

Universal Dependencies for Learner Russian



Alla Rozovskaya

Queens College
City University of New York

Studies on learner language

- A lot of work on non-standard texts written by language learners
- Most work focuses on grammatical error correction (GEC)
 - Most of the research is devoted to errors made by learners of English



Studies on learner language

- More recently, learner datasets in other languages annotated for errors have been created
 - Czech
 - Arabic
 - Ukrainian
 - Russian



Challenges of annotating linguistic structures and creating treebanks of learner language

- Very little work on annotating linguistic structure in learner language
 - Berzak et al. (2016) created a Treebank of Learner English manually annotated with dependency trees
- The annotation follows Universal Dependency formalism (Nivre et al., 2016)
- Berzak (2016) discuss challenges of correcting non-standard structures in English
 - They formulate an additional set of annotation conventions aiming at uniform treatment of ungrammatical learner language
 - Two-layer annotation is proposed for the annotation of the original and the corrected version of each sentence
- No resources available of learner language annotated for linguistic structure in other languages

Our contributions

- We present a pilot Learner Treebank of Russian
 - 500 sentence pairs
 - Both the source and the corrected version of each sentence are annotated, resulting in parallel dependency learner corpus
 - The sentences are taken from two datasets of learner Russian (RULEC, and RU-Lang8)
 - The annotated errors allow us to tie the non-standard structures and their treatment to the specific linguistic phenomena and language misuse
- We describe challenges arising from annotating non-standard syntactic constructions
- We propose how to handle non-standard constructions that are not found in standard Russian language specific to Russian
- We also identify error-specific challenges affecting the annotation of syntactic structures in learner data

Overview of the Russian grammar

- Russian belongs to the Slavic subgroup of the Indo-European language family
- It has highly fusional morphology
 - It includes case, gender, and number markings for adjectives, nouns, pronouns, and numerals
 - It also has a complex verb conjugation system
- Russian has free word order

Russian learner datasets

- We use two datasets of Russian learner data annotated for errors
 - RULEC-GEC (Rozovskaya and Roth, 2019)
 - **Essays written by learners whose native language is English**
 - RU-Lang8 (Trinh and Rozovskaya, 2021)
 - **Data from Lang8 website from a variety of learners and first language backgrounds**

Common Russian learner errors

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.

Common Russian 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”

Dependency annotation

- We annotated 500 sentence pairs
 - 300 sentences from RULEC-GEC
 - 200 sentences from RU-Lang8
- Since our focus is on the annotation of non-canonical structures, we excluded sentences that have no errors
- We then sorted the remaining sentences by the number of edits in decreasing order (excluding short sentences)
 - The average number of corrections in the annotated sample is 3.4 (although these can include spelling and punctuation errors that do not affect the dependency annotation)

Dependency annotation of non-standard structures

- We follow the UD guidelines
- The literal reading principle (Berzak et al., 2016)
 - The UD guidelines do not cover non-canonical structures arising due to learner errors
 - Non-standard structures are annotated based on the relations exhibited in the original sentence produced by the learner, and not based on the relations in the corrected version of the sentence
 - E.g. “we waited him” -> “him” is annotated as the direct object of the verb “wait”
- There are specific non-canonical structures that are observed in Russian learner data and we discuss their treatment next

Annotation of non-canonical structures

- Errors resulting in non-canonical structures
 - Preposition errors
 - Missing words (e.g. the verb есть/to be/to have)
 - Extraneous words
 - **Connectors, markers, and expletives**
 - Errors in derivational morphology

Distributions of dependency relations

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

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).

Conclusion

- We have presented the first pilot annotation for dependency relations of Russian learner data
- We have identified specific challenges of annotating non-canonical structures in learner Russian and proposed how to treat those within the UD framework
- Please see the paper for more detail!

Thank you!