# REALEC learner treebank: annotation principles and evaluation of automatic parsing

**Olga Lyashevskaya**
School of Linguistics
National Research University
Higher School of Economics,
Vinogradov Institute of the Russian
Language RAS Moscow
olesar@yandex.ru

**Irina Panteleeva**
School of Linguistics
National Research University
Higher School of Economics
Moscow
impanteleyeva@gmail.com

## Abstract

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 original corpus is manually annotated for learners' errors and gives information on the error span, error type, and the possible correction of the mistake provided by experts. The syntactic dependency annotation adds more value to learner corpora since it makes it possible to explore the interaction of syntax and different types of errors. Also, it helps to assess the syntactic complexity of learners' texts.

While adjusting existing dependency parsing tools, 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 academic texts. In our experiments, we compared the output of the dependency parser Ud-pipe (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, Ud-pipe performed reasonably well (UAS 92.9, LAS 91.7). We provide the analysis of several cases of erroneous parsing which are due to the incorrect detection of a head, on the one hand, and with the wrong choice of the relation type, on the other hand. We propose some solutions which could improve the automatic output and thus make the syntax-based learner corpus research and assessment of the syntactic complexity more reliable.

The REALEC treebank is freely available under the CC BY-SA 3.0 licence.[1]

## 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 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 language learning. For example, Cobb and Horst point out the importance of such analysis of learners' essays (Cobb and Horst, 2015). Berzak et al. (2016) introduce a publicly available syntactic treebank for English as a Second Language (ESL), which provides manually annotated POS tags and Universal Dependency (UD), with 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 have been created in cooperation with Daniella McNamara, cf. (Graesser

[1] https://github.com/olesar/REALECtreebank



*Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 80–87,
Prague, Czech Republic, January 23–24, 2018. 

**Keywords:** learner corpus, dependency annotation of learner treebank, UD, evaluation of parser quality

et al. (2011), in which the results on linguistic evaluation of complexity are presented. One more complexity analyzer is made by (Lu and Haiyan, 2016). This work provides a set of simple criteria such as the length of each clause, the number of dependent clauses, and so on. In ((Ragheb and Dickinson, 2017) authors discuss how to improve 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 (Rosén and Smedt, 2010), who explore how dependency annotation complements the annotation of errors, and (Schneider and Gilquin, 2016), who focus on innovations in learner's grammar revealed by parsing, to name just a few. In (Rooy and Schäfer, 2002) Bertus van Rooy and Lande Schäfer present the idea that spelling errors cause errors in parsing. Also they show how learners' errors influence the performance of the taggers. Our research, as we hope to show, also confirms this.

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

In the section 'Original data' we present data on which we based for this research. The next part of the text named 'Dependency annotation scheme' shows how we worked on the examples from the corpus. Section 'Choice among alternatives' explains how we chose the option of the annotation. The next chapter presents the sample of our research and also reports which tool we have used. In the section 'Confusion matrix and causes of errors' we show the relations which are confused frequently in students' essays. In 'Constructions that require attention' the examples from corpus that cause the errors in the parser's work are brought in.

## 2 Original data

The treebank annotations 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).[2] It is an open-access collection of English texts written by Russian-speaking students of English. The resource consists of more than 3,500 pieces written by Bachelor students while preparing for the English 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, with the last three further subdivided according to a detailed categorisation scheme). 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.

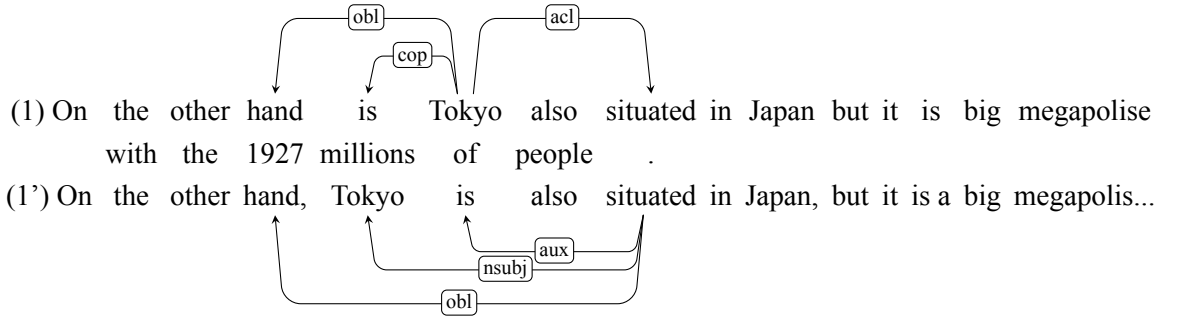
## 3 Dependency annotation scheme

We have chosen Universal Dependencies framework ((Nivre et al., 2016) since it allows one to present typologically diverse treebanks in a comparable format and provides certain matching of 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 insufficiently edited data: 'literal' labeling describe the way the two words are related given their formal properties (Lee et al., 2017)), whereas an alternative design bears on the notion of 'intended' usage, and experts are asked to consider functional rather than formal side of the utterance and to try and reconstruct what the intended meaning of the author was. (1) and (1') below illustrate an original sentence and its 'intended'

[2] http://realec.org

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.



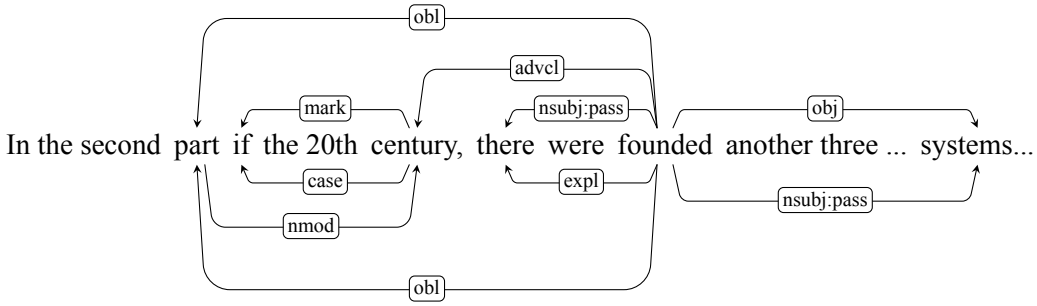

In schemes that follow we show the automatic output (edges above the text) and gold parses (edges below the the text), respectively.

## 4 Choice among alternatives

There can be multiple alternatives for possible corrections, in which case the principle of minimal editing distance seems to be releveant. For example, in sentence (2), two readings can evoke.

(2) *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 grammarwise it is not quite correct. We have chosen the option where we change *if* for *of*. In this case we also have to change the label of the primary relation 'mark' for 'case'.




## 5 Parsing and manual corrections

We needed an easy-to-use parser which would provide the information about part-of-speech, syntactical groups, dependency relation between words and which would represent the syntax trees for more convenient counting, so the choice fell on Ud-pipe (Straka et al., 2016; Straka and Straková, 2017)[3] trained on english ud 2.0 treebank. Like any parser, Ud-pipe makes mistakes, and it was important to evaluate the output for the purposes of our 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 processed with the Ud-pipe parser. The parser detected the heads correctly for 6688 out of 7196 nodes (UAS 92,9 %), of which 6600 were labeled corectly (LAS 91,7 %). Overall, 6894 nodes (95,8 %) were labeled correctly, which suggests that it was the disfluencies that affected the tree structures, rather than functions.

[3] http://ufal.mff.cuni.cz/Parsing

## 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.

| | **acl** | **nsubj** | **num mod** | **amod** | **case** | **obj** | **obl** | **root** | **nmod** | **com pound** | **conj** | **others** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **acl** | 36 | | | | | 1 | | 4 | | | 1 | |
| **nsubj** | | 475 | | 1 | | 5 | 1 | 9 | 1 | 2 | 7 | 5 |
| **num-mod** | | | 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 |
| **com-pound** | | 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 | |

**Table 1:** Confusion matrix of relation types.

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. There are different causes of incorrect detection of relation type, some of them depend on failures in other parsing stage - for example, incorrect detection of the head of the sentence (confusion between root and other relations), incorrect detection of the syntactic group, incorrect detection of part of speech, while still others are the result of learner errors.

## 7 Constructions that require attention

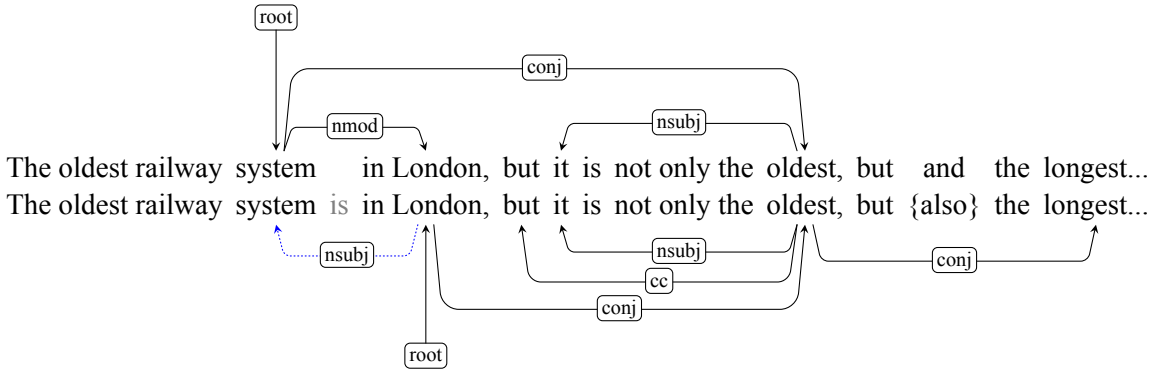
We have identified the cases in which the parser most often makes mistakes. The following examples present the errors that arise because of ungrammatical nature of sentences, or because of the parser's deficiency.

### 7.1 Typical errors made by Russian students

In a learner corpus essay, L1-interference mistakes often occur. 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 in L1, for example, articles, uses of perfect forms of the verb, several types of relative clauses are all absent in Russian, to name just a few.

For example, sentence (3) has a 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*.

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




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 is omitted after the subject *The oldest railway system*).

The next example presents the frequent mistake made by Russian students - the usage of large amount of specifying words. Because of them the parser determines the head of the sentence incorrectly.

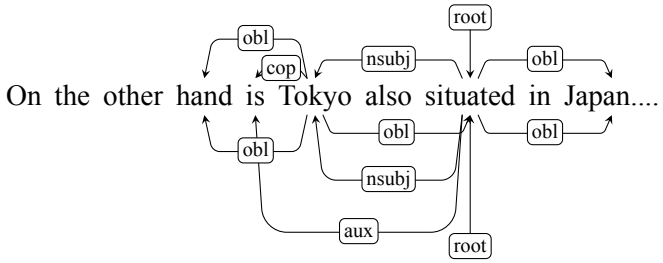
(4) *Accordingly , the same situation as in the proportion of skilled vocational diploma is in postgraduate diploma.*

The parser determines the noun *situation* as the head of the word *accordingly*, but the right choice here is the 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.

### 7.2 Errors influenced by word order

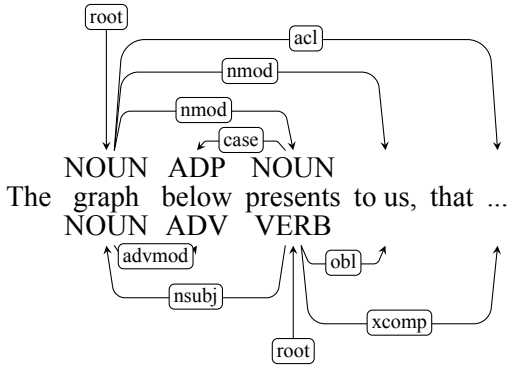
Sentence (5) demonstrates the wrong SV word order typical of students' writing. In a gold representation, this mistake is reflected in a non-projective tree.

(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) has an ambiguity in reading *presents* as a noun or as a verb, the former being provided by the parser. As a result, the adverbial modifier *below* comes after its nominal head (*graph*), thereby evoking the reading of the segment *below present* as PP.

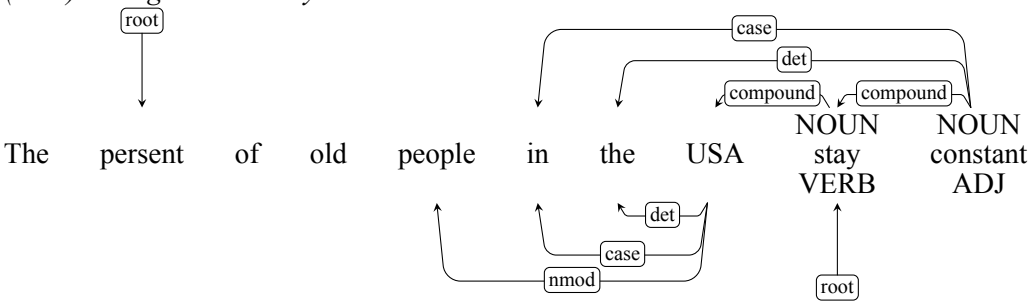
(6) *The graph below presents to us, that between 1983 and 2030 in Japan it rise from 3 procent to 10 procent, but in Sweden it is a little fall to 13 procent , but there was a high growth to 20 procent in 2010.*



### 7.3 Spelling and grammar mistakes made by students

We investigated to what extent misspelt words affect the parser's quality. Comparison of automatic and gold parses in (7) with those of its 'improved' version (7') 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.*
(7') *The percentage of old people in the USA stays constant (14 %) from 1980 to 2020 and rises quickly (23%) during the 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 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 causes the wrong detection of sentence root, which is considered critical for parsing and entails other errors (in head detection and consequently in type of relation). Accordingly, spelling correction made before parsing would reduce the number of errors made by the parser.

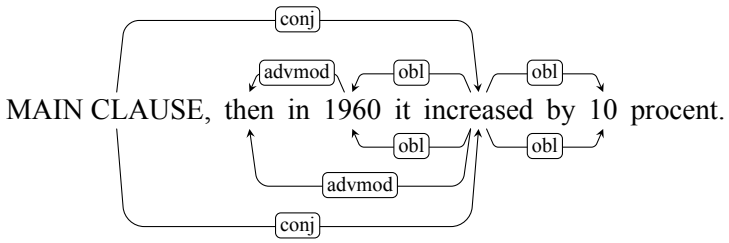
### 7.4 Participial construction not recognized by the parser

(8) *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 (8), 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 of choosing *opened* as a verb and the head of the sentence is higher than the probability of choosing *kilometres* as the head of the sentence.

### 7.5 Syntactic homonymy

(9) *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 adverbial modifier of time *in 1960*. This is not a critical mistake but the automatic parsing slightly changes the meaning of the statement.

## 8 Conclusion

This paper presents the REALEC learner treebank automatically annotated by Ud-pipe and then manually corrected. We provide evaluation of the automatic parsing output and explore what types of learners' errors are critical for the parser.

We confirmed the idea of van Rooy and Schäfer, who claimed 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 annotated learner corpora should have a spellchecker which analyses not only the spelling, but also improves the work of various automatic tools.

Studying the output of the Ud-pipe parser, we found out that phrases like *a chart below* or *7 years old*, which occur frequently in academic register of English, are parsed incorrectly. In such cases, the parser fails to identify the head of the phrase, which is in turn the cause of further parser errors, and involves a large amount of manual corrections.

The obtained results will help to improve the quality of the parser and the annotation in the learner corpora. Firstly, we have identified a list of typical error-provoking patterns based on the collection of reannotated sentences. In the future the inventory of such patterns will be expanded. Secondly, as the amount of annotated learner data in the open access grows, we will conduct a series of experiments on parser training and compare the models trained on grammatically correct texts vs. those involving learner data.

For future work, we also plan to increase the size of our treebank taking more samples from the learner corpus REALEC. We would also like to use dependency parsing to improve the quality of corpus annotation.

## Acknowledgements

The paper 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.

## 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 characteristics. *Educational Researcher*, 40(5):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*, volume 135.

Lu, X. and Haiyan, A. (2016). Universal dependencies for learner English. In *Journal of Second Language Writing*, volume 29, pages 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*, pages 965–974.

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

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):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)*, pages 4290–4297.

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*.

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*, volume 15, pages 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*, volume 16, pages 382–396.