NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Olga Lyashevskaya, Irina Panteleeva

AUTOMATIC DEPENDENCY PARSING OF A LEARNER ENGLISH CORPUS REALEC

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS
WP BRP 62/LNG/2017

This Working Paper is an output of a research project implemented at the National Research University Higher School of Economics (HSE). Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.

SERIES: *LINGUISTICS*

Olga Lyashevskaya[1], *Irina Panteleeva*[2]

AUTOMATIC DEPENDENCY PARSING OF A LEARNER ENGLISH CORPUS REALEC[3]

The paper presents a Universal Dependencies (UD) annotation scheme for a learner English corpus. The REALEC dataset consists of essays written in English by Russian-speaking university students in the course of general English. The essays are a part of students' preparation for the independent final examination similar to the international English exam.

While adjusting existing dependency parsing tools to a learner data, one has to take into account to what extent students' mistakes provoke errors in the parser output. The ungrammatical and stylistically inappropriate utterances may challenge parsers' algorithms trained on grammatically appropriate written texts.

In our experiments, we compared the output of the dependency parser UDpipe (trained on UD-English 2.0) with the results of manual parsing, placing a particular focus on parses of ungrammatical English clauses. We show how mistakes made by students influence the work of the parser. Overall, UDpipe performed reasonably well (UAS 92.9, LAS 91.7). The following cases cause the errors in automatic annotation a) incorrect detection of a head, b) incorrect detection of the relation type, as well as c) both. We propose some solutions which could improve the automatic output and thus make the assessment of syntactic complexity more reliable.

JEL Classification: Z.

Keywords: learner corpus, dependency annotation of learner treebank, Universal Dependencies, evaluation of parser quality, L2 English.

1. Introduction

The diversity of research based on learner corpora is increasing in the fields of language acquisition and language teaching methodology. The manual and automatic analysis of texts written

[1] National Research University Higher School of Economics. School of Linguistics. Professor; E-mail: olesar@yandex.ru.

[2] National Research University Higher School of Economics. School of Linguistics. Bachelor Student; E-mail: impanteleyeva@gmail.com.

[3] The publication was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2016–17 (grant No 16-05-0057) and by the Russian Academic Excellence Project "5-100". We would like to thank the participants of the Research Team Project “Learner corpus REALEC: Lexicological observations” and in particular, Olga Vinogradova for discussions and suggestions. An extended version of this paper will be published in the Proceedings of TLT 2018.

by learners leads to the creation of various tools used for pedagogical purposes, namely, for improvements in teaching techniques achieved by paying attention to frequent errors that have been made by generations of learners. Linguistic data obtained in the analysis of the learner corpora texts serve as a basis not only for teaching but also for evaluating the works written by people learning a language.

Using different automatic tools in learner corpus is a frequent idea of works aimed at checking the progress of learning language. For example, Cobb and Horst point out the importance of such analysis of learners' essays (Cobb, Horst 2015). In (Berzak et al. 2016) Berzak introduces a publicly available syntactic treebank for English as a Second Language (ESL) that provides manually annotated POS tags and Universal Dependency (UD), thanks to which the data obtained from the parser can be checked. Moreover, ESL annotation allows for consistent syntactic treatment of ungrammatical English texts. Many applications based on syntactic parsing are made in cooperation with Daniella McNamara, cf. McNamara 2011, among others, in which the results on linguistic evaluation of complexity are presented. One more complexity analyzer is made by (Lu, Ai 2016). This work provides a set of simple criteria such as the length of clause, the number of dependent clauses and so on. In (Ragheb 2012) authors discuss improving syntactic annotation for learner language by dint of clarifying the properties which the layers of annotation refer to. They also show the mistakes of annotation that could be corrected with the help of some tools. The list of the studies in learner data syntactic parsing also includes (Rosen, DeSmedt 2010) who explore how dependency annotation complements the annotation of errors, and (Schneider, Gilquin 2016) who focus on innovations in learner's grammar revealed by parsing, to name just a few. In (Rooy, Schäfer 2002) Bertus van Rooy and Lande Schäfer present the idea that spelling errors cause errors in parsing. Also they show how the errors of learners influence the performance of the taggers. This will be confirmed in our research.

In (Vinogradova et al. 2017) syntax complexity is already discussed based on the examples from corpus REALEC. The paper presents the results of the syntactic analysis made by parsing the sentences regarding mean sentence depth and the average number of relative, other adnominal, and adverbial clauses. There we cleared up how much these criteria influence on syntactic complexity of the essay. The analysis showed that the mean sentence depth is insignificant for evoluation of text and the average number of clauses, on the contrary, is considered to be feature of better works (scored 75 % and higher).

The article is structured as follows. In Section 2 we present the original data from the corpus REALEC on which we based for this research. Section 3 is focused on the dependency annotation principles and tackles a number of difficult examples from the corpus. Section 4 ('Choice among alternatives') explains how we choose the variant of annotation. Section 5 and 6 present the data sample of our research and also reports which tool we used for automatic parsing. In the Section 6 ('Confusion matrix and causes of errors') we analyze the relations which are labeled incorrectly in the parser output. In Section 7 we focus on the constructions that require particular attention since they cause frequently errors in automatic annotation. Section 8 concludes.

2. Original data: corpus REALEC

The treebank annotation reported in this article are based on the materials from the publicly available corpus REALEC (Russian Error-Annotated English Learner Corpus), see Vinogradova 2016, Vinogradova et al. 2017, available at: http://realec.org). It is an open access collection of English texts written by Russian-speaking students of English. The resource consists of more than 3,500 essays written by bachelors students while preparing for international examination. Students' errors are annotated manually by experts (EFL instructors and trained students). Error labels are divided into groups depending on the type of error (spelling, punctuation, grammar, vocabulary, and discourse that include further detailed division). Experts mark the error span, assign to it one error tag or a few tags, and suggest the corrected version of the span. The original corpus is also equipped with tools for searching and downloading the text.

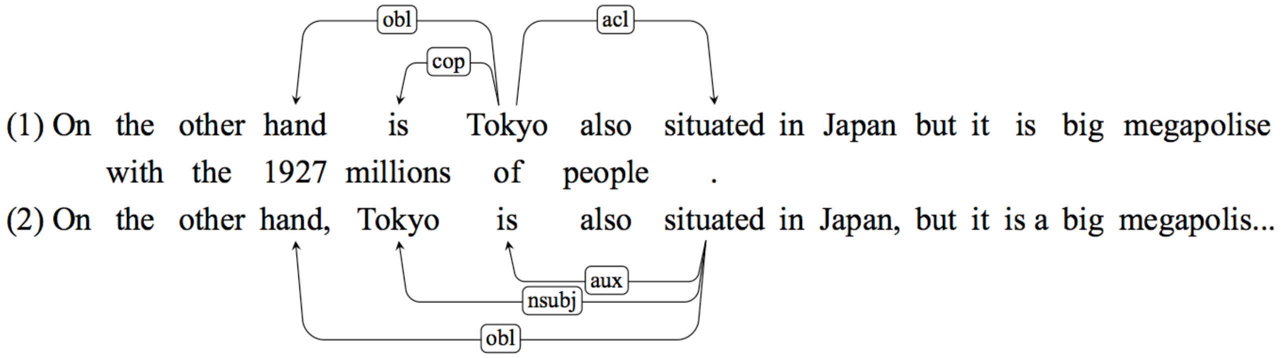
3. Dependency annotation scheme

We choose Universal Dependencies framework (Nivre et al. 2016) since it allows one to present typologically diverse treebanks in a comparable format and allows for matching different types of dependency relations in different languages. There are 32 dependency relation types provided by parsers trained on English UD 2.0 data, among them subject and object, relative, adverbial and adnominal clauses, conjunction, auxiliary and copula, parataxis).

There exist two common approaches to syntactic annotation of learner and other not-well-edited data: 'literal' labeling describe the way the two words are related given their formal properties (Lee et al. 2017), whereas and alternative design bears on the notion of 'intended' usage, and experts are asked consider functional rather than formal side of the utterance and to reconstruct what was intended to be said. The REALEC annotation scheme follows the latter in both building the tree structure and labeling the dependency relations.

(1) and (2) below illustrate an original sentence and it's 'intended' reading (a partly corrected version). In (1), the phrases *On the other hand is Tokyo* and *Tokyo situated in Japan* present two locally well-formed syntactic structures, but their combination within the whole tree is problematic for the 'literal' approach.

As for the 'intended-usage' approach, it is prone to the word order related issues that reflect native patterns of Russian speakers. What is convenient, the corpus is already annotated for students' errors, so our experts can get use of 'the suggested corrections' provided in that layer. However, we do not ask the treebank annotators to rewrite sentences in the correct way, as the intended reading is only implied.


obl
cop
acl
(1) On the other hand is Tokyo also situated in Japan but it is big megapolise
with the 1927 millions of people .
(2) On the other hand, Tokyo is also situated in Japan, but it is a big megapolis...
aux
nsubj
obl

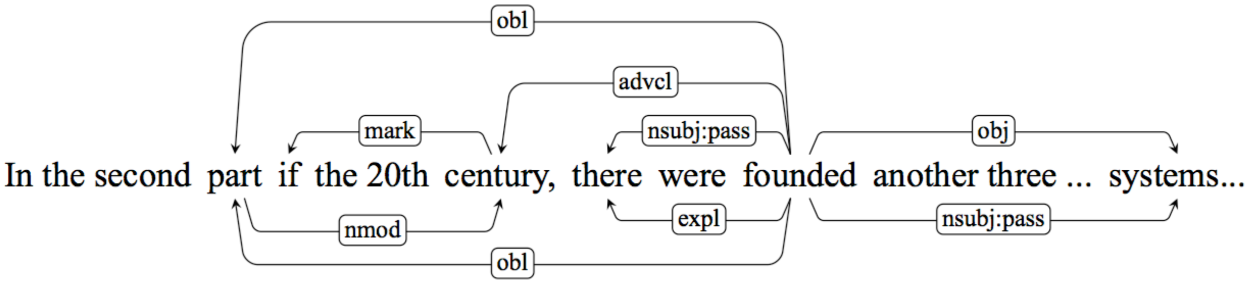
In schemes that follow we show the automatic output (edges above the text) and gold parses (edges below the the text), respectively.

4. Choice among alternatives

There can be multiple alternatives for possible corrections, in which case the principle of minimal edit distance seems to be releveant. For example, in sentence (4), two readings can evoke.

(3) *In the second part if the 20th century, there were founded another three major railway systems, which although had significantly worse harasteristics.*

The first one is the situation that is chosen by the automatic parser but grammatically it is not quite grammatically correct. We have chosen the variant where we change *if* for *in*. In this case we also have to change the label of the primary word *if* for 'case'.


obl
advcl
mark
nsubj:pass
obj
In the second part if the 20th century, there were founded another three ... systems...
nmod
expl
nsubj:pass
obl

5. Parsing and manual corrections

We needed an easy-to-use parser that would provide the information about part-of-speech, syntactical groups, dependency relation between words and to represent the syntax trees for more convenient counting, so the choice fell on UDPipe (Straka 2016, available at: http://ufal.mff.cuni.cz/udpipe) trained on English UD 2.0 treebank. Like any parser, UDPipe makes mistakes, and it was important to evaluate the output for the purposes of pur project and assess to what extent these mistakes are imposed by students' errors in orphography, morphology, and syntax.

For the research, 373 random sentences (7196 tokens, including 756 punctuation marks) from students' essays were checked to evaluate the UPPipe parser quality. The parser detected the heads correctly for 6688 nodes (UAS 92,9 %), of which 6 600 were labeled correctly (LAS 91,7 %). Overall, 6894 nodes (95,8 %) were labeled correctly, which suggest that the disfluencies affected the tree structures rather than functions.

6. Confusion matrix and causes of errors

Table 1 illustrates the confusion matrix for the most frequent mismatches in relation types. The totals are calculated for all relations.

Tab. 1. Confusion matrix of relation types

gold auto	acl	nsubj	nummod	amod	case	obj	obl	root	nmod	compound	conj	others
acl	36					1		4			1	
nsubj		475		1		5	1	9	1	2	7	5
nummod			227				3		2	1	1	
amod			3	387	1		1	4		6		1
case					994							7
obj	2	2				246	1	1	2	4	7	1
obl		1	1			1	405	1	10		1	
root	1	1	1					348	5	3	8	9
nmod	3	1	4	1		1	15	1	465	6	6	6
compound		1		5			1		5	141	3	
conj	2			2	4	2	2	3	1	5	270	7
others	2	4	2	3	7	9		2		4	15	

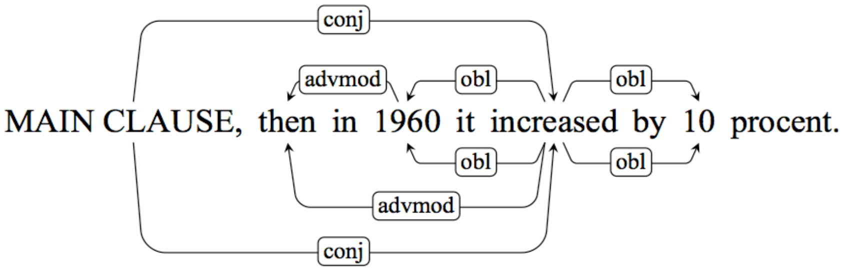
The most frequent relation errors are mismatches between root and adjectival modifier, root and nominal subject, object and nominal modifier, root and nominal modifier, conjunction and root, adnominal modifier and conjunction. Their are different causes of incorrect detection of relation type, for example, incorrect detection of the head of the sentence (confusion between root and other relations), incorrect detection of the syntactic group, incorrect detection part of speech, errors of the student.

7. Constructions that require attention

We have identified the cases in which the parser most often makes mistakes. In what follows we consider errors influenced by syntactic homonymy, ungrammatical word order, spelling and grammar mistakes, including those typical for the Russian students of English. In addition, we analyze some distance effects in the wrong attribution of the heads of participle and introductory phrases.

7. 1. Syntactic homonymy

(4) *Meanwhile, in USA there was 9 procent of people aged 65 and over in 1940, then in 1960 it increased by 10 procent.*

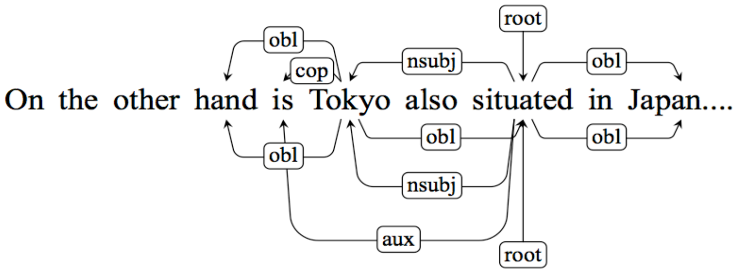


Here we can see that the linking word *then* refers not to the whole sentence. It is parsed as the clarification of the circumstance of time *in 1960*. This is not a critical mistake but the automatic parsing slightly changes the meaning of the statement.

7. 2. Errors influenced by word order

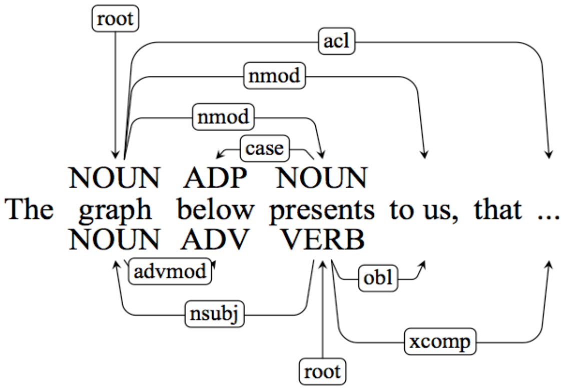
Sentence (5) demonstrates the wrong SV word order typical in students' writing. In a gold representation, this mistake is reflected in a non-projective tree.

(5) *On the other hand is Tokyo also situated in Japan but it is big megapolise with the 1927 millions of people.*



However, it can be seen that even in well-formed sentences the parsing errors can be explained by non-standard word order patterns. Sentence (6) is ambiguous between a noun and a verb reading, the former being provided by the parser. Here, the adverbial modifier *below* comes after it's nominal head (*graph*), thereby evoking the reading of the segment *below present* as PP.

(6) *The graph below presents to us, that between 1983 and 2030 in Japan it rise from 3 procent to 10 procent, but in Sweden it is a little fall to 13 13 procent, but there was a high growth to 20 procent in 2010.*

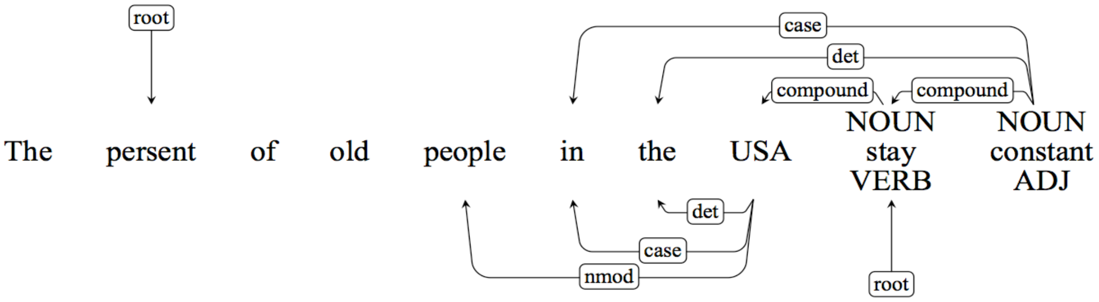


7. 3. Spelling and grammar mistakes made by students

We checked to what extent spelling of words affects the parser's work. Comparison of automatic and gold parses in (7) and it's 'improved' version (8) demonstrates that verb agreement is critical for parsing.

(7) *The persent of old people in the USA stay constant (14 %) from 1980 to 2020 and rising quicly (23 %) during next 20 years.*

(8) *The percent of old people in the USA stays constant (14 %) from 1980 to 2020 and rises quicly (23 %) during next 20 years.*



The schemes show that grammatically correct sentences are parsed better than those with spelling and grammatical mistakes. We suggest that this problem could for the most part be solved with the help of a common spellchecker. It will allow us to analyze the syntactic structure of the sentences ignoring the students' grammar and spelling errors that do not influence syntactic complexity.

Generally, the modification in grammar of the sentences showed that the grammatically correct statements are parsed more accurately than those that contain errors. The main mistake of the parser is the wrong detection of part of speech. It influences the wrong detection of sentence root which is considered as a critical for parsing and entails other errors (in head detection and consequently in type of relation). Accordingly, spelling correction before parsing would reduce the errors made by the parser.

7. 4. Participial construction

(9) *Tokyo railway, opened in 1927, was only 155 kilometres on route but, compare to previous system, helped to travel to almost 2000 millions passengers.*

In (9), the participle *opened* is parsed as the root of the sentence. As the parser chooses the part of speech incorrectly, the error arises: *opened* is defined as a verb and it becomes more and more probable that this word will be the root of the sentence. The probability that the conjunction is the head of a sentence is less than the probability that the head is a verb.

7. 5. Introductory prases

(10) *Accordingly, the same situation as in the proportion of skilled vocational diploma is in postgraduate diploma.*

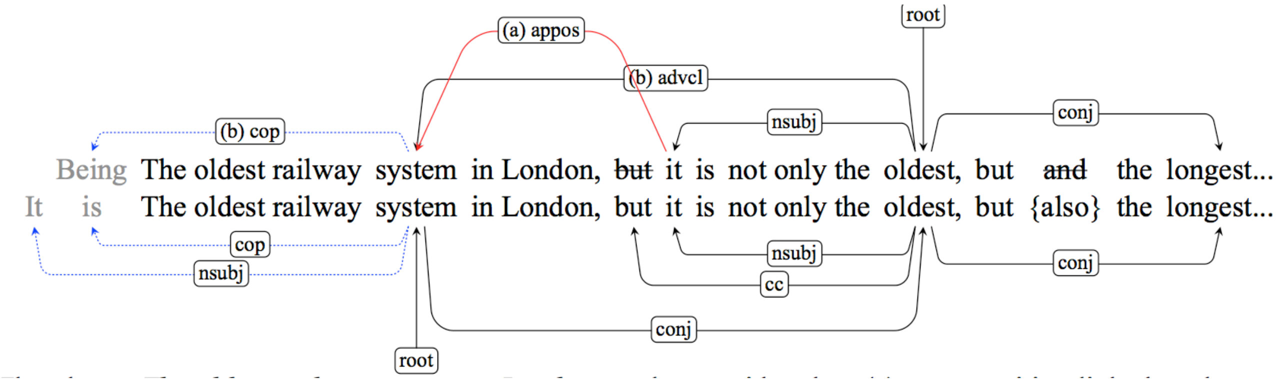
The parser determines the noun *situation* as the head of the word *accordingly* whereas the right host here is root of the whole sentence, *diploma*. As the head of the introductory phrase is too far, parser take the closest possible word as a head. The head of the introductory word should be always the root of the whole sentence.

7. 6. Typical errors made by Russian students

In a learner corpus essays with L1-interference mistakes often occurs. In our sample we also have such cases. The errors can be connected with calques, or the possibility of omitting the auxilary verb in Russian when in English it is not possible, or the absence of category L1, for example, articles, uses of perfect forms of the verb, several types of relative clauses, etc.

For example, sentence (10) has calque mistake critical to building an appropriate syntactic structure: there is a conjunction (*but*) between the noun phrase and the clause, and there is a double coordinating conjunction *but and* between two adjectives, *oldest* and *longest*.

(11) *The oldest railway system in London, but it is not only the oldest, but and the longest – three hundred ninety four kilometres of route.*


(a) appos
root
(b) advcl
nsubj
conj
(b) cop
Being The oldest railway system in London, ~~but~~ it is not only the oldest, but ~~and~~ the longest...
It is The oldest railway system in London, but it is not only the oldest, but {also} the longest...
cop
nsubj
nsubj
conj
cc
conj
root

The phrase *The oldest railway system in London* can be considered as (a) an appositive linked to the pronoun *it* in the main clause; (b) a part of the concessive clause (with *being* being omitted), or (c) a part of the main clause where the copula and *it* are omitted).

8. Conclusion

Overall, this paper has presented the REALEC learner treebank, which was automatically annotated by UDPipe and then manually corrected. The paper included features involved in evoluation of automatic parsing. We explored what types of errors that students make are critical for the parser.

We confirmed the idea of van Rooy and Schäfer, who claim that if we check the spelling in essays before applying a parser, errors that are not related to the syntax will not affect the evaluation of the syntactic complexity. This conclusion leads to the idea that advanced annotator learner corpora should have a spellchecker which analyzes not only the spelling, but also improves the work of various automatic tools.

Studying the results during the work of the UDPipe parser, we found out that the problems arise in phrases which occur very frequently, especially in texts written in academic register of English. Examples given above show common phrases like *a chart below* or *7 years old*, in which the parser fails in the detection of the head, which usually leads to a large number of manual corrections.

The obtained results will help to improve the work of the parser and the process of the annotation in the learner corpora. Firstly, we suggest a list of typical error-provoking patterns based on the collection of reannotated sentences. In the future the number of such cases will be expanded. Secondly, as the amount of annotated learner data in open access grows, we will conduct a series of experiments on parser training and compare results obtained on grammatically correct texts and learner data.

As a future work, we intend to take more samples from the learner corpus REALEC to increase the volume of our treebank. We would also like to use dependency parsing to improve the quality of corpus annotation.

REALEC treebank is freely available under the CC BY-SA 3.0 licence.

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, pages 737–746.

Cobb, T. and Horst, M. (2015). Learner corpora and lexis. In The Cambridge Handbook of Learner Corpus Research, pages 185–206. Cambridge University Press.

Graesser, A., McNamara, D., and Kulikowich, J. (2011). Coh-metrix: Providing multilevel analyses of text char- acteristics. Educational Researcher, 40(5), pp. 223–234.

Lee, J., Leung, H., and Li, K. (2017). Towards universal dependencies for learner chinese. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, 22 May 2017, Vol. 135.

Lu, X. and Haiyan, A. (2016). Universal dependencies for learner english. Journal of Second Language Writing, volume 29, pp. 16–27.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of Language Resources and Evaluation Conference (LREC'16).

Ragheb, M. and Dickinson, M. (2017). Defining syntax for learner language annotation. In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Poster Session, pp. 965–974.

Rosén, V. and Smedt, K. D. (2010). Syntactic Annotation of Learner Corpora, pages 120–132.
Schneider, G. and Gilquin, G. (2016). Detecting innovations in a parsed corpus of learner english. International Journal of Learner Corpus Research, 2 (2), pp. 177–204.

Straka, M., Hajič, J., and Strakova, J. (2016). Ud-Pipe: trainable pipeline for processing CONLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 4290–4297.

Van Rooy, B. and Schäfer, L. (2002). Universal dependencies for learner English. Southern African Linguistics and Applied Language Studies, 20 (4), pp. 325–335.

Vinogradova, O. (2016). The role and applications of expert error annotation in a corpus of english learner texts. In Computational Linguisitics and Intellectual Technologies. Proceedings of Dialog 2016, vol. 15, pp. 740–751.

Vinogradova, O., Lyashevskaya, O., and Panteleeva, I. (2017). Multi-level student essay feedback in a learner corpus. In Computational Linguisitics and Intellectual Technologies. Proceedings of Dialog 2017, vol. 16, pp. 382–396.

Contact details:

Olga Lyashevskaya
National Research University Higher School of Economics (Moscow, Russia). School of Linguistics. Professor;
E-mail: olesar@yandex.ru. Web: https://www.hse.ru/staff/olesar. Tel. +7 (906) 798-60-21
Irina Panteleeva
National Research University Higher School of Economics (Moscow, Russia). School of Linguistics. Bachelor Student;
E-mail: impanteleyeva@gmail.com.

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.