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

Evgeniya Smolovskaya
National Research University HSE
esmolovskaya@hse.ru

Evgeniy Mescheryakova
National Research University HSE
eimescheryakova@hse.ru

Olesya Kisselev
Pennsylvania State University
ovk103@psu.edu

Ekaterina Rakhilina
National Research University HSE
erakhilina@hse.ru

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. A new linguistic resource, known as Russian Learner Corpus (RLC, http://web-corpora.net/heritage_corpus), is now available for researchers of L2 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). The corpus includes comparable sub-corpora 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. The new corpus provides a unique opportunity to conduct comparative studies in Russian SLA and pedagogy, as well as methodological studies that have relevance for learner corpora annotation, analysis and management.

2 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 main discussions vis a vis systematic errors in learner language are currently focusing on the following two issues: 1. methodological issues such as creating annotator-friendly tagging systems and automated and semi-automated methods of error identification in non-standard texts, and 2. theoretical issues of error identification, categorization and explanation of error source. These two lines of work are not entirely independent of each other; in fact, they feed into one another, ideally, resulting into creation of a unified, automated, and comprehensive error tagging system. Error analysis of the texts in the Russian Learner Corpus has been thus far attempted from these two perspectives. Klyachko et al. (2013) tested a protocol for automated error identification, which consisted of comparison of lists of bi- and tri-grams found in the learner corpus to the lists of bi- and tri-grams found in a native corpus. This approach was found to be fairly successful in identification of such errors as noun-adjective agreement and prepositional and verbal government. However, it comes with certain limitations: for instance, it provided far less accurate results for discontiguous structures compared to contiguous strings (possibly due to the size and characteristics of the baseline corpus) and, more importantly, left a large repertoire of non-grammatical structures out of its scope.

Another approach, discussed in this paper, begins with manual annotation of a sample of learner texts. The annotators first read and tag deviant forms using a tagging software developed for the project (see the illustration of the program interface below, Figure 1). Importantly, the error tags include the information about the source of an error (calque, semantic extension, etc.), in addition to the information about the structural property of an error (e.g. lexical, aspectual, morphological).

Those erroneous structures that reach a frequency threshold that reliably points to a systematic rather than a random nature of these errors are then examined and grouped according to structural and functional properties. To illustrate how this approach works we refer to examples below:

(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)

(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)

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.

Example 2 (ETO+AVD+INF) 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?

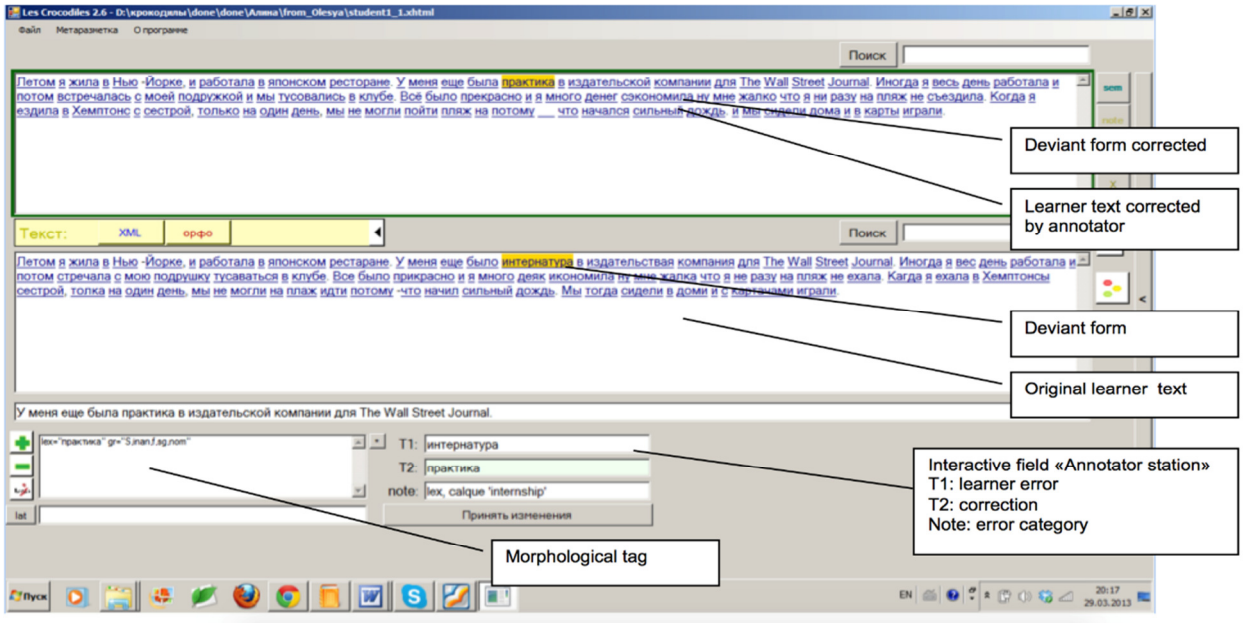
However, when it comes to the examination of oral sub-corpus of the RNC, we find the construction

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.

3 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".

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.

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.

Klyachko, E., Arkchangelskiy, T., Kisselev, O. and Rakhilina. 2013 Automatic error detection in Russian learner language. Conference presentation, CL2013

Discourse and politics in Britain: politicians and the media on Europe

Denise Milizia

University of Bari "Aldo Moro"

denise.milizia@uniba.it

This research is part of a larger project that investigates the sentiment of the UK towards the European Union, the British "à la carte" attitude to the EU, this cherry-picking attitude, as it has been called, which sees Britain opting in, opting out, in many ways half in, half out (Musolff 2004). It cannot be denied that Britain has always been an awkward partner in EU affairs (George 1994), agreeing to some policy areas, disagreeing to some other European policies, for the sake of what has now become the signature of this government: 'in the national interest', 'in Britain's national interest' (Milizia 2014a).

This investigation is based on two political corpora, a spoken corpus and a written corpus. The spoken corpus includes all the speeches of the Labour government from 1997 to 2007, led by Tony Blair, and from 2007 to 2010, led by Gordon Brown; it also includes all the speeches of the coalition government formed in 2010, in which Conservative Prime Minister David Cameron and Liberal Democrat Deputy Prime Minister Nick Clegg were more often than not at odds over the position that the UK will have to take in the near future; it also includes some speeches of the current government, the Conservative government led by David Cameron, who is back in Downing Street after winning the general election of May 2015. Furthermore, the corpus includes some speeches delivered by Nigel Farage, former leader of UKIP (United Kingdom Independence Party), who wants the UK "unequivocally out of Europe", promising that "an exit is imminent" (Milizia 2014c), and some speeches by Ed Miliband, former Labour leader who, in the 2015 Manifesto, maintained that David Cameron "is sleepwalking Britain towards exit from the European Union", and that "Britain will be better off remaining at the heart of a reformed EU".

At the time of writing the spoken corpus totals slightly more than 5 million words.

The written corpus relies on articles from *The Economist*. The data selected comes from the section World Politics, Europe, and at the time of writing it counts 2 million words.

The purpose of the present investigation is to analyse and compare how British politicians and this élite magazine mediate specialized knowledge, European political knowledge in the case in point, how they disseminate knowledge and how they