abstract

CzeSL (Hana et al 2010, http://utkl.ff.cuni.cz/learncorp/) 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.

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.

For these reasons, in our project, we have focused on syntactic annotation of the non-native text within the framework of Universal Dependencies (http://universaldependencies.org). 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.

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.

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.

Key words:

learner corpus, second language, syntax annotation, universal dependencies, second language acquisition

Highlights: (not more than 3 bullet points)

- Syntactic annotation of a morphologically rich non-native language

---

Introduction

Universal Dependencies (UD) is a unified approach to grammatical annotation that is consistent across languages. It has been used to annotate treebanks in more than 60 languages, thus facilitating 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.

CzeSL corpus

CzeSL (Rosen et al 2013, Hana et al 2010, http://utkl.ff.cuni.cz/learncorp/) 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.

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 (https://bitbucket.org/czesl/czesl-man/) is a subset of CzeSL, manually annotated for errors. 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: an anonymized transcript of the hand-written original with some properties of the manuscript preserved (variants, illegible strings),
- Tier 1: 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 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 (1) – showing the original text (T0) and the target hypothesis (T2). The full error analysis, including error tags is in Figure 1.

(1)	T0:	Myslim	že	kdy	by	byl	se	svim	ditem	...
	T2:	Myslím ,	že	kdybych		byl	se	svým	dítětem	...
		think.$_{SG1}$	that	if.$_{SG1}$		was.$_{MASC}$	with	my	child	...

`I think that if I were with my child ….'

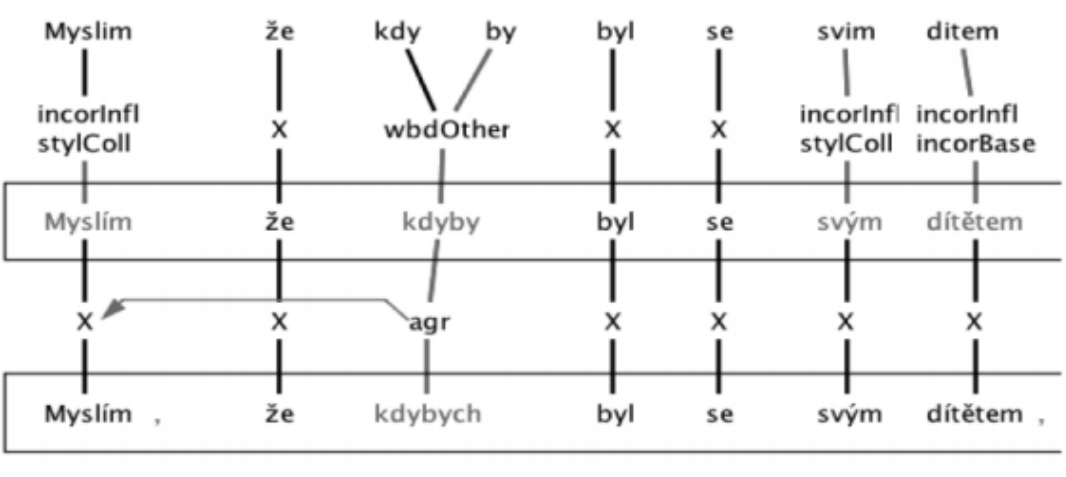

Figure 1 – Full error annotation corresponding to (1) `I think that if I were with my child ….'

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, 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.

Grammatical Annotation of Learner Czech

For these reasons, in our project, we have focused on syntactic annotation of the non-native text within the framework of Universal Dependencies (Nivre et al., 2016, http://universaldependencies.org). As far as we know, this is the first project annotating a richly inflectional non-native language.

Universal Dependencies

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.

Related Work

Corpus	TLE	REALEC Kuzmenko et a 2014	TWEEBANK Liu et al 2018	CFL Lee et al 2017
Language	English	English	English	Mandarin
Size (annotated)	5,124 sentences 97,681 words	373 sentences 7,196 words	3,550 tweets 45,607 words	451 sentences 7,256 words

Table 1 – Relevant UD-annotated Corpora

Table 1 summarizes UD-annotated corpora relevant for our task. 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 no 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.

Approach

Similarly as the projects above, we follow the basic annotation principle of the SALE project (Ragheb and Dickinson, 2014), 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.

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 (https://ufal.mff.cuni.cz/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., 2012). 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.

Lemmata

Ideally, we would use lemmata from the author's interlanguage. For example, in (2), 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 (3) the lemma of *krasivaja* is *krasivyj*, based on Russian.

(2)	T0:	Praga	je	hezké	město
	T1:	Praha	je	hezké	město
		Prague	is	nice	City
		'Prague is a nice city'			

(3)	T0:	Praga	je	krasivaja
	T1:	Praha	je	krásná
		Prague	is	beautiful
		'Prague is beautiful'		

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.

The alternative is to use the correct lemma (Praha in example (2)). This would obviously make the situation clearer and the annotation more reliable. However, the benefit would be minimal – we already know the correct forms so we can easily derive their lemmas using available approaches for standard native language.

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 (4), 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’.

(4) vstoupit místnost
enter room
intended: enter a/the room

(5) vstoupit do místnost
enter into room
enter a/the room

Conclusion

We are in the process of creating a syntactically annotated corpus of learner Czech. So far, we have annotated around 2000 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 by 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:

- When form and function clash, form is considered less important.
- 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.

Acknowledgments

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

References

Berzak, Y.; Kenney, J.; Spadine, C.; Wang, J. X.; Lam, L.; Mori, K. S.; Garza, S. & 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)', Association for Computational Linguistics, Berlin, Germany, pp. 737–746.

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

Kuzmenko, E. & 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', Linköping University Electronic Press, Linköpings universitet, pp. 87–97.

Lee J., Leung H., & Li K. (2017), Towards Universal Dependencies for Learner Chinese. In Proc. Workshop on Universal Dependencies.

Liu Y., Zhu Y, Che W., Qin B., Schneider N., & Smith N. A.. (2018), Parsing Tweets into Universal Dependencies. In Proc. of NAACL.

Nivre, J.; de Marneffe, M.-C.; Ginter, F.; Goldberg, Y.; Hajic, J.; Manning, C. D.; McDonald, R. T.; Petrov, S.; Pyysalo, S.; Silveira, N.; Tsarfaty, R. & Zeman, D. (2016), Universal Dependencies v1: A Multilingual Treebank Collection, *in* 'LREC', European Language Resources Association (ELRA).

Petrov, S.; Das, D. & McDonald, R. T. (2011), 'A Universal Part-of-Speech Tagset', *CoRR* abs/1104.2086.

Ragheb, M. & Dickinson, M. (2014), Developing a Corpus of Syntactically-Annotated Learner Language for English, *in* 'Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)', pp. 292--300.

Rosen, A.; Hana, J.; Štindlová, B. & Feldman, A. (2013), 'Evaluating and automating the annotation of a learner corpus', *Language Resources and Evaluation*, 1-28.