**Russian in the English mirror: (non)grammatical constructions in learner Russian**

1. Introduction

Learner corpora have truly become an irreplaceable resource in the study of second language acquisition (SLA) and second language pedagogy in the recent decades.

Although the majority of learner corpora to date represent English as a Foreign (FL) or Second (L2) language, many well-designed corpora of learner languages other than English have appeared in the past decade.

This new linguistic resource is now available for researchers of L2 Russia: Russian Learner Corpus (RLC, <http://web-corpora.net/heritage_corpus>) is an annotated multilingual corpus. RLC comprises texts produced by two categories of non-standard speakers of Russian: learners of Russian as a Foreign language and speakers of Heritage Russian. RLC is a collaborative project between the Heritage Russian Research Group (Higher School of Economics) under E. Rakhilina and a team of American researchers associated with the Heritage Language Institute (M. Polinsky, O. Kisselev, A. Alsufieva, I. Dubinina, and E. Dengub).

RLC includes texts produced by RFL and RHL learners with various native/dominant languages (English, German, Italian, Finnish, French, and Korean). Currently, most texts entered into the corpus are authored by learners with English (US) as dominant language: for example, the largest sub-corpus is RULEC (Russian Learner Corpus of Academic Writing), a longitudinal corpus of advanced student writing (compilers O. Kisselev and A. Alsufieva, the number of words: ~ 400 000) (Alsufieva, Kisselev, and Freels, 2012). It was created by speakers of FL Russian and speakers of Russian as a Heritage language (HL), across different levels of language command, linguistic modes (written and oral) and genres. Data on Heritage Russian oral production include the results of experimental studies: frog stories (based on the methodology developed by Berman & Slobin 1994; Slobin 2004) and narratives based on a short cartoon (“Nu pogodi!”) (see Isurin & Ivanova-Sullivan (2008) and Polinsky (2008) for more details).

Corpus data and its flexible search system provide a sound basis for comparative research in Heritage and L2 production and enables a deeper insight into complicated phenomena, such as non-standard use of Russian aspect, cases, prepositional phrases, as well as lexical and semantic misuse in multi-word constructions.

Apart from telling a lot about non-standard Russian, RLC is a powerful tool for opening new facets of Standard Russian grammar: deviations in language use help uncover subtle rules that previously have been paid no attention to.

Moreover, the new corpus provides a unique opportunity to conduct not only comparative studies in Russian SLA and pedagogy, but also methodological studies that have relevance for learner corpora annotation, analysis and management.

1. Error analysis of Learner Russian

The idea of usefulness of error analysis has been largely -- if not uncritically -- embraced by the field of Learner Corpus research (Granger 1998).

The preliminary linguistic analysis and tagging at RLC were done by the members of the Heritage Russian Research Group (Higher School of Economics), with technical support provided by Timofey Arkhangelsky and Elmira Mustakimova.

The texts of RLC include linguistic and meta-linguistic information. Linguistic tags include POS annotation and error annotation.

* 1. Meta-linguistic tags include information about **authors** (gender, level of RL command, L1 or dominant language, etc.) and about **texts** (genre, mode, course name, time restriction, etc.)

<http://web-corpora.net/heritage_corpus/admin/annotator/document/add/>

* 2. Linguistic tags
  + 2.1. POS annotation aided by Mystem tagging software (Yandex)
  + 2.2. Error annotation (manual)

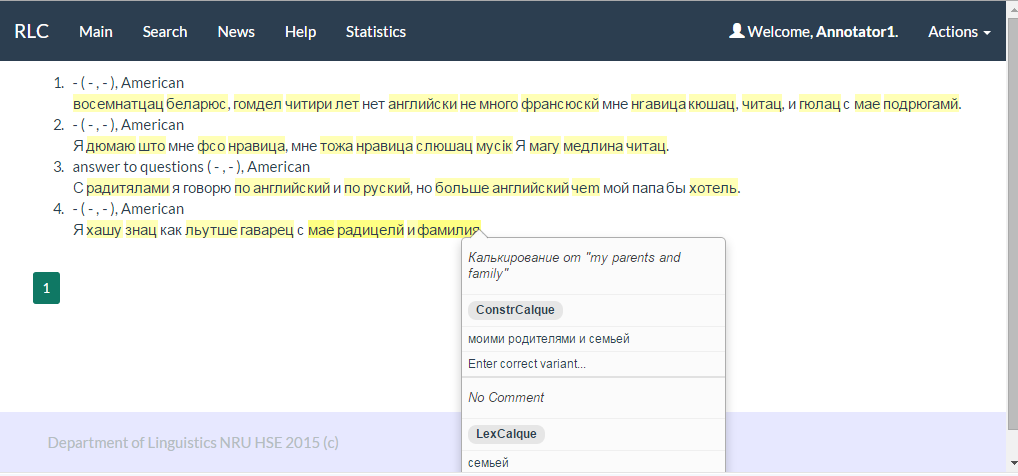
The annotation platform is a compilation of tools that are integrated into one online interface. Through a number of annotation steps, the platform allows for detection and tagging of formal errors, as well as lexico-semantic, syntactic, discourse and pragmatic errors. The system keeps both the original and the corrected versions of each sentence, allowing to search by standard forms. Moreover, the system allows to track tagger comments which he/she can leave while tagging errors: the first field is designed for comments, the second one is for tags, and the third one is for correct Russian variant (see Figure 1).

The steps of analysis are:

1. texts are automatically parsed into sentences and are tokenized;
2. word forms are automatically POS-tagged;
3. annotated files are sent into the tagged database;
4. the files are channeled into the interactive interface for manual error annotation.

<http://web-corpora.net/heritage_corpus/document-annotations/>

Figure 1. Error tag sample (with tagger’s comments)



The users of RLC can be divided into two groups: registered and trained taggers and registered users. Their access levels to the RLC platform are different:

* **Registered and trained taggers** have a full access to all operations within the error-tagging platform, such as adding and tagging new texts, editing previously tagged texts, and correcting the tags.
* **Registered users** have access to tagged files; can search the files by error types, tags, wordforms, ect.; can restrict sub-corpora by genre, task, and other metalinguistic information; have access to statistical data.

We have developed a new classification of error tags to clarify its structure (e.g. the correlation of tags at different levels). The previous classification contained tags of different hierarchical levels at the same level that led to divergence of tagging of the same cases (e.g., often substitution of such tags as Syntax and Coord, Morph and Asp). The detailed description of a new system of classification was made by D. Loshkareva in her MA thesis written under the supervision of E. Rakhilina (Loshkareva, 2015):

**1. Mechanical errors:**

1.1. Graphic errors, Graph;

1.2. Phonetic and Orthographic errors, Orpho/Phon;

1.3. Mistakes and typos, Typo;

1.4. Code switching, Cs;

**1.5. Lexical errors, Lex.**

**2. Grammatical errors:**

2.1. Morphological and derivational errors, Morph;

* + 2.2. Morphological and derivational errors in Nouns and Adjectives, NF/AF;
  + 2.3 Morphological and derivational errors in Verbs, VF;
  + 2.4. Aspect errors, Asp;
  + 2.5. Tense errors, Tense;

**3. Grammatical errors:**

3.2. Syntactic errors, Syntax;

3.2. Errors in verbal government, Gov;

3.3. Errors in agreement, Coord;

3.4. Errors in reference, Ref;

3.5. Word order errors, WO;

**4. Errors in constructions**, Constr;

**5. Errors as a result of L1 interference: calques and loans.**

5.1. Orthographic calque, OrphoCalque;

5.2. Morphological calque, MorphCalque;

5.3. Lexical calque, LexCalque;

5.4. Constructional calque, ConstrCalque;

5.5. Syntactic calque, SyntaxCalque;

5.6. Quasicalque Qs;

5.7. Loan translations, LoanTr.

Tag types are not hierarchical; two or more types can be assigned to one error (e.g. Morph + Constr). If the error type is unclear and could possibly be assigned to more than one category, a slash mark is used (e.g. Orpho/Phon):

E.g. \* vam podderzivaet drug postoyanno

\* you (DAT) support (VPres) friend (NOM) always

drug (NOM) vas (ACC) postoyanno podderzivaet (VPres)

friend (NOM) you (ACC) always support (VPres)

In this case there are two mistakes in one construction (Gov+WO): incorrect government (\* *vam* DAT instead of *vas* ACC and non-grammatical word order (\*Pron ACC +V+N NOM+ADV instead of N NOM + Pron ACC + ADV + V).

When choosing the right tag for a mistake a tagger should strive to use a tag of lower levels rather than of higher ones (if any). For instance, incorrect usage of word order is marked as WO (Word Order) not Syntax (v.s.).

It is worth noting that we can use one annotation for a combination of words in our corpus (see Figure 1). There you can find an example of construction calque where all five words have one tag Constrcalque:

\**гаварец* *с мае радицелй и фамилиа*

*говорить с моими родителями и семьей*

*to talk with my parents and family*

Thus a tagger can use one tag for errors in constructions or syntax (but only in one sentence).

The platform of RLC is working in testing mode now.

**3 Typical errors in constructions**

One of the most difficult language structures for acquisition is a construction. By construction we mean "a combination of words which has an aspect of the plane of content and the plane of expression, not equal to the meaning or form of its components» (Fillmore, 1988; Goldberg, 2006; Rakhilina, 2010), i.e., a complex in which the semantic and syntactic elements are closely related to each other. According to E. Rakhilina, very often heritage speakers not knowing how to express some meaning correctly, choose the following strategy: they do not turn to the familiar patterns of the dominant language, and build their own structure (Rakhilina, 2008). Corpus data suggest that such language behavior is not typical for native speakers.

The mistakes in constructions give us a lot of material for analysis. Here are some interesting frequent construction calques (metaphoric construction, discourse words, and syntax structure) (see Figure 2).

Figure 2. Examples of frequent constrcalques.

|  |  |  |
| --- | --- | --- |
| Russian construction | Frequency (450 phrases) | Examples |
| Metaphoric construction быть рядом (\*там за – be there for smb) | 9 | \*Друг – человек, который всегда там за вас, который хочет вас слышить.  Друг – это человек, который всегда рядом… |
| Discourse words (\*в дополнение – in addition, etc.) | 7 | \*В дополнение, верные друзья почти всегда храбрые люди, которые ничего не боятся.  Кроме того, … |
| Comparative and parallel construction: \*не тот же самый (not) the same - разный | 7 | \*Но, не смотря на тот, что мы не живём в тот же самом городе, я всегда могу позвонить ей по телефону и поговорить с ней.  Но, несмотря на то, что мы живем в разных городах, … |
| eto+ADV+inf | 3 | \*С этим человеком, это не трудно говорю, потому что мы понимаем друг друга.  С этим человеком не трудно говорить, … |

Today I would like to talk about these two constructions - (1) и (2).

(1) \* eto vredno svoim pal’cam

*\* it is bad one’s* DAT PL  *fingers* DAT PL

cf. eto vredno dlya PREP pal’cev GEN PL

*it is bad for fingers*

*\*Но, по-моему, это вредно своим пальцам, поскольку часто встречающиеся буквы не находятся близко к центру клавиатуры* (L2 speaker)

*But I think it is bad for one’s fingers since the most frequent letters are not located towards the center of the keyboard.*

(2) \* eto ne trudno govorit’

*\* it is not hard to speak*

cf. NULL ne trudno govorit’

*(it’s) not hard to speak*

\*С этим человеком, это не трудно говорю, потому что мы понимаем друг друга. (L2 speaker)

*With this person, it’s not hard to speak because we know each other.*

In analyzing errors like these, we attempt to establish those patterns and rules present in the interlanguage of the learner that allow us to hypothesize (and in some cases predict) the source of the non-native-like construction. Thus, in example 1, the likely source of error is the English (albeit infrequent) construction *to be bad (harmful)+to+something.* For instance:

(a) *Ayscough felt that white glass created an offensive glaring light that was bad to the eyes.* ([GloWbE)](http://corpus.byu.edu/glowbe/help/intro_e.asp?w=1024&h=600)

(b) *On the other hand, we may find out 3D is truly harmful to children's eyes, at which point it will likely lose the interest of the public and die.* ([GloWbE)](http://corpus.byu.edu/glowbe/help/intro_e.asp?w=1024&h=600)

The transfer is likely to be supported by the existence of two possible constructions in Russian as well, *dlya*(cf. for)+GEN and NULL PREPOSITION+DAT. These two constructions are close semantically and may be interchangeable (Ladygina 2014) under the right circumstances, i.e. if the experiencer is animate (Bonch-Osmolovskaya 2003). In example 1, the requirement of animacy is not upheld (likely because no such restraint exists in English). Interestingly, HL learners (at least at advanced levels) appear to be sensitive to the restraint of animacy and do not exhibit errors of this type.

Additionally, the Russian National Corpus contains similar pattern not only for adverbs HARMFUL FOR and HARD FOR but also for 24 other adverbs (N=100 (see Figure 3). Hence, the difficulty for the learner is not only in the semantics of the adverb but in the pattern itself.

Figure 3. Russian National Corpus: frequent constructions ADV+FOR

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Frequency** | **Number of documents** | **construction** |
| 1 | 1251 | 960 | NEEDED FOR |
| 2 | 1019 | 845 | REQUIRED FOR |
| 3 | 971 | 809 | ENOUGH FOR |
| 4 | 602 | 511 | NOT FOR |
| 5 | 455 | 410 | JUST FOR |
| 6 | 409 | 371 | NOTHING FOR |
| 7 | 407 | 365 | BETTER FOR |

Example 2 (ETO+AVD+INF) demonstrates that L2/HL can make a mistake in this construction regardless an adverb. This mistake belongs to a type of learner errors known as null subject errors; it is frequently mentioned in the works on negative transfer. Although this error type is most often explained by the negative transfer from English, persistence of such errors in HL interlanguage indicates that it is also preempted by the fact that Russian allows for pronoun *eto* in certain constructions, i.e. INF(COP)+ADV-o/ INF (COP) – *eto +*ADV-o:

(c) Купить в супермаркете пищу и из-за нее потом едва не протянуть ноги – это сейчас несложно. (Russian National Corpus)

*To buy groceries in a supermarket and then almost die as a result – it’s not difficult these days.*

More importantly, the previous research in this area of grammar disregarded diachronic development of the use of *eto* in the Russian language. Thus in the main corpus of the Russian National Corpus, we find the following dynamic: in the text authored in the 19th century, the frequency of *eto*-construction is 1.4\*10^-5, in the 20th century texts it becomes 2.87\*10^-5, and in the texts authored in the first decade of the 21st century the frequency reaches 3.35\*10^-5. Thus, the construction under examination has become 105.3% more frequent in the 20th century when compared to the 19th century, and 16.4% more frequent in the 21st when compared to the 20th.

(d) Думаешь, это было просто ― бросить все и прилететь сюда? (Russian National Corpus)

*You think it was easy – to drop everything and fly here?*

However, when it comes to the examination of oral sub-corpus of the RNC, we find the construction ETO+ADV-o – INF(COP):

(e) Это тяжело очень сказать / когда достроят (Russian National Corpus)

*It is hard to say / when they will finish building.*

Additionally, in constructions that employ *kak* (cf. how) the word order is the same as in English: *kak*+ *eto*+ADV – INF (cf. Eng., how it is+ADJ+INF)

(f) … как же это сложно: говорить так, чтобы тебя слышали и слышали. (Russian National Corpus)

*… how it is difficult – to speak in a way that you are listened to and heard.*

In other words, the learners (and error-taggers) have to follow two sets of rules for *eto*-constructions: one in writing, another in speech.

Such error analysis is not methodologically simple: it requires extensive analysis of errors and comparable or similar constructions in the native and target language. However, we believe that this approach will allow us to build a detailed and comprehensive repertoire of error types and to build a library of error “models” (effectively represented by strings of morphological tags such as *eto*+ADV+INF). These models will be subsequently incorporated into a tagging software used to automatically detect and annotate errors in constructions in non-standard varieties of Russian.

4 Conclusions

The paper illustrates the general approach to the identification, categorization and explanation of errors in learner Russian. Although this approach comes with a list of challenges and limitations, we believe that it will not only significantly improve the Russian Learner Corpus but will provide a new model for error-annotation for other corpora “with noise in the signal”. Moreover, we can conduct comparative studies in Russian SLA and pedagogy.

References

Alsufieva, A., Kisselev, O. and Freels, S. 2012. “Results 2012: Using Flagship Data to Develop a Russian Learner Corpus of Academic Writing”. Russian Language Journal, 62: 79-105.

Bonch-Osmolovskaya, A. 2006. Dativnyj subject v russkom yazyke: korpusnoe issledovanie. Unpublished PhD thesis, Moscow State University.

Fillmore, Ch. J., Paul Kay and Mary Catherine O’Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of *Let Alone*. Language 64 3. 501-538.

Goldberg, A. 2006. Constructions at Work. Oxford: Oxford University Press.

Granger, S. 1998. Learner English on Computer. Addison Wesley Longman, London and New York.

Ladygina, A. 2014. Russkie heritazhnye konstruktsii: korpusnoe issledovanie. Unpublished MA thesis, Moscow State University.

Loshkareva, D. 2015. Unpublished MA thesis. National research university “Higher School of Economics”.

Rakhilina, E. V. (Eds.) 2010. Construction linguistics. Russian Academy of Sciences V. V. Vinogradov. Russian Language Institute. Azbukovnik. Mocow. (In Russian)