Universal Dependencies for Learner Russian

Alla Rozovskaya
Department of Computer Science
Queens College, CUNY
CUNY Graduate Center
arozovskaya@qc.cuny.edu

Abstract

We introduce a pilot annotation of Russian learner data with syntactic dependency relations. The annotation is performed on a subset of sentences from RULEC-GEC and RU-Lang8, two error-corrected Russian learner datasets. We provide manually labeled Universal Dependency (UD) trees for 500 sentence pairs, annotating both the original (source) and the corrected (target) version of each sentence. Further, we outline guidelines for annotating learner Russian data containing non-standard erroneous text and analyze the effect that the individual errors have on the resulting dependency trees. This study should contribute to a wide range of computational and theoretical research directions in second language learning and grammatical error correction.

Keywords: Russian learner grammar, dependency parsing, syntactic annotation of learner Russian

1. Introduction

There has recently been a lot of work in the Natural Language Processing (NLP) community with a focus on non-standard texts, written by language learners. Most of the work in this area focuses on Grammatical Error Correction (GEC), the task of detecting and correcting mistakes in text (Chollampatt and Ng, 2018; Yuan and Briscoe, 2016; Grundkiewicz et al., 2019; Grundkiewicz and Junczys-Dowmunt, 2019; Kiyono et al., 2019; Zhao et al., 2019; Jianshu et al., 2017; Yuan and Briscoe, 2016; Katsumata and Komachi, 2019; Xie et al., 2018; Palma Gomez et al., 2023; Rozovskaya and Roth, 2016). Most of the research in GEC has focused on mistakes made by English as a Second Language writers, but datasets in other languages have also been created recently, including Arabic (Mohit et al., 2014), Chinese (Zhang et al., 2022), Russian (Rozovskaya and Roth, 2019; Trinh and Rozovskaya, 2021), Ukrainian (Syvokon and Romanyshyn, 2023; Syvokon and Nahorna, 2021), and Czech (Naplava et al., 2022). Although there are quite a few studies on building computational models for correcting learner errors, there has been very little work on the linguistic analysis of learner errors, with the exception of studies that focus on the automatic classification of learner errors based on their syntactic and morphological characteristics, e.g. (Bryant et al., 2017; Choshen et al., 2020; Rozovskaya, 2022).

In this work, we study the problem of generating dependency parse trees for Russian learner data. We conduct a pilot study and annotate a 500-sentence subset from two Russian manually-corrected learner corpora. One challenge of learner data annotation is the non-canonical language use resulting from the errors in the data. Although existing dependency frameworks have detailed guidelines on the annotation of syntactic relations, these do not address the issue of annotating non-standard constructions and relations that are not present in well-formed texts. To address this shortcoming, we study the challenges of dependency annotation in Russian learner texts. We adopt the Universal Dependency framework (UD, Nivre et al. (2016)), that is particularly conducive to annotating non-canonical language use arising in learner data, due to the influence of the learner first language syntactic properties on the underlying constructions used in the second (foreign) language. Although prior work exists on learner language syntactic annotation, this is the first work that addresses the issue in Russian.

This paper makes the following contributions: (1) We present a pilot Learner Treebank of Russian that consists of 500 sentence pairs; both the source and the corrected sentence are annotated, resulting in a parallel dependency learner corpus; (2) We describe challenges arising from annotating non-standard syntactic constructions and propose how to handle constructions that are not found in standard Russian language; (3) We also identify error-specific challenges affecting the annotation of the syntactic structure in learner data.[1]

We envision this resource to support work in GEC, second language acquisition and corpus analysis of non-standard data. The analysis of non-standard constructions and their treatment following the UD framework should help develop similar resources for other languages. The study should also contribute to a further an understanding of syntactic errors and the influence of the native language of the learner on the use of syntactic con-

[1]The annotations are available for research at https://github.com/arozovskaya/dependency-learner-russian.

LREC-COLING 2024, pages 17112–17119

structions in the second language.

2. Background and Related Work

Berzak et al. (2016) were the first to create a Treebank of Learner English, a first of its kind resource for non-native English, containing 5,124 sentences manually annotated with dependency trees. The annotation followed the Universal Dependency formalism (Nivre et al., 2016), which provides a unified annotation framework across different languages and is geared towards multilingual NLP (McDonald et al., 2013). Thus, using the UD framework allows for a unified approach that can relate the linguistic structures in the second language to those used the native language of the learner (Berzak et al., 2016) and to account for the effect of second language interference (Leacock et al., 2010). Studies in second language learning find that learners may generate a sentence in the second language by translating it from their native language (Watcharapunyawong and Usaha, 2013; Derakhshan and Karimi, 2015), thereby incorrectly transferring structures and expressions into the second language.

Berzak et al. (2016) discuss the challenges of correcting non-standard structures in English, building on earlier work in learner language analysis (Ragheb and Dickinson, 2012; Dickinson and Ragheb, 2013; Dıaz-Negrillo et al., 2010) to formulate an additional set of annotation conventions aiming at a uniform treatment of ungrammatical learner language. The annotation scheme in Berzak et al. (2016) uses a two-layer analysis, providing a distinct syntactic annotation for the original and the corrected version of each sentence. We adopt the same approach, but introduce additional decisions, to support the specific non-canonical constructions arising from the complex morphology of Russian.

3. Russian Learner Data

Overview of the Russian grammar Russian belongs to the Slavic subgroup of the Indo-European language family. It has complex highly fusional morphology that includes case, gender, and number marking for adjectives, nouns, pronouns and numerals, as well as complex verb conjugation systems. Russian does not have definite and indefinite articles, which makes it challenging for learners whose native language has an article system. This is also a language with free word order.

Russian learner data We use two datasets authored by speakers of a variety first language backgrounds. Both datasets are manually corrected for errors: the RULEC-GEC corpus (henceforth RULEC) (Alsufieva et al., 2012) manually corrected for errors (Rozovskaya and Roth, 2019; Palma Gomez and Rozovskaya, 2024), and RU-Lang8, a dataset of Russian learner writing collected from the online language learning platform Lang-8 (Mizumoto et al., 2011) and annotated by native speakers (Trinh and Rozovskaya, 2021).

RULEC contains essays written by learners of Russian, mainly native English speakers. RU-Lang8 data is drawn from the Lang-8 corpus and contains data from a variety of first language backgrounds (Mizumoto et al., 2012).

While the errors in RULEC are manually labeled with error type at the level of syntax, morphology, and lexical usage (a total of 22 categories), the annotation of RU-Lang8 is performed at the level of four operations: Replace, Insert, Delete, and Word Order. To assign error categories in RU-Lang8, we apply an error classification tool developed for Russian (Rozovskaya, 2022), that uses a part-of-speech (POS) tagger and a morphological analyzer (Sorokin, 2017) to automatically classify the edits into appropriate linguistic types (the tool follows the error taxonomy adopted in the original RULEC annotation. Common Russian learner errors are illustrated in Table 1. Appendix Table A.2 illustrates common Russian learner errors within sentential context.

4. Dependency Annotation

The pilot annotation includes a sample of 500 sentences from the test partitions of RULEC (300 sentences) and RU-Lang8 (200 sentences). The dependency annotations are provided for both versions of each sentence: the *original* sentence authored by the learner, and the *corrected* version. Since some of the original sentences do not contain corrections, and our goal is to focus on the annotation of non-canonical structures, we exclude sentences that do not contain errors. We then select a sample of 1,000 sentences from each dataset that does not include short sentences (we exclude sentences that are shorter than 12 tokens in RULEC and shorter than 8 tokens in RU-Lang8). We then sorted each list of 1,000 sentences by the number of edits in decreasing order, and selected the top 300 sentences in RULEC and the top 200 sentences in RU-Lang8. The smallest number of corrections in the resulting sample is two edits per sentence, and the average number of edits is 3.4. The average sentence length in the annotated sample is 19 tokens. Table 3 lists the relative frequencies for the top-12 most common error types in the 500-sentence sample.

Dependency annotation for non-standard structures The general annotation scheme for dependency structures follows the UD guidelines and label inventory. However, these guidelines do not cover non-canonical syntactic structures that arise due to learner errors.

Error type	Example
Punctuation	∅ →,
Extraneous word (open-class)	был “was” → ∅
Missing word (open-class)	∅ → для того “with the purpose of”
Prep. (ins.,del.,repl.)	в “in” → из “from, out of”
Noun case/number	иде-и (“idea” (sg.,gen/pl.,nom.)) → иде-й (“idea” (pl.,gen))
Noun case	специалист-ы “experts” (pl.,nom) → специалист-ам (pl.,dat.)
Noun number	пол-а“gender” (sg.,gen.) → пол-ов “gender” (pl.,gen.)
Adj. case	главн-ая “main” (sg., fem., nom.) → главн-ую (sg., fem., acc.)
Adj. number	дальнейш-ие “future” (pl.,nom.) → дальнейш-ее “future” (sg.,nom.))
Verb agr. (number/gender/person)	жив-ут “live” (3rd person pl.) → жив-ет (3rd person sg.)
Morphology (deriv.)	вдохнов-ленным “inspired” → вдохнов-енной “inspiring”
Lex. choice (word)	предлагает “proposes” → утверждает “claims”

Table 1: Some common error types in Russian learner data. Partial changes on a word are shown with a hyphen. The following error labels are used to denote errors in the use of morpho-syntactic categories (inflectional morphology): *noun case/number, verb agreement (number/gender/person), adj. case/number/gender*. The category *morphology (deriv.)* denotes errors in derivational morphology. Sample sentences from the learner data containing these errors are shown in Appendix Table A.2.

The “literal reading” principle We follow Berzak et al. (2016) and adopt the “literal reading” principle, whereby syntactic structures are annotated based on the observed language usage. The “literal reading” principle means that non-standard structures are annotated based on the relations exhibited in the original sentence produced by a learner, and not based on the relations in the corrected sentence. We illustrate the application of the “literal reading” principle with the following example: consider an expression involving a missing preposition, as in “we waited him”, with the preposition “for” missing: “him” would be annotated as the direct object of the verb “wait” in the original sentence (the *obj* label in UD), whereas it would be labeled as *obl* in the corrected version “waited for him”. Example 1(c) in Table 2 illustrates this in Russian.[2] This annotation strategy is motivated by work in second language acquisition advocating for centering analysis of learner language around morpho-syntactic surface evidence (Ragheb and Dickinson, 2012). Due to Russian being a morphologically-rich language with free word order, there is a greater variety than in English, of non-canonical structures that deserve special treatment, and we discuss this below.

Annotation decisions and disagreement resolution We hired an annotator, who is a native Russian speaker with a Master’s degree, basic background in Linguistics and previous annotation experience.[3] The annotator carefully studied the dependency grammar formalism and the UD guidelines. To establish guidelines for annotating various ungrammatical constructions, for each error type, the annotator performed an initial round of annotation, consisting of a sample of about 100 sentences pairs containing about 10 occurrences of each of the most frequent error types. The annotator identified recurring issues in the treatment of specific constructions tied to individual error types and proposed their annotation decisions. These annotations were reviewed by the author of the paper, and disagreements were resolved through discussions. The remaining 400 sentences were annotated following the proposed guidelines. The annotation of the source and of corrected sentence versions was conducted in parallel: the annotator first annotated the corrected version of the sentence, and then proceeded to annotate the original non-canonical sentence. We found that this approach helped maintain consistency in the annotation of identical structures and allowed the annotator to focus on the differences between the two sentences.

Annotation of non-canonical structures Below, we discuss the decisions on the dependency annotation for the most common error types as shown in Table 1. Among the errors shown in the table, mistakes in verb number/gender agreement, verb aspect, spelling, punctuation, adjective case/number agreement do not typically alter the syntactic dependency tree. For the other common errors, we identify the most prominent issues, discuss how we treat those non-canonical structures, and show how these errors affect the dependency annotation.

Preposition mistakes include preposition replacement errors, extraneous and missing prepositions. As most of the prepositions have a verb or noun attachment, we consider the effect of a preposition error occurring in those contexts. Preposition replacement errors do not affect the dependency relations. Furthermore, an incorrect preposition headed by a noun also typically does not affected the dependency structure. However, when a preposition introduces a dependency with the head being

[2] An exception to this principle are orthographic (spelling) errors that are annotated according to the intended meaning of the word.

[3] The annotator previously contributed to the annotation of the RULEC and RU-Lang8 datasets.

(1a) Preposition error (extraneous, predicate dependency)
Согласно-1 (rel=parat.) *с-2/case автора-3 (rel=obl; head=1) ,-4 они-5 ели-6 (root) ... капусту-7
Согласно-1 (rel=case) ∅ автору-2/(rel=parat.,head=5) ,-3 они-4 ели-5 (root) ... капусту-6
'According *with/the author, they ate... cabbage'
(1b) Preposition errors (extraneous, noun dependency)
места-1 жительства-2 (rel=nmod,head-1) *для-4 (rel=case, 'for') мигрантов-4 (rel=nmod;head=2)
'places of residence for migrants'
место-1 жительства-2 ∅ мигрантов-3 (rel=nmod, head=2)
'migrants place of residence'
(1c) Preposition error (missing, verb dependency)
язык-1 всегда-2 влияет-3 (root) *∅ наше-4 мышление-5 (rel=obj; head=3)
язык-1 всегда-2 влияет-3 на-4 (rel=case, head=6,'on') наше-5 мышление-6 (rel=obl; head=3)
'language always influences *∅/on our thinking'
(2a) Noun case (nominal dependency)
в-1 результате-2 влияния-3 (rel=nmod) языковой-4 *политикой-5 (rel=nmod, head=3, case=instr.
в-1 результате-2 влияния-3 языковой-4 политики-5 (rel=nmod, head=3, case=gen.)
'as a result of the influence of the linguistic (language) policy'
(2b) Noun case (predicate dependency)
обеспечивать-11 (root) пресную-2 *воду-3 (rel=obj, head=1, case=acc.)
обеспечивать-1 пресной-2 водой-3 (rel=obl, head=1, case=instr.)
'to provide with drinking water'
(3a) Missing word - dependent
...я-1 должна-2/root была-3 (rel=cop,head=2) *∅ богатой-4/rel=xcomp; head=2
...я-1 должна-2 (root) была-4 (rel=cop,head=2) быть-5 (rel=cop,head=6,'to be'), богатой-6 (rel=xcomp,head=2)
'I was supposed *∅/(to be) rich'
(3b) Missing word - dependent
В-1 каком-2 фильме-3 *∅ хотели-4 (root) бы-5 сняться-6 ?-7
В-1 каком-2 фильме-3 вы-4 (rel=nsubj,head=root, 'you') хотели-5 (root) бы-6 сняться-7 ?-8
'In what movie would *∅/you like to star?'
(3c) Missing word - root
У-1 нас-2 (rel=obl, head=ROOT) *∅ (ROOT) человеческая-3 ответственность-4 (rel=nsubj,head=ROOT)
У-1 нас-2 (rel=obl, head=3) есть-3 (root, 'there is') человеческая-4 ответственность-5 (rel=nsubj, head=3)
'To-us *∅/there is human responsibility'
(4a) Extraneous word
Я-1 не-2 знаю-3 пока-4 ,-5 *если-6 (rel=mark, head=8) мне-7 придётся-8 убегать-9
Я-1 не-2 знаю-3 пока-4 ,-5 придётся-6 ли-7 (rel=advmod,head=9) мне-8 убегать-9
'I don't know yet if I will have to run away'
(4b) Extraneous word
Для-1 них-2 ,-3 *это-4 (rel=expl, head=6) очень-5 важно-6 (root) знать-7 этого-8 автора-8
Для-1 них-2 ,-3 очень-4 важно-5 (root) знать-6 (rel=csubj, head=6) этого-7 автора-8
'For them, it is very important to know this author...'
(5a) Morphology
Люди-1 ...не-2 знают-3 (root) *стихию-4 (rel=obj, head=3) бедствий-5 (rel=nmod, head=4)
Люди-1 ...не-2 знают-3 (root) стихийных-4 (rel=amod, head=4) бедствий-5 (rel=obj, head=3)
'People...do not know *nature/natural disasters...'

Table 2: Examples of non-canonical structures and their dependency relations. Incorrect words are marked with *. ∅ denotes a missing word. Each example shows the source sentence, followed by the corrected sentence, and followed by the English translation. The relevant dependency relations in the non-canonical structures are listed. Indices correspond to word position in the sentence.

Error type	Rel. freq. (%)
Spelling	20.6
Lex. choice (word)	11.4
Noun case	6.9
Punctuation	9.2
Missing word	4.3
Extra. word	1.8
Noun case/num.	7.4
Preposition	5.0
Lex. choice (phrase)	12.4
Adj. case	3.0
Verb agreement	1.9
Morphology (deriv.)	0.9
Total errors	1,707

Table 3: List of top-12 error types and their relative frequencies in the 500-sentence sample annotated with dependency relations.

the verb, the dependency relation changes. See examples below 1(a)-1(c) in Table 2.

Noun case errors are the most frequent type of inflectional errors among Russian learners. When a noun with a case error modifies another noun, there are typically no changes in the dependency structure (example 2(a) in Table 2). However, when a noun is a core argument or an adjunct of a verb, a noun case error usually causes a change in the type of a dependency relation (examples 2(b)). Note that direct objects (*obj*) in Russian typically are associated with accusative case (rarely genitive or nominative), nominals in dative case are labeled as *iobj* (indirect object), and nouns in instrumental case that are not introduced by a preposition are labeled as *obl*.

Missing words are some of the most common errors. A few prominent examples of commonly omitted words include the verb есть/'to be/to have' that does not have a direct correspondence in English and many other languages, personal pronouns in subject positions, and an expletive это. issing word errors typically change the structure of the sentence in a major way and thus cause changes in dependency relations. We treat these instances similar to the treatment of ellipses in UD. If a predicate is missing (see example 3(c) in Table 2), we create an artificial ROOT node.[4]

Extraneous word errors can be broken down into two categories: extraneous words that are typically modifiers, whose omission those does not alter the syntactic structure of a sentence. The other group includes errors that involve the incorrect use of connectors and markers, such если, что, то, бы, как, and expletive это. These errors affect the overall syntactic structure of the sentence. See examples 4(a) and 4(b) in Table 2. Note that in example (4c) the extraneous 'it' (это) is marked as expletive, although this construction does not exist in standard Russian.

[4]https:universaldependencies.org/v2/ellipsis.html

Morphology errors are mistakes where the base form of the word used is correct but the derivational morphology is incorrect. The use of an incorrect derivational suffix or prefix could result in a different part-of-speech, thereby affecting the overall syntactic structure (see example 5(a) in Table 2), where noun стихия 'nature' is used instead of the correct adjective стихийное 'natural'.

Distributions of the dependency relations We have computed the distribution of the syntactic relations in the original and the corrected sentences. Overall, we did not observe major differences in distribution between the syntactic relations of the original and corrected sentences in the resulting corpus. However, three types of relations are underused in the original sentences (*nmod, nsubj,punct.*, and *advmod,cc,mark* are slightly overused, compared to the corrected sentences. Relative frequencies for the most common relations are shown in Appendix Table A.1. We believe that the reason for the lack of major differences in distribution is due to the "minimal edit principle" that the annotators followed in the correction of the original sentences (Palma Gomez and Rozovskaya, 2024).

5. Conclusion and Future Work

In this work, we present a pilot annotation of 500 sentences of Russian learner texts annotated with dependency relations. We adopt the "literal reading" principle, whereby non-canonical structures resulting from errors in the data are annotated according to the observed language usage. This approach facilitates the relation between second language usage and the native language of the learner. We have identified and discussed the unique challenges of annotating dependency relations in Russian and connected those to specific error types based on a linguistically-motivated classification schema. We believe that our annotation reflects the most common ungrammatical structures in learner Russian and the annotation decisions for such structures.

Future work will include extending the annotation to include more data from learner corpora and to address less common errors. We also plan to hire a second annotator, which will allow us to compute the inter-annotator-agreement. We also plan to use the annotation to develop a parser for noisy learner Russian data. Finally, while we have focused on annotating learner data, there is a related line of work on creating learner translation corpora, for example, (Kutuzov and Kunilovskaya, 2014) that develop Learner Russian Translator corpus; annotating this data for dependency structures would be another interesting direction for future work.

In addition to the computational applications of this dataset, the described resource should be of interest to researchers working on the computational and cognitive aspects of language acquisition.

6. Ethics Statement

The annotation presented in this work is performed using data from existing datasets that are available for research (Mizumoto et al., 2012; Rozovskaya and Roth, 2019). The annotation presented in this work was manually generated by a native Russian speaker hired to perform that annotation for a compensation. The amount of the compensation was established based on a compensation that was offered for similar annotation efforts, and that amount was deemed acceptable by the annotators. The authors are not aware of any potential problems that could result from the use of the data and the annotations.

Acknowledgments

The author is grateful to the anonymous reviewers for their insightful comments.

7. Bibliographical References

A. Alsufieva, O. Kisselev, and S. Freels. 2012. Results 2012: Using flagship data to develop a russian learner corpus of academic writing. *Russian Language Journal*, 62:79–105.

Y. Berzak, J. Kenney, C. Spadine, J.X. Wang, L. Lam, , K.S. Mori, S. Garza, and B. Katz. 2016. Universal Dependencies for learner English. In *ACL*.

C. Bryant, M. Felice, and T. Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *ACL*.

S. Chollampatt and H.T. Ng. 2018. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction . In *Proceedings of the AAAI*. Association for the Advancement of Artificial Intelligence.

L. Choshen, D. Nikolaev, Y. Berzak, and O. Abend. 2020. Classifying syntactic errors in learner language. In *CoNLL*.

A. Derakhshan and E. Karimi. 2015. The interference of first language and second language acquisition. *Theory and Practice in Language Studies*, 5(10):2112–2117.

M. Dickinson and M. Ragheb. 2013. Annotation for learner English guidelines. In *Technical report*.

A. Dıaz-Negrillo, D. Meurers, S. Valera, and H. Wunsch. 2010. Towards interlanguage pos annotation for effective learner corpora in sla and flt. *Language Forum*, 36(1-2):139–154.

R. Grundkiewicz and M. Junczys-Dowmunt. 2019. Minimally-augmented grammatical error correction. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT*.

R. Grundkiewicz, M. Junczys-Dowmunt, and K. Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA-19)*.

J. Jianshu, Q. Wang, K. Toutanova, Y. Gong, S. Truong, and Jianfeng J. Gao. 2017. A nested attention neural hybrid model for grammatical error correction. In *ACL*.

S. Katsumata and M. Komachi. 2019. (almost) unsupervised grammatical error correction using synthetic comparable corpus. In *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA-19)*.

S. Kiyono, J. Suzuki, M. Mita, T. Mizumoto, and K. Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *EMNLP-IJCNLP*.

A. Kutuzov and M. Kunilovskaya. 2014. Russian learner translator corpus. In *Text, Speech and Dialogue*, pages 315–323. Springer International Publishing.

C. Leacock, M. Chodorow, M. Gamon, and J. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, and H. Zhang et al. 2013. Universal dependency annotation for multilingual parsing. In *ACL*.

T. Mizumoto, Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of esl writings. In *COLING*.

T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto. 2011. Mining revision log of language learning SNS for automated japanese error correction of second language learners. In *IJCNLP*.

B. Mohit, A. Rozovskaya, N. Habash, W. Zaghouani, and O. Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *ANLP Workshop*.

J. Naplava, M. Straka, J. Strakova, and A. Rosen. 2022. Czech Grammar Error Correction with a Large and Diverse Corpus. In *TACL*.

J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *LREC*.

F. Palma Gomez and A. Rozovskaya. 2024. Multi-reference benchmarks for russian grammatical error correction. In *EACL*.

F. Palma Gomez, A. Rozovskaya, and D. Roth. 2023. A low-resource approach to the grammatical error correction of ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop in conjunction with EACL*.

M. Ragheb and M. Dickinson. 2012. Defining syntax for learner language annotation. In *COLING*.

A. Rozovskaya. 2022. Automatic Classification of Russian Learner Errors. In *LREC*.

A. Rozovskaya and D. Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *ACL*.

A. Rozovskaya and D. Roth. 2019. Grammar error correction in morphologically-rich languages: The case of russian. In *Transactions of ACL*.

A. Sorokin. 2017. Spelling correction for morphologically rich language: a case study of russian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*.

O. Syvokon and O. Nahorna. 2021. UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language.

O. Syvokon and M Romanyshyn. 2023. The UNLP 2023 shared task on grammatical error correction for Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop in conjunction with EACL*.

V. A. Trinh and A. Rozovskaya. 2021. New dataset and strong baselines for the grammatical error correction of russian. In *ACL Findings*.

S. Watcharapunyawong and S. Usaha. 2013. Thai EFL students? writing errors in different text types: The interference of the first language. *English Language Teaching*, 6(1):67–78.

Z. Xie, G. Genthial, S. Xie, A. Y. Ng, and D. Jurafsky. 2018. Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction. In *NAACL*.

Z. Yuan and T. Briscoe. 2016. Grammatical error correction using neural machine translation. In *NAACL*.

Y. Zhang, Z. Li, Z. Bao, J. Li, B. Zhang, C. Li, F. Huang, and M. Zhang. 2022. MuCGEC: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. In *NAACL*.

W. Zhao, L. Wang, K. Shen, R. Jia, and J. Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *NAACL*.

Appendix

Dep. type	**Rel. freq. (%)**	
	Orig. sents.	**Corrected sents.**
punct. (↓)	17.4	18.5
case	9.9	9.7
nsubj (↓)	8.8	9.3
obl.	7.9	7.8
nmod (↓)	7.4	8.3
amod	7.2	7.1
conj.	5.4	5.3
advmod (↑)	5.4	4.8
cc (↑)	4.5	4.0
mark (↑)	4.0	3.7
det	3.7	3.5
obj.	3.5	3.6

Table A.1: List of most frequent relation labels in the 500-sentence sample annotated with dependency relations. ↑ indicates a relation that is overused in the original learner data, whereas ↓ indicates an underused relation, compared to the corrected sentences (we use a difference of 0.3 or greater in the relative frequency to define an overused/underused relation).

Noun case

Это	зависит	от	*показания/показаний	очевидцев
This	depends	from	$testimony_{gen,*}$**sg**$/_{gen,}$**pl**	$eyewitness_{gen,pl}$

'This depends on the testimony of eyewitnesses'

Preposition

Слова	*от/из	прошлых	уроков
$word_{nom,pl}$	***from/out of**	$previous_{gen,pl}$	$lesson_{gen,pl}$

'Words from previous lessons'

Verb agreement (number)

Все	новые	здания	*разваливается/разваливаются	
All	$new_{nom,pl}$	$building_{nom,pl}$	$*fall_{pres,imperfect,}$**sg**$/fall_{pres,imperfect,}$**pl**	$apart$

'All new buildings are falling apart'

Verb agreement (gender)

Лера	*пробовал/пробовала		флиртовать	с	ним
Valerie	$*try_{past,imperfect,}$**masc**$/try_{past,imperfect,}$**fem**		to flirt	with	him

'Valerie tried flirting with him'

Lexical choice (word)

Тогда	люди	стали	*спрашивать/задавать	вопросы
Then	$people_{nom,pl}$	started	$*toinquire/$**to ask**	$questions_{acc,pl}$

'Then people started to ask questions'

Morphology (deriv.)

Такие	окна	не	*пускают/пропускают	свет
Such	$windows_{nom,pl}$	do not	$*allow_{animate}/allow_{}$**inanimate**	light

'Such windows do not allow light'

Missing word

Много	необходимо	сделать	*∅/чтобы	решить	эту	проблему
$Much_{nom}$	must	to do	$*\varnothing/$**in order**	to solve	this	$problem_{acc,sg}$

'A lot needs to be done to solve this problem'

Table A.2: Examples of common errors in the Russian learner corpus. Incorrect words are marked with an asterisk.