CORPUS OF RUSSIAN STUDENT TEXTS: DESIGN AND PROSPECTS[1]

Zevakhina N.A. (natalia.zevakhina@gmail.com)

Dzhakupova S.S. (svetlanads@yandex.ru)

National Research University Higher School of Economics (Moscow, Russia)

Abstract

The Corpus of Russian Student Texts (CoRST) is a computational and research project started in 2013 at the Linguistic Laboratory for Corpora Research Technologies at HSE. It comprises a collection of Russian texts written by students from various Russian universities. Its main research goal is to examine language deviations viewed as markers of language change. CoRST is supplied with metalinguistic, morphological and error annotation that enable to customize subcorpora and search by various error types. Its error annotation is based on the modular classification: lexis, grammar and discourse, within which most frequent error phenomena are further distinguished. In total, the error classification encompasses 39 (20 higher-level and 19 lower-level) error tags. The crucial characteristic of CoRST is that the error annotation is multi-layered. Typically, since an error section can be corrected in a few ways, it is annotated with a few error tags respectively. Moreover, the corpus provides search by two possible explanation factors – typo and construction blending. The perspectives of CoRST development have both computational and research aspects, including qualitative and statistical comparative analysis of language phenomena in CoRST and NRC.

Key words: corpus, learner corpus, corpus linguistics, errors, error annotation

КОРПУС РУССКИХ УЧЕБНЫХ ТЕКСТОВ: АРХИТЕКТУРА И ПЕРСПЕКТИВЫ

Зевахина Н. А. (natalia.zevakhina@gmail.com)

Джакупова С.С. (svetlanads@yandex.ru)

Национальный исследовательский университет «Высшая школа экономики» (Москва, Россия)

Ключевые слова: корпус, учебный корпус, корпусная лингвистика, ошибки, разметка ошибок

[1] The results of the project "Corpus studies of language variation: from deviations to linguistic norm", carried out within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2015, are presented in this work.

1. *Introduction*

In the last 30 years, a new field of corpus linguistics known as learner corpora has been widely acknowledged. Usually, a *learner corpus* consists of texts written by people who study a second language (L2) and speak different first languages (L1). The language of such texts is not a target language L2 in a proper sense of a word; rather, it contains a great amount of L2 deviations, or errors. In Second Language Acquisition studies, the deviation variant of L2 is traditionally called *interlanguage* (Selinker 1983). The primary goals of learner corpora development are linguistic research and their educational application. The former includes (although is not limited to) investigating linguistic features of interlanguage and possible factors that underlie the errors usage, which are, according to Daniel and Dobrushina (2013), language interference (L1 affects interlanguage) and/or inherent non-trivial properties of target language L2 system open to deviations (e.g., non-trivial distribution of Nominative and Genitive in Russian object position, see ibid.: 195). Educational application of learner corpora encompasses developing textbooks, tests, and electronic resources for students.

The principal technical feature of learner corpora that makes them unique in corpus linguistics is the possibility to search by the type of errors, which are either only manually annotated (as in the majority of learner corpora, see Díaz-Negrillo and Fernández-Domínguez 2006) or partially automatically and partially manually annotated (e.g., CzeSL, see Hana et al. 2010).

Errors annotations are based on errors classifications that vary from corpus to corpus and essentially hinge upon the corpora goals. Thus, it is not surprising that there is no unified classification of errors in learner corpora research. However, all possible error classifications can be partitioned into several main types that we suggest to coin as *combinatorial*, *standard*, *modular* and *POS* (parts of speech) respectively.

Combinatorial error classification covers formal characteristics of errors: a proper word/construction can be omitted, has an incorrect form (or used in an incorrect order), or an error word/construction can be added. *Standard error classification* includes writing methods, i.e., orthography and punctuation. *Modular error classification* comprises all basic language levels: phonetics, lexis, grammar (morphology and syntax), discourse, and pragmatics (style). Among them grammar and lexis are most commonly distinguished. Further division within each level is also possible. Some learner corpora provide *POS error classification*. Several error classification types can be simultaneously implemented in the same corpus.

Error classification provides the basis for error description with a particular error tag. Besides, some corpora (e.g., Falko, see Lüdeling et al. 2005) might contain combinations of the three more annotation domains: error section (i.e., a text fragment where an error is identified), error explanation (i.e., a potential cause of why the error occurred), and/or error correction (i.e., in what way the text fragment can be amended).

The amount of error tags in learner corpora range from 30 to 100.

2. CoRST: Main goals and general description

The Corpus of Russian Student Texts (CoRST; http://web-corpora.net/CoRST/)[2] is part of a general project of non-standard variants of Russian developed by the Linguistic Laboratory for Corpora Research Technologies (headed by Ekaterina Rakhilina) at the National Research University Higher School of Economics, Moscow. The project started in 2013 and includes the following corpora subprojects: (i) CoRST that contains the texts written by Russian native speakers who study at various Russian universities; (ii) Russian Learner Corpus that is a collection of texts written by heritage and L2 speakers (see Vyrenkova et al. 2014); (iii) Corpus of regional variants of Russian (see Daniel and Dobrushina 2013).

Unlike Russian Learner Corpus and most of the learner corpora that consist of L2 texts, CoRST is a collection of L1 speakers' texts. Some of the authors are bilinguals and their second L1 might affect their Russian writing. Bilingual status of texts is, however, captured by metalinguistic annotation (see section 3.1).

The texts of both L1 monolingual and bilingual authors of CoRST contain errors and plausible reasons for that fact are as follows. First, the texts demonstrate the phenomenon similar to *imperfective learning* (Thomason 2008). While studying at the university, young people have to produce texts of various genres they are unfamiliar with. Acquisition of novel genres yields producing an impressive amount of errors. Second, part of the errors can be explained by wrong interpretation of communicative situation. Most students write fast, never proofread or improve their own texts. However, even if they do so, not many of them are able to make corrections. Third and the most important, we assume that at least part of the errors (especially, the most frequent ones) reflects internal language processes. In particular, this suggests that errors can be regarded as instances of language change (Glovinskaya 2000, Rakhilina in press). The supporting evidence comes from the fact that the corpus texts' authors are quite young (17-25 years old) and the peculiarities of their language might significantly differ from the norms of elder generations.

Consequently, the primary research goal of CoRST is to study errors as markers of language change. Moreover, each error type is supposed to have limits of variation and the studies based on CoRST data are aimed at capturing them. In accordance with the goal, CoRST has the following distinctive features.

[2] The corpus software developers are Timofey Arkhangelsky and Elmira Mustakimova.

First, like the Russian National Corpus (RNC), CoRST is an open web-resource available to everyone. In order to preserve author rights, corpus texts are all anonymous. Moreover, corpus users cannot see full texts since the maximal available context for any query is limited to five sentences.

Second, CoRST does not provide an annotation layer for corrections since most of the error texts cannot be amended just in one way. Another reason for this is that corrections often affect more than one word, i.e., they may require reorganizing a whole construction and (at least for now) that might cause technical causes while texts annotating. Furthermore, the principle of a multi-layered annotation, which we have been adhering to while developing the corpus, provides clues of how exactly a given error can be corrected.

Since 2013, the size of CoRST has reached about 2 639 298 tokens (2649 texts). Part of the corpus containing about 500 000 tokens is error annotated: in total, about 10 246 errors are tagged. The next section discusses main principles of CoRST annotation.

3. Annotation

The texts of CoRST are supplemented with metalinguistic, morphological and error annotation.

3.1. Metalinguistic annotation

Metalinguistic annotation contains information about a text (a type of a text, year, semester/module) and its author (age, gender, first language, region of residency, faculty/department, year of studying, bachelor/master, academic major). The corpus includes the following types of texts: course paper, bachelor or master thesis, abstract, essay, report, summary, autobiography, paragraph, application, business proposal[3]. The texts are written by students of the following academic majors: history, linguistics, logistics, political science, sociology, economics, journalism, psychology, law, philology, design, management, culture studies, and mathematics.

3.2. Morphological annotation

Morphological annotation is carried out automatically with help of the morphological analyzer MYSTEM (Segalovich and Titov 1997-2014). However, morphological ambiguity is not resolved: every ambiguous word is provided with all possible grammatical analyses. The tag set of 52 morphological labels meets the standards established by RNC.

[3] We are planning to add some more text types to the list.

3.3. Error annotation

CoRST is supplied with modular error annotation that, among other error markup systems based on error classifications (see section 1), seems to be the most relevant for linguistic research. Due to morphological annotation, we do not make use of POS classification. The combinatorial way of error annotating is partially implemented in the modular system. Omission corresponds to ellipsis. Since Russian word order is relatively free and reflects topic-focus structure of a text fragment, incorrect word order belongs to the discourse layer. Abundant words are errors of the discourse layer, as well. Incorrect words are viewed as lexical errors. We also mark up the beginning and end of each citation in order to rule out the fragments not authored by a student from morphological search.

As for standard error classification, CoRST tag set does not contain labels for orthographic or punctuation errors. Surely, they do occur in student texts. However, first, they do not constitute a large set since the majority of students are literate and their orthographic errors are primarily typos that we mark up. Second, they do not meet key research CoRST objectives. Last but not least, annotating them would not facilitate the morphological analyzer as much as disambiguation can do and, in this regard, we are planning to integrate the latter as well as automatic correction of typos.

Moreover, following some of the learner corpus developers (e.g., Lüdeling et al. 2005), we employ a multi-layered annotation system.

First, the text fragment that contains an error may be improved in different ways and, consequently, it may have several different error tags; in that case, all possible tags are provided.

Second, in addition to tags by which error types are classified, we also introduce tags that provide a possible explanation for some errors. Another annotation layer is so-called “weight” of an error. To the best of our knowledge, this layer is not implemented in any learner corpora yet. Further, we plan to incorporate this level into our annotation system.

In total, CoRST annotation comprises 39 tags: 20 higher-level tags and 19 lower-level tags. When appropriate, both level tags are specified.

The error annotation is carried out manually using the interface of *Les Crocodiles 2.6* (Arkhangelsky 2012).

3.4Error classification

While developing the error annotation system, we tried to balance theoretical views on error classification and practical purposes, i.e., annotation convenience. In doing so, we design the following modular error classification: lexical, grammatical, discourse, and stylistic errors. Except for style, all other error types correspond to the well-established language levels: lexis, grammar, and discourse. Since the distinction between morphology and syntax seems to be debatable and mainly hinges upon a particular framework, we unified them under the grammar label.

What is important is that linguistic errors have been traditionally viewed as a vague zone of language studies; the same applies to its opposite – the linguistic norm. For us, the basic reference point in this regard is academic texts, since the corpus was intended to be a collection of academic texts (see section 3.1). It goes without saying that some of the errors exemplified below can occur in texts of various genres, however, their occurrence in academic texts does not seem to be appropriate, at least according to stylistics handbooks and to our colleagues who run the courses of academic writing at universities. Another reference point is RNC. Presumably, a potential error should not occur or should not have many occurrences in RNC, which comprises texts that are supposed to represent the linguistic norm. However, first, genres, styles, and time intervals of texts should definitely be taken into account since what is the norm for texts of one genre or time interval is not considered to be so for texts of another genre or time interval. Second, there does not seem to exist a quantitative criterion: how many occurrences should a linguistic phenomenon have in RNC in order to be considered as a norm? 10, 100, or even 1000? All these questions deserve future work.

In CoRST, lexical errors, first, comprise errors in any given word or phrase. Second, they include four specific error types: intensifiers, metonymy, nominalizations, and so-called “light” verbs (semantically they are empty and usually accompany nominalizations). Third, word-formation errors (i.e., paronymy and aspectual errors) are associated with lexical errors. As for the aspect, we follow the most influential tradition that ascribes Russian aspect to word formation rather than to inflection, i.e., considers Russian aspectual verbal pairs as different lexical units.

Sentences (1) and (2) exemplify lexical errors. In (1), a non-standard form *udobnost’* instead of *udobstvo* ‘convenience’ is employed. (2) illustrates wrong use of the light verb *nesti* in place of the standard light verb *vlech’* ‘entail’. See (6) and (10) for some further lexical errors.

(1) *Bezuslovno, my ne mozhem otritsat’* ***udobnost’*** *testov, potomu chto oni pomogayut prepodavatelyam znachitel’no sekonomit’ vremya kak na samom ekzamene, tak i pri dal’neyshey proverke rabot.*

‘Undoutedly, we cannot deny convenience of using tests since they help teachers to save a lot of time both at the exam and at further checking students’ papers.’

(2) *Kuril’shchiki bol’she ostal’nykh lyudey podverzheny tuberkulezu, zabolevaniye kotorym* ***neset*** *za soboy tyazhelye posledstviya.*

‘Smokers are exposed to tuberculosis to a greater extent than non-smokers; being sick with this may have severe outcomes.’

Grammatical errors comprise deviations in comparative and superlative constructions, coordination and subordination errors (sentential arguments and relative clauses among the latter ones), reference violations (including errors in pronouns and converbs), ellipsis, construction violations, agreement and government errors, nominal and verbal inflection errors, errors involving voice and argument structure alternations (reflexives, decausatives, etc). A few brief comments need to be made here. First, by inflection errors we mean all inflection errors, apart from agreement and government errors labeled with a different tag. Second, voice and argument structure alternations semantically exhibit different grammatical phenomena; however, morphologically at least some of them are marked with the suffix *-sya* in Russian. This is the main reason for why we unified them under the same label.

Grammatical errors are illustrated in (3)-(5). In (3), the author employs the non-standard combination of comparative connectors *nezheli chem* ‘than’ instead of using a standard one – either *chem* or *nezheli*. Moreover, the subject of the sentence *ponyatiye* in the singular form does not agree in number with the verb *poyavlyayutsya* ‘appear’ that is in the plural form.

(3) *V XXI veke u chelovechestva* ***poyavlyayutsya*** *sovsem drugoye ponyatiye o sem'ye, o kar'yere,* ***nezheli chem*** *u sovetskikh grazhdan.*
‘In the XXI century, the humankind understands the notions ‘family’ and ‘career’ quite differently than the Soviet people.’

In (4), the null subject of the converb *oglyadyvayas' na real'nost'* ‘taking into account the real facts’ and the subject of the main clause *yazyk* ‘language’ are not coreferential. Semantically, the null subject can refer to the noun phrase *chelovek* ‘human’.

(4) *Yazyk kak neot'emlimaya chast' cheloveka,* ***oglyadyvayas' na real'nost'***, *takzhe preterpevayet nemalye izmeneniya.*[4]
‘Language as a human essential, taking into account the real facts, also undergoes considerable change.’

In (5), the wrong combination of the pronoun *to* and the standard complementizer *chto* introduces the second sentential argument. *To chto* is a well-spread phenomenon in colloquial modern Russian that should be interpreted as a new complex complementizer (see Korotaev 2013 among

[4] We cite examples in their original orthographical and punctuation forms.

others). The evidence comes from coordination of the embedded clauses introduced with *chto* and *to chto* respectively. Also, the main clause contains a prepositional form *s tem* of the same pronoun *to*, which is completely impossible in standard Russian.

(5) *Ya soglasen s tem, chto ya ne stalkivalsya so mnogimi problemami, s kotorymi stalkivalis' drugiye lyudi, i* **to chto** *ne mne ikh sudit'.*
'I agree that I haven't faced many problems other people faced and that it is not for me to pass judgment on them.'

Discourse errors include errors in meta-textual comments, mixing direct and indirect speech, incoherent sentences, tautology and pleonasm, parcellation (division of sentences into uncompleted units), logical errors, wrong use of linking words, wrong word order, and inappropriate topicalization.

Consider the following examples. (6) illustrates the phenomenon of parcellation: the connector *to yest'* 'that is' cannot take the initial position in a sentence, like, e.g., *drugimi slovami* 'in other words'. Rather, it has to be a continuation of the first sentence[5].

(6) *Primery, privedennye v otryvke, naglyadno pokazyvayut, naskol'ko chasto lyudi nesposobny primenit' raneye* ***zauchennye znaniya*** *kogda eto, deystvitel'no, neobkhodimo.* ***To yest'*** *ne sopostavlyayut teoriyu s praktikoy.*
'The examples cited in the excerpt clearly show how often people are unable to apply the acquired skills when this is truly needed. That is, they don't associate theory with practice.'

In (7), the connector *takzhe* 'also' takes a non-standard sentence initial position[6]. Interestingly, English *also* is appropriate in that position.

(7) ***Takzhe*** *tekst mozhet byt' prednaznachen dlya uchenykh, studentov, zanimayushchikhsya issledovaniyem povedeniya cheloveka na dorogakh.*
'Also, the text can be intended for researchers and students who investigate psychological aspects of driving.'

[5] There is yet a lexical error in (6) – *zauchennye znaniya* 'received (lit. learnt) knowledge'. A plausible explanation for this wrong collocation is that if one acquires some new information (e.g., rules), she usually learns it.

[6] According to (Baranov et al. 1984: 214), Russian *takzhe* has two functions and none of them is realized in a clause initial position.

Sentence (8) illustrates a pleonasm error (*perezhivat' i nervnichat'* 'be anxious and nervous') tagged as tautology errors.

(8) *V sovremennom mire nam chasto prikhodit'sya* ***perezhivat' i nervnichat'*** *iz-za chego stradayet nasha nervnaya sistema*.
'In modern world, we often have to be anxious and nervous; the nervous system suffers from that.'

Stylistic errors encompass any mismatch between the style of a text and its type, primarily inappropriate use of official or colloquial style. To illustrate, in an outdoor advertisement, official style constructions are regarded as infelicitous; in an essay or a course paper, a colloquial style construction is not tolerated, see *u menya yest' voprosy* 'I have questions' in (9). Any redundancy is marked up, as well. For example, in (10) every single meta-textual commentary fits the official text style but one can hardly use three of them in the same sentence[7].

(9) *K tomu zhe* ***u menya yest' voprosy*** *po klassifikatsiyi osey: pochemu v neprostranstvennuyu os' popadayut prostranstvennye primery s glagolami dvizheniya?*
'Besides, I have questions about the classification of axes: why do spatial examples with the motion verbs illustrate the spatial axis?'
(10) ***Takim obrazom, podvodya itog vysheskazannomu, vazhno zametit'****, chto slovo yavlyaetsya* ***osobym klyuchevym instrumentom v*** *upravleniyi lyud'mi.*
'Thus, summing up, it is worth noting that a word is a special key tool in management.' (lit.)

The explanation level has two types: typo and construction blending exemplified in (11) and (12) respectively. The latter seems to be very important since discussion of errors makes sense only with respect to a given construction environment.

(11) illustrates an agreement error in gender between the possessive *moya* 'my (fem.)' and the noun phrase *lichnoye mneniye* 'personal opinion (neut.)'. Being marked with an agreement tag,

[7] The sentence also contains another error sequence – *osobym klyuchevym instrumentom v upravleniyi* 'a special key tool in management' (lit.). First, there appears to be a pleonastic error (*osobyy* and *klyuchevoy*); second, there is a government error (*instrument v upravleniyi*); third, there seems to be a lexical error (*instrument* instead of *sposob* 'way'). The ways in which the whole sentence can be improved are, e.g., as follows: *Slovo yavlyaetsya odnim iz klyuchevykh instrumentov, s pomoshch'yu kotorogo mozhno upravlyat' lyud'mi* 'A word is one of the key tools with help of which one can manage people' (lit.) or *Vozdeystviye pri pomoshchi slova – eto odin iz klyuchevykh sposobov upravleniya lud'mi* 'Influence with help of a word is one of the key ways to manage people' (lit.).

this error sequence is most naturally explained in terms of a typo since a native speaker can hardly make this error.

(11) ***Moya*** *lichnoye mneniye po povodu dannoy problemy ochevidno.*
'My (fem.) personal opinion (neut.) on this problem is obvious.'

(12) exemplifies *vypolnit' tsel'* as a result of blending of the two standard constructions *vypolnit' zadachu* 'perform a task' and *dostich' tseli* 'achieve a goal'.

(12) *Ya shchitayu, chto mne, pust' i ne polnost'yu, no vse-taki udalos'* ***vypolnit'*** *postavlennuyu mnoyu* ***tsel'****.*
'I think I succeeded in achieving the goal I set, although not fully.'

4. Conclusion

The perspectives of CoRST development have both computational and research aspects. In what follows, we list only crucial ones. One computational goal is to implement the identification system that helps backtrack students' individual errors and customize students' individual learner corpora. Since error annotating involves offline consistent cooperation of an annotator[8] and an expert who work on the same whole text one by one, another goal is viewed as online rechecking by an expert who sees all the contexts that contain errors with an identical tag. Besides, as long as the annotation convention is to tag the opening word of an error phrase only, the third goal is to mark up the whole error section. We see the fourth goal as further developing the system of multi-layered annotation, whereby alternative interpretations of an error section are visually separated on the screen (for the time being, all the alternative tags are listed in one string).

With assistance of students, we are currently working on a few research case studies. One of the main research principles is to detect and account for language tendencies in modern Russian in comparison to NRC, regarding identical metalinguistic information (e.g., year and genre).

References

Arkhangelsky T.A. (2012), Les Crocodiles 2.6. [Software]. Moscow.

[8] Most of the annotators are students of the School of Linguistics at HSE who do this work for university practice credits.

Baranov M.T., Kostyaeva T.A., Trubnikova A.V. (1984), Russian language. Handbook for students [Russkiy yazyk. Spravochnik dlya uchashchegosya], Moscow, Prosveshcheniye.

Daniel M.A., Dobrushina N.R. (2013), A corpus of Russian as L2: the case of Daghestan [Russkiy yazyk v Dagestane: problemy yazykovoy interferentsii], Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference “Dialogue 2013” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy ezhegodnoy Mezhdunarodnoy Konferentsii “Dialog 2013”], Bekasovo, Vol. 1, pp. 186–199.

Díaz-Negrillo A., Fernández-Domínguez J. (2006), Error Tagging Systems for Learner Corpora, Spanish Journal of Applied Linguistics [Revista Española de Lingüística Aplicada], Vol. 19, pp. 83–102.

Glovinskaya, M.Y. (2000), Active processes in grammar [Aktivnye protsessy v grammatike], The Russian language in the end of the 20th century [Russkiy yazyk kontsa XX stoletiya], Yazyki russkoy kul’tury, Moscow, pp. 237–304.

Hana J., Škodová S., Rosen A., Štindlová B. (2010), Error-tagged Learner Corpus of Czech, Proceedings of the Fourth Linguistic Annotation Workshop, Uppsala, pp. 11–19.

Korotaev N.A. (2013), Clauses-combining with *to chto* in spoken Russian [Polipredikativnye konstruktsii s *to chto* v nepublichnoy ustnoy rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference “Dialogue 2013” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy ezhegodnoy Mezhdunarodnoy Konferentsii “Dialog 2013”], Bekasovo, Vol. 1, pp. 358–367.

Lüdeling A., Adolphs P., Kroymann E., Walter M. (2005), Multi-level error annotation in learner corpora, The Corpus Linguistic Conference Series, 1(1), pp. 105–115.

Rakhilina, E.V. (in press), Comparative degrees in light of the Russian grammar of errors [Stepeni sravneniya v svete russkoy grammatiki oshibok], Trudy Instituta Russkogo Yazyka imeni V.V. Vinogradova, West-Consulting, Moscow.

Segalovich I., Titov V. 1997-2014. Automatic morphological annotation MYSTEM. [Software]. Available from https://tech.yandex.ru/mystem/

Selinker L. (1983), Interlanguage, Second Language Learning: Contrastive analysis, error analysis, and related aspects, The University of Michigan Press, Ann Arbor, MI, pp. 173–196.

Thomason S. (2008), Pidgins/Creoles and Historical Linguistics, The Handbook of Pidgin and Creole Studies, Blackwell Publishing Ltd, pp. 242–262.

Vyrenkova A.S., Polinsky M., Rakhilina E.V. (2014), Grammar of errors and construction grammar: “heritage” Russian [Grammatika oshibok i grammatika konstruktsiy: “eritazhnyy” (“unasledovannyy”) russkiy yazyk], Voprosy yazykoznaniya, Vol. 3., pp. 3–19.