Russian Language Journal

Volume 62 | Issue 1 Article 6

2012

Results 2012: Using Flagship Data to Develop a Russian Learner Corpus of Academic Writing

Anna A. Alsufieva

Olesya V. Kisselev

Sandra G. Freels

Results 2012: Using Flagship Data to Develop a Russian Learner Corpus of Academic Writing

Anna A. Alsufieva (Yatsenko)
Olesya V. Kisselev
Sandra G. Freels

This paper presents a project developed at the Russian Flagship Center at Portland State University, the pilot Russian Learner Corpus of Academic Writing (piRULEC). PiRULEC is the first of its kind Russian learner corpus that contains academic texts written on a variety of topics produced by advanced learners of Russian from a variety of linguistic backgrounds (heritage speakers of Russian and mainstream American students).

We begin the article with a short introduction to the field of corpus linguistics followed by a closer look at corpus resources available for the Russian language. We then focus on learner corpora research in particular and offer a discussion on the advantages of using learner corpora in the study of language acquisition of Russian as a Foreign Language (RFL). Using the example of piRULEC, we examine possible applications of a developmental Russian learner corpus and provide examples from piRULEC.

Corpus Linguistics and Corpus-informed Language Teaching and Learning

Corpus linguistics has gained a sure footing in linguistic research in the past two decades as computer-aided analyses of collections of authentic texts, known as language corpora, brought about new insights into the nature of language. Language corpora are not simply large; these databases are meant to be principled representative collections of authentic (i.e., naturally-occurring) language. Since corpora comprise a variety of texts, each text comes with meta tags that supply information about such parameters of a text as author, author's gender, genre, time of text creation, length of text, and/or other characteristics depending on the design criteria put forth by the creators of a corpus. Many sophisticated corpora contain elaborate systems of grammatical annotation, which may

provide morphological, semantic, syntactic and/or discoursal information. Meta tags and grammatical tags allow users and researchers to go beyond word searches (not to say that word searches are limiting; many interesting studies can be conducted on un-annotated or raw corpora). Yet, annotation allows for more finely customized searches: for instance, one may choose to search only texts created by female authors of a particular time period in a particular genre or to investigate a grammatical feature such as the use of participles across different genres.

Corpus studies require the use of special software. Even when a corpus itself can be stored in a rather simple database, its analysis will most usually be conducted with the help of a text retrieval program such as WordSmith Tools (Scott, 2010) or MonoConc (Barlow, 2003), although many corpora are nowadays available online and come with built-in text analyzing software. These programs provide general statistics on the texts (such as numbers of words and word tokens, and numbers of sentences and paragraphs), analyze corpora according to the parameters set by the researcher, retrieve the search items (words, phrases or grammatical constructions), create concordance lines and collocation lines, supply information on search item frequency and the like. Using a text-retrieval program, one can sort and compare information obtained from these analyses.

Such flexibility of data analysis done on vast volumes of data in a relatively quick manner turned corpora into unique platforms for theoretical and applied language studies. Many national languages are now represented by large national corpora (British National Corpus, The National Corpus of Polish, Russian National Corpus, etc.) and multiple types of specialized corpora.

Corpus approach to investigation of language has had a significant impact on the field of language teaching (Conrad, 200; McCarthy, et. al., 2005), offering educators new teaching techniques and materials based on authentic language. Among the numerous examples of corpus-based pedagogical resources for English are new grammars such as The Longman Grammar of Spoken and Written English (Biber, et. al., 2002) and Real Grammar (Conrad, 2009); pedagogical materials developed for English as a Second Language (ESL) such as The Academic Word List (Coxhead, 2000), corpus-based textbooks such as Focus on Vocabulary (Schmitt & Schmitt, 2005) and the Touchstone ESL series (McCarthy, et. al., 2005), and many more. Many English-language

corpora are also available for free public use and can thus become a resource for language learners with or without guidance from teachers.

Although the majority of available corpus resources today represent the English language, other language corpora are developing rapidly. The flagship project in the Russian corpus linguistics today is the Russian National Corpus (RNC, http://www.ruscorpora.ru/index.html), which sets out to represent the contemporary Russian language across genres and styles. The corpus is composed of multiple sub-corpora such as literary, legal, dialectal, technical and other corpora. RNC is continuously updated and boasts one of the most sophisticated grammatical tagging systems providing detailed morphological and semantic information for each word in the corpus (a large sub-corpus in RNC also contains syntactic annotation). This diversity, complexness and sophistication of annotation of the data collected in RNC provides an enormous platform for theoretical study of the Russian language and holds the promise of bringing new corpus-based materials such as contemporary dictionaries (for example, dictionaries currently available at http://dict.ruslang.ru/) and new corpus-informed grammars of the Russian language (see project Русская Грамматика at http://rusgram.ru/, currently under development).

Fully realizing the potential impact of corpus linguistics on pedagogical practice, the team of the RNC has developed a unique sub-corpus specifically for pedagogical purposes (Обучающий корпус русского языка, http://www.ruscorpora.ru/search-school.html). This "educational" sub-corpus contains texts that are correlated to the Russian educational programs, and the texts are annotated to reflect the demands of Russian language school curriculum. The pedagogical applications of this sub-corpus as well as corpus data in general are reviewed in detail in Dobrushina (2005, 2009) and Savchuk & Sichinava (2009). All articles describe an array of corpus-based tasks appropriate for Russian language courses taught at Russian schools and universities: examination of lexemes, analyses of grammatical forms, identification of text register, and the like.

RNC has also been suggested as a pedagogical tool for teaching advanced levels of Russian as a Foreign Language (RFL) and Russian for professional purposes (Levinzon, 2007), as well as a platform for unguided exploration of linguistic phenomena by advanced learners of RFL (Janda, 2007).

Another Russian language corpus project that holds a promise for RFL classroom practice is HANCO (Хельсинский аннотированный корпус русского языка, http://www.ling.helsinki.fi/projects/hanco/). HANCO represents a collection of texts published in the early 2000s in the Russian language political and social-media magazine Itogi. The corpus is relatively small (100,000 words) but contains exact and detailed grammatical annotation accessible for instructors and students alike. Some ideas for the use of HANCO in teaching RFL are listed in Kopotev (2008).

The development and constant improvement of authentic corpora has undoubtedly changed the study of linguistic phenomena; corpus linguistics research has also revolutionized the related field of language pedagogy. It is possible that this success of corpus linguistics has prompted a new direction in corpus research, namely the study of learner corpora. The next section (and effectively the rest of this article) will be largely devoted to the discussion of learner language corpora and their applications.

Corpus Linguistics and Learner Language

Similarly to a native corpus, a learner corpus is a large, digitized principled collection of texts produced – in the case of learner corpora – by foreign language (FL) or second language (L2) speakers of a given language. It is important to mention that learner corpora are usually smaller than native corpora: the volume is often constrained by data availability, but is often governed by the research design. At the same time, learner corpora often contain a more detailed system of meta tags addressing the fact that language production of learners is influenced by a far larger combination of factors than that of native speakers. The question of grammatical tagging in case of learner corpora is more complicated: most automated taggers have been developed for standard linguistic forms and are most likely inapplicable to learner language with its many "deviations," although many learner corpus projects aim exactly at error detection and tagging, and subsequent categorization of errors.

The learner corpus "revolution" similarly started in the field of English as a Foreign or a Second Language (EFL/ESL), with Sylvain Granger's ground-breaking work on The International Corpus of Learner English (ICLE, 1996). ICLE, which contains 3,640 argumentative essays written by advanced learners of English from 11 different first language backgrounds, was envisioned as a tool to contribute to better

understanding of the universal and language- and group-specific patterns of EFL/ESL acquisition. ICLE spurred an array of studies – mostly in vocabulary, vocabulary frequency and discourse areas – that set the standard in learner corpus research. Soon after, other small and large studies of learner corpora in languages other than English followed, creating a new research agenda and establishing new lines of inquiry and application. A survey of learner corpora reveal various applications of learner corpora, such as identifying general patterns of language acquisition by groups of learners of various first languages; developing linguistic portraits of non-native speakers of various linguistic levels; comparing patterns of errors of heritage and traditional learners of language at different levels; comparing frequencies of cohesion devices in writing of L2 and L1 speakers; comparing usage of collocations in L1 and L2 production; studying the effectiveness of pedagogical intervention; analyzing linguistic progress over time; developing pedagogical materials; and more (see Dagneaux, et. al., 1998; Granger, 1996; Granger, 1999; and Hinkel, 2001, for ESL/EFL; Stritar, 2009, for Slovene as a Foreign Language; Hana, et. al., 2010, for Czech as a Non-Native Language; Tenfjord, 2008, for Norwegian as a Second Language).

The corpus studies of learner Russian are still few (Kopotev & Mustajoki, 2008). Contrastive learner corpus analysis (CLCA) was applied in a research study by Pavlenko and Driagina (2007), who compared oral narratives produced by American speakers of RFL (30 learners), Russian monolinguals and American monolinguals (as base groups). The oral narratives, experimentally collected retellings of a silent film, were transcribed and compared among the three groups with the goal of identifying patterns of emotion talk. The authors compared frequencies and appropriateness of emotion words and lemmas among the three groups and found that unlike Russian monolinguals, who show strong preference for emotion verbs, the American RFL learners prefer adjectival constructions, violate sociolinguistic register, and generally use a smaller register of emotion lemmas.

Hasko (2010) used the CLCA approach in a study of Russian motion verbs. In this study, Russian 30 native speakers of Russian and 30 advanced American RFL speakers were asked to produce spontaneous oral stories based on the picture book Frog, Where Are You? (Mayer, 1969). Once transcribed, the texts were selectively annotated for verbs of motion, and the patterns of grammatical representation of motion were compared in the two sub-corpora. The author concludes that the learners

exhibit a lack of systemic conceptual structuring of unidirectional motion, despite the fact that they are experienced learners of Russian.

These two studies undoubtedly contributed to a better understanding of the lacunas that American RFL speakers might still have, even at advanced levels of linguistic competence. More importantly, these studies introduced an approach new to the field of Russian language acquisition. Both corpora are, however, relatively small and were collected with specific research goals, which limits the use of these data for future research. And, of course, the field of Russian language studies, theoretical and applied, requires a variety and diversity of learner corpora, encompassing different levels, genres, modes, and designs. We believe that proliferation of learner corpora will be advantageous to the field of Russian language studies; learner corpora may be an especially beneficial resource for the programs that teach advanced levels of Russian, such as Flagship programs.

Russian Flagship at Portland State University

The Russian Flagship at Portland State University, like other Russian Flagship programs, prepares students for participation in the Russian Overseas Flagship at St. Petersburg State University. The Portland State program is composed of four levels of study modeled on the university's general education program:

- Level 1 "Globalization" (30 weeks, prerequisite Intermediate-Mid)
- Level 2 "American Studies," "European Studies," Environmental Sustainability" (30 weeks, prerequisite Intermediate-High)
- Level 3 "Russian in the Major" (30 weeks, prerequisite Advanced-Low)
- Level 4 "Effecting Change" (taught in coordination with the Russian Overseas Flagship, prerequisite Advanced)

Note that in order to begin the Flagship Sequence, students need to possess linguistic skills at the Intermediate-Mid level, at least. Those who complete the entire program, including one academic year in the Russian Overseas Flagship, expect to graduate with ACTFL Superior proficiency in Russian.

The four levels of the PSU Russian Flagship are designed to accommodate students from diverse backgrounds ranging from traditional seat learners of Russian to heritage speakers. Each level of study is organized in part around the subject matter and in part around

providing students with the linguistic tools they need to accomplish certain tasks. Level 1 students focus on the formation of paragraphs. Level 2 students work on essays and classroom presentations. Level 3 students learn to conduct and present research, and Level 4 students work collaboratively to create a product, such as a website or a video, that addresses a community need.

Since its inception in 2008, the PSU Russian Flagship has dealt with two related problems: the relative lack of Intermediate High/Advanced instructional materials, especially on such diverse topics, and the relative lack of empirical research on advanced interlanguage, especially among traditional and heritage learners of Russian. The authors hypothesized that a corpus of the Flagship learners' written Russian would form an important tool for eventual research on language acquisition, but also would meet the more immediate needs of facilitating the development of instructional materials and the assessment both of students' progress and of the effectiveness of pedagogical interventions at each level of study.

PiRULEC

In 2008, the faculty of the RFP at PSU began a compilation of a pilot Russian Learner Corpus of Academic Writing (piRULEC). To this end, all written assignments produced by PSU's Russian Flagship students both in class and at home were collected, digitized, assigned codes, and entered into the corpus.

To allow for the complexity and the diversity of the research questions and the assessment tasks we had in mind, the design criteria of piRULEC had to be thoroughly thought through. In general, learner corpora specialists note the increased variability of learner language (compared to native language). This variability of linguistic product is influenced by a wide range of linguistic, situational, and psychological factors, such as language level, time restriction, type of a task, familiarity with topic, etc., and the range and impact of these factors is greater in learner language production than in that of native speakers (Granger, 2004; Tono 2003; Gass & Selinker, 2001). Granger (2004) suggests that the learner corpus variables largely fall into two categories: those pertaining to the learner (language level, language background and the like) and those pertaining to the task (time restriction, type of assignment); within these categories each researcher distinguishes factors that are most relevant for the learners in question and the research questions.

Having carefully considered the various factors that may influence the language of Flagship learners, we distinguished the following variables: language background and language experience of the student, point in time (week and academic year), time limit under which the paper was written, text type and text function, and whether a paper was written individually or in a group. Each variable and the reasoning for its importance for the current project are described below.

Language Background and Language Experience of a Learner

The students whose works are included in piRULEC have a lot in common: all are young adults studying Russian at an American university with the goal of achieving Superior-level proficiency. The majority of students are undergraduates who major in fields other than Russian. All student-authors in the corpus are at least Intermediate-Mid speakers of Russian, with the majority of students being Advanced-Low and Advanced-Mid speakers of Russian as established by unofficial OPI interviews conducted by the faculty. However, the students differ in the kinds of language background and language experience they possess. The most important language-background variables are: current linguistic level, first, second, foreign language(s), language(s) of schooling, age of exit (if a student is a heritage speaker of Russian), visits to Russian-speaking countries and the purpose of the visits, courses taken in Russian, and Russian language use outside of classroom. These factors have shown to have an immediate impact on language attainment.[1]

The information on linguistic background is collected through a comprehensive student survey and is stored in a database in the form of a sociolinguistic passport (see Appendix A) to use in research. Teachers' comments and information on the results of the external tests are also added to the sociolinguistic passport. The actual identity of each learner is carefully protected through assignment of pseudonyms, which correspond to the sociolinguistic passport stored in the database of the corpus.

Time Stamp

Whether a student is only beginning the program or is ready to join the overseas Flagship site is, naturally, an important variable in the general level of linguistic performance. To allow for very close tracking of linguistic development, we record not only the academic year but also

[1] For a more detailed discussion of language background as a variable, see Tono (2003).

the number of the week in the school year (academic weeks run 1 through 33). The name of the course will also provide an idea of how close the student is to graduation from the RFP, in addition to providing a general topic on which the texts are written.

Time Limit

Time limit is considered to be one of the most important variables influencing the accuracy and complexity of writing (Ellis, 2002). The RFP students complete short writing assignments in class approximately once every week and at least one writing assignment a week at home. Each text in the corpus has an identification of whether the paper was written at home in a non-timed manner or in class in a timed manner.

Text Type

The language data collected for piRULEC represents one register – student academic writing. The data are restricted to academic writing,[2] and the topics discussed in the RFP classes – issues of globalization, historic events, cultural phenomena, or topics in the student's major – go beyond personal experience discourse typical of lower linguistic levels (e.g., My Day or My Friend) and require college-level cognitive and linguistic skills. The types of texts, however, differ depending on the pedagogical goals of particular assignments that may request an essay, an outline (of an essay or oral presentation), or short answer to a question. The topics on which texts are written vary greatly, since they reflect the subject matter of each particular course. As a result we do not treat "topic" as a variable feature.

Text Function

Assignments developed by RFP instructors typically target one or another text function, i.e., a goal of communication, such as describing an object or constructing an argument. Following the ACTFL guidelines, we distinguish the following text functions: definition, paraphrase, summary, narration, description, expository writing, comparison and contrast, cause and effect, supported opinion, argumentation, process analysis, and hypothesis. Blended types are also represented in the corpus, e.g., research papers. It is important to note that the recorded function reflect the one intended by the teacher, not necessarily what the student produced.

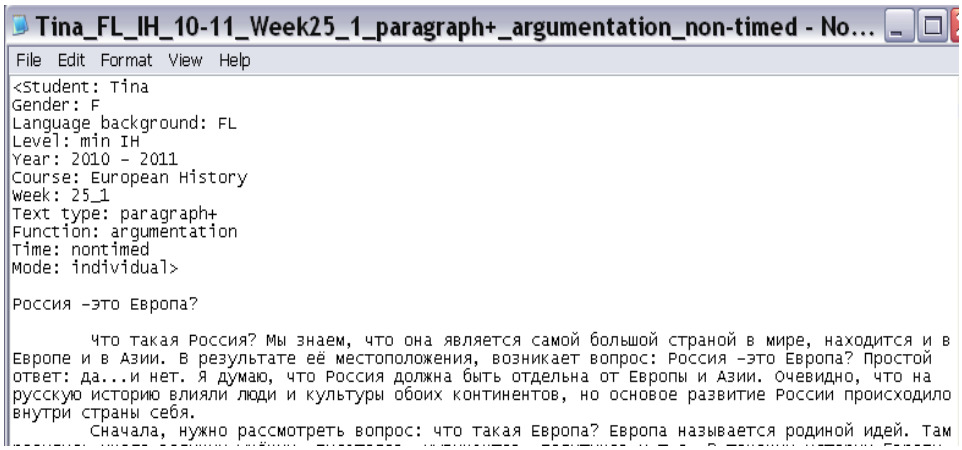
[2] By academic writing, we consider formal papers such as essays, terms papers, book reports, and other types of college-level writing assignments that require college-level cognitive and language skills (Hinkel, 2001).

Mode

Although the greater majority of texts represent the individual effort of each student, a small number of texts was produced by students while working in pairs or small groups.

Most of the characteristics described in the pages above are reflected in the Header Identification Box (Header ID) of each text entered in the corpus as well as in the file name, with the full sociolinguistic information available in the separately stored sociolinguistic passport. The text header ID and the name of piRULEC files are illustrated below (see Illustration 1).

Illustration 1. Text header ID



The information provided in the header ID and in the file names is especially useful when a researcher needs to group texts in piRULEC according to categories (for example, FL learners vs. HL learners, argumentative essays vs. descriptive essays, etc.) for comparative analyses. This information also comes into play when the results of corpus searches are interpreted and discussed. The next section of the paper will illustrate various analyses of piRULEC materials.

Using piRULEC

PiRULEC is at present a relatively small corpus: currently containing 800 texts composed of up to 200,000 words. Texts vary in length from fewer than 40 words to up to 2,000 words. This variability is due to the fact that some files may only include one sentence, while others contain full research papers. The texts are authored by 36 learners; 17 of the 36 are mainstream American learners who have started learning

Russian as adults, and 19 are heritage speakers of Russian, born in a Russian-speaking country and brought to the U.S. as children. The relatively small number of authors in the corpus may preclude the researchers from drawing a generalized conclusion about an "average" advanced learner of RFL. Nevertheless, the creation of piRULEC is an important first step in the study of advanced learners through the use of corpus linguistics methods. Moreover, the relatively large number of works representing each learner may become an advantage for longitudinal studies, ethnographic studies, or studies that require close tracking of interlanguage development.

Currently, PiRULEC is a non-annotated or raw corpus, since tagging software is not readily available for the Russian language. Yet, practical applications and the possibility of theoretical investigations even of this small un-annotated corpus are very broad, spanning studies of vocabulary, grammar, and syntax for a variety of purposes, from theoretical descriptions to the development of pedagogical materials. We distinguish four major applications of piRULEC. First of all, we see piRULEC as a tool that may help uncover universal or group-specific patterns of Russian language acquisition and build profiles of various groups of RFL learners. Secondly, we use piRULEC in assessment of students' linguistic progress and assessment of pedagogical techniques. Thirdly, we view piRULEC as a tool for lingua-pedagogical investigations, and, finally, as a resource for language instructors. These four applications of piRULEC are inherently interdependent and are listed separately, primarily for the purposes of presentation, which follows in the sections below.

Building Linguistic Profiles of Learner Groups and Individual Learners of Russian

One of the most important applications of learner corpora in general and piRULEC in particular is the possibility of creating comprehensive linguistic profiles of various groups of learners. Uncovering patterns of language usage by students of different levels, different language learning histories, different ages, and so on, will shed light on the processes of second language acquisition and the nature of language in general.

Since piRULEC includes texts created by heritage and non-heritage learners of advanced levels of Russian, it can conceivably contribute to better understanding of these two groups, their similarities and differences of acquisition patterns found in the lexicon, grammar,

syntax or discourse. We began the comparison of the two student groups with one of the most compelling topics in comparing heritage and traditional learners of Russian: the topic of usage of phraseologically-bound words (e.g., внимание, проблема, мнение, etc.). To establish whether the heritage and mainstream learners in the RFP differ in their abilities to use such words in native-like collocations, we ran a search for the word "внимание" in two sub-corpora, Heritage Learners and Non-Heritage Learners, created from the piRULEC texts, and analyzed the concordance lines. This manipulation produced the following results: there are 11 occurrences of the word "внимание" in the Non-Heritage sub-corpus and 43 occurrences in the Heritage sub-corpus. Given that the size of the two sub-corpora is approximately the same, the difference in the number of attempts to use the abstract noun is significant. A closer look at the samples reveals more qualitative differences: 10 of 11 instances of "внимание" in the Non-Heritage appear in a native-like collocation (e.g., обращать/обратить внимание, принимать во внимание, в центре внимания), and only one was erroneous (…[автор] негативно относится к внимании на идею, что всё можно продать и купить). The percentage of errors in the Heritage sub-corpus is higher and includes different types of lexico-grammatical deviations: the use of wrong verb (e.g., ...высококачественные рекламы привлекут широкое демографическое поле и вызовут внимание многих людей) and wrong case government (e.g., Другое, чему важно обратить внимание, это отличающие черты стран…). At the same time, however, the heritage learners employ a greater diversity of native-like collocations: in addition to "обращать/обратить внимание" they have used "уделять внимание," "привлекать внимание," and "принять во внимание." This short analysis raises an array of interesting questions: Does this pattern hold with other abstract nouns? If so, does it mean that formal instruction results in more accurate but more restricted "collocating" of abstract nouns, as we saw in the case of non-heritage learners?

Obviously, an analysis like the one above is only one step in the direction of compiling comprehensive portfolios of various groups of learners, and yet, we would like to argue that it is an important step that sets new protocols for the study of learner of RFL and opens new opportunities for such studies. In addition to providing the insight into the nature of language acquisition and particularly of learners' interlanguage, the learner corpora, especially such small developmental corpora like piRULEC, can have an immediate impact on the pedagogical

practices. The further sections will provide an overview of pedagogical applications of learner corpora based on the example of piRULEC.

Learner Corpora as a Platform for Lingua-Pedagogical Investigations

By lingua-pedagogy, we understand an approach to the study of the learner language that is directed at the understanding of formal features of interlanguage (gaps and strengths) of a particular group of learners with the goal of adjusting pedagogical practices employed with this group of learners. Thus, piRULEC offers the instructors of the RFP at PSU a way of validating (or disproving) certain ideas about the linguistic difficulties of the RFP students. For example, after noticing some erroneous usages of the preposition "через" in our students' work, we conducted a short study of the said preposition to establish (1) whether the difficulty with preposition "через" was individual or typical of all learners, (2) if any particular meaning of "через" was more problematic than the other meanings, and (3) if there was a difference in the usage of "через" between the heritage and non-heritage students. If the difficulty was to be found general, we would then develop a language activity to address the gap.

Having analyzed the concordance lines obtained from a search of "через" in the two sub-corpora, Heritage and Non-Heritage, we established that both groups used the preposition with approximately the same frequency: 44 instances in the non-heritage corpus and 58 instances in the heritage. The following three meanings of the preposition were present in the speech of the learners: temporal (e.g., через несколько минут я была на автобусной остановке), transitive (e.g., слово пришло в русский язык через польский), and mediative (e.g., Я думаю, что через СМИ люди стали узнавать больше о ситуациях в мире); all uses are typical of the native Russian speech (Zolotova, 1988).

However, if the sentences with "через" in temporal and transitive meaning were correct, the use of this preposition in the meaning of abstract medium was not native-like (e.g., через анализ советской музыки Сталинского периода мы увидели…, Америка показала свою прогрессивность через избрание афро-американца в Белый дом) (see table 1).

Table 1. Preposition "через" in Non-Heritage and Heritage sub-corpora of piRULEC

	Non-heritage learners	Heritage learners
Temporal meaning	10 correct occurrences 0 incorrect occurrences	3 correct occurrences 0 incorrect occurrences
Transitive meaning	3 correct occurrences 0 incorrect occurrences	7 correct occurrences 0 incorrect occurrences
Mediative meaning	31 correct occurrences 25 incorrect occurrences	48 correct occurrences 30 incorrect occurrences

In the "educational" sub-corpus of the Russian National Corpus, the vast majority of the contexts with "через" were found to be temporal (через месяц, через некоторое время) followed by a smaller number of the contexts, in which the preposition is used in the transitive meaning (через костер, через реку, через окно); the contexts in which "через" is used in the mediative meaning are relatively few in the RNC. The tendency was inverted in the speech of our students. It appeared that the students were transferring an English construction with preposition "through" in the mediative meaning (e.g., through songs and poems one learns about culture) instead of expressing the mediative meaning with the help of Russian constructions such as "с помощью," "используя," or other constructions. This clearly identified the need for pedagogical intervention.

In order to address this gap in the students' written interlanguage, we developed a sequence of exercises using the same concordance lines that we used in the study (see below).

> 1. Look at the sentences with preposition "через" and group them in groups that make sense to you. How many groups do you have? Which categories did you use for grouping the sentences?
>
> 1) Через несколько минут я была на автобусной остановке.
>
> 2) Договор предлагает международное сотрудничество через культуру.
>
> 3) Англия не разрешила российскому флоту пройти через Суэцкий канал.

4) Через СМИ люди стали больше узнавать о происходящем в мире.[3]
5) etc.

2. Compare the sentences in each pair. Underline the phrase that was replaced in sentence 1 of the pair and the phrase that replaced it in sentence 2 of the pair.
1) Через изучение культуры мы учимся уважать других == Изучая
культуру, мы учимся уважать других.
2) Невозможно относится позитивно к Америке через роман "Великий Гетсби" == После (про)чтения романа "Великий Гетсби" невозможно относиться к Америке позитивно.
3) Америка показала свою прогрессивность через избрание афро-американца в Белый дом == Избрав президентом афро-американца, Америка показала свою прогрессивность.
4) etc.

3. Read the sentences and identify the meaning of preposition "через." Paraphrase, if possible.
1) Он вернулся в Магадан через 20 лет.
2) Импортную одежду покупали через знакомых.
3) Через СМИ люди стали больше узнавать о происходя-
щем в мире.
4) Эпишура -- это маленький рачок, который через свой фильтр очищает воду.
5) etc.

As one can see, the learner corpus approach not only addresses global issues of second language acquisition, it also allows the instructor to uncover issues relevant to a particular group of students and address them in real-time using the "real" authentic language in a time-efficient manner.

[3] All samples are retrieved from piRULEC with spelling mistakes corrected.

Learner Corpus as a Source of Pedagogical Material

The sequence of activities presented in the previous section was a result of a linguistic study. However, a teacher often does not require a study to know that his or her students need practice with a particular structure or concept. As long as the area of difficulty is recognized, the corpus can become a source of development of various tasks and exercises. These include lexical and grammatical exercises, as well as those that target spelling, morphology, punctuation and syntax, discourse, and register variation. A teacher might choose to create a fill-in-the-gap type activity or have learners conduct their own guided searchers in corpus. Below, we suggest a few activities that aim at three different areas of difficulty: punctuation mistakes motivated by intonation, punctuation mark with conjunction "который," and choice between adjectives "русский" and "российский."

Heritage Russian learners are known to have the tendency to put commas according to the intonation patterns that they would use pronouncing the sentences aloud (Zemskaia, 2001). At the same time, the traditional students as well as heritage students schooled in English have a tendency to transfer English-language punctuation rule into Russian marking introductory phrases with a comma. To address this problem, a teacher can quickly assemble an exercise by pulling sentences containing this type of mistake from the corpus and making an activity. For example:

> Task: Find the mistakes in punctuations and correct them. Explain your choice.
>
> 1) Очень часто при открытии малого бизнеса, компания берет коммерческий кредит, с минимальной процентной ставкой для покупки товара.
>
> 2) В бухгалтерских счетах, кредиты находятся с правой стороны и эта кредитовая сторона счета содержит доходы.
>
> 3) За использование мобильного телефона в Европе, клиенты компании АТТ должны платить тариф плюс 5 коп. за минуту разговора.
>
> 4) Когда Пётр Первый путешествовал по Европе, европейские идеи не были важны для России. Однако, Пётр I стремился изменить систему российского государства.
>
> 5) В 18 веке в России, стали появляться учебные заведения для девушек.
>
> 6) etc.

To address another type of typical mistake in punctuation with conjunction "который," students may be given the following exercise:

Task: Find the mistakes in punctuations and correct them. Explain your choice.

1) В данный момент налог корпораций доход которых 250,000 или выше составляет 6.6%.
2) Философия художника стала основой его социальных действий в число которых входит и Пакт о защите международных культурных ценностей.
3) Нашу жизнь заполняют различные товары и услуги рекламы которых убеждают нас выбрать именно этот продукт, а не какой-то другой.
4) Карьерист – такой человек который идет по головам других.
5) etc.

To raise awareness of the choice between adjectives "русский" and "российский" and the typical collocations of these adjectives in the Russian language, one may create the following exercises based on piRULEC concordance lines:

a) Task: Insert the missing adjectives "российский" or "русский." If both adjectives are applicable, explain the differences in the meaning of phrase/context.

1) ..... эмигранты каждой волны отличаются друг от друга
2) В эту организацию входят известные .....ученые
3) Жизнь .....крестьян после отмены крепостного права
4) Огромные размеры .....территории
5) После развала СССР ……..культура резко изменилась
6) etc.

b) Task: Insert the missing nouns from the given list (молодежь, правительство, человек, интересы, земля etc.)

1) В наше время происходят изменения в картине мира русского .....
2) Николай II стремился защитить российские ..... на Дальнем Востоке.

3) В XVI в. российское ..... начало формировать систему образования
4) В 13 веке татары начали занимать русскую .....
5) В радиопередаче обсуждались проблемы российской .....
6) etc.

The nature of pedagogical activities depends on many factors, from the goal of the activity to the instructor's pedagogical style to the availability of instructional resources. We suggest that piRULEC as well as other learner corpora may be successfully used as an instructional resource providing ideas and material for exercises.

PiRULEC as an Assessment Tool

As mentioned previously, learner corpora may become an advantageous tool for assessment of students' linguistic progress. It does not require additional tests and can assess the current state of the language as well as linguistic progress of an individual student or a whole group of learners; the corpus-based assessment may be more comprehensive than a test (one can assess different language categories in different contexts) and more flexible (through the use of different base-lines such as native speech, the performance of the cohort, and the performance in previous terms).

Corpus approaches to assessment of writing in Russian, however, are not thoroughly developed; the example that is reviewed on the pages below has as much to do with establishing a protocol for corpus-based assessment of students' writing as with the actual assessment. Since the study was not designed to provide a comprehensive analysis, we focused on one formal feature of advanced writing; namely, complex sentences. Since piRULEC does not have syntactic annotation, we analyzed complex sentences through the use of subordinating conjunctions. For the study, we chose a cohort of eight students (heritage and traditional), who have gone through two years of instruction in the RFP, and created two sub-corpora: sub-corpus 1, which included papers that students wrote at the end of their first year at RFP, and sub-corpus 2, which contained texts from the end of the second year. Perusing the word lists created off the two sub-corpora, we retrieved all subordinating conjunctions, conducted separate searches for each conjunction, and analyzed the concordance lines.

We found that the overall quantity of subordinating conjunctions did not increase over the course of the year, despite the fact that the number of sentences in the sub-corpus 2 grew by 36 percent. The usage of conjunctions, however, has changed qualitatively. For example, there was a decrease of the most frequent conjunction "что" and an increase in the number and variety of causal conjunctions such as "потому что," "поэтому," "поскольку" and "потому." Within the class of subordinating subjunctions, the instances of incorrect usage decreased. The most dramatic move towards accuracy was observed in the case of "однако": a misplaced comma appears in 91 percent of all instances of usage at the end of year 1; by the end of the second year, the comma is incorrectly used in only 40 percent of cases. Additionally, by the end of year 2, the students began to use compound conjunctions and use conjunctions within complex syntactic structures such as "как ..., так и," "так же ..., как и," "такие как…, а также," and other similar constructions (e.g., Знак триединства можно найти как на храмах Западной Европы, так и на восточных изображениях Будды.) Students also began to use multi-lexeme conjunctions for added emphasis (e.g., "и потому," "но и," "но при этом"), moving towards stylistic variation through the use of grammatical structures. This corpus study showed a tangible and measurable progress in students'syntax and established a working protocol for assessment of formal features of complex writing.

Just as one can track the "combined" progress of a whole group, it is possible to assess progress (or lack thereof) in the language performance of an individual student. The example that follows shows an attempt to assess the progress in lexical choices of one RFP student. We selected a number of phraseologically-bound words that were used incorrectly in the early works (13 texts) of the learner and then ran the searchers for these words in the later assignments (nine texts). We found that the usage of the verb "состоять," for example, was incorrect in the early works, where the student used this word in place of other phraseologically-bound verbs (*состоит частью vs. является частью) or similar-sounding words (*состоит работать vs. *предстоит работать). After one academic year, the student uses the phrases correctly and appropriately. See the concordance lines below:

Concordance lines from Fall 2009 - Winter 2010

1. Я думаю, что Россия состоит частью Европы, но при этом я имею введу...

2. Иногда, такой выбор не состоит возможным, из-за этого, многие бывают захватанными идеями
3. ... что американцам состоит еще работать над ликвидацией сегрегации населения.
4. ... непонимание авантгардных произведений искусства состоит еще и в том, что зрители не знают о его происхождении.

Concordance lines from Spring 2011

1. ...и ее [науки] главная цель состоит в изучении непознаного
2. … подходящее название текста состоит из комбинации двух предлагаемых названий

We believe that corpus approach to assessment provides an important angle in writing assessment. In addition to using the criteria of general impression of writing, we can look for particulars, for formal features of writing such as separate lexemes, collocations, cohesive devices, punctuation and more. Corpus approach may provide a tangible, measurable result, and not just a sense of the overall impression of students' writing.

Conclusion

The applications of learner corpora and the advantages of the learner corpora to the field of SLA and pedagogy are vastly broader than the ones described here. The scope and the nature of a corpus-based investigation is only somewhat limited by technology and depends on the needs and, often, on the imagination of the researcher. Despite its limitations of size and representativeness and the current lack of annotation, piRULEC may provide a rich resource for theoretical studies and practical work in the field of Russian language acquisition. We plan to have the finished version of piRULEC by summer 2013, and to share the corpus with colleagues and students of Russian. All materials will be stored on CD-ROMs in text file format and accompanied by a manual describing the corpus design and offering ways of corpus utilization.

More importantly, we hope that the arguments that we advanced in this article and the examples of what one can do with a learner corpus of Russian will encourage language researchers as well as Russian language teachers to investigate corpus approaches to the study of

language and the learners and, through this, enrich the field of theoretical study of Russian as well as teaching Russian as a Foreign Language.

Appendix A

Student Sociolinguistic Passport Sample 1

Name : Daniel
Gender: M F
Age: 30
Major: Russian

I. Language background. Please provide details if you can.

–What was the first language or languages you learned to speak? English

–What language(s) your family spoke when you were growing up? English

–What was the language or languages you were schooled in? English

–Did you learn a foreign language(s) in school or college? 1 term of Spanish when 16. I dropped the class or completed at a poor level.

II. Russian language experience.
–Age when you left Russia or another country where Russian is a primary language (if born in the US, please put 0) 0

–List all Russian language courses taken up to this date (please provide name and year): Russian 101, 102, 103 – 1998-199; Russian 201, 202, 203 – 1999-2000; Russian 301, 2007, Russian 411, 412, 416 2010-2011; currently in Russian 416 (Russian Flagship First level)

–As a child, did you attend any Russian language classes? (Russian pre-school, grade school; or if born and raised in the US: home classes/church/Sunday school/private lessons/etc.) Please specify: no

–How many times did you visit a country where Russian is spoken by the majority of speakers? How long was each visit? When was it? What was the goal of your visit (studies, work, tourism, family visit). 10+; 2 months in 1999 for study and tourism w/WSU Dr. Bill Richrdson and Pskov Volny University; 200 – moved to Russia, study, then marriage, then work, then left in 12/2011; several visits to Ukraine and Estonia.

–When was the last time when you visited a country where Russian is spoken by the majority of speakers? How long was this visit? 2006, Voronezh, 2 weeks.

–Where and how often do you use your Russian? (Circle all that apply and provide an estimation of time you are engaged in this activity each week)
Classes: 2+hours
Home: 0 hours
Place of worship: 0 hours
Friends: ~ 1 hours
Russian Immersion Dorm: 0 hours
Tutor: 1–2 hours
Extra-curricular activities: 0 hours
Other (please specify): N/A

Student Sociolinguistic Passport Sample 2

Name:
Gender: M F
Age: 20
Major: International Studies

I. Language background. Please provide details if you can.
–What was the first language or languages you learned to speak? Russian

–What language(s) your family spoke when you were growing up? Russian

–What was the language or languages you were schooled in? Russian, then English.

–Did you learn a foreign language(s) in school or college? Russian, Italian, English.

II. Russian language experience.
–Age when you left Russia or another country where Russian is a primary language (if born in the US, please put 0) 11 yrs old.

–List all Russian language courses taken up to this date (please provide name and year): Russian 416, three-course sequence On Democracy (2009-2010, Russian Flagship First level), Russian Grammar (Spring 2010), Rus 421 Contemporary Russia (Summer 2010), Rus 416 American Studies and Rus 416 Environmental Sustainability (Russian Flagship Second level) currently in Rus 416 European History (Russian Flagship Second level), also currently in Rus 416 three-term sequence Russian in the Major (Russian Flagship Third level).

–As a child, did you attend any Russian language classes? (Russian pre-school, grade school; or if born and raised in the US: home classes/church/Sunday school/private lessons/etc.) Please specify: went to Russian school in Estonia (grades 1–5).

–How many times did you visit a country where Russian is spoken by the majority of speakers? How long was each visit? When was it? What was the goal of your visit (studies, work, tourism, family visit): 2: 2 weeks in Estonia, family visit, 2 months in Estonia, family visit.

–When was the last time when you visited a country where Russian is spoken by the majority of speakers? How long was this visit? 2007, 2 weeks.

–Where and how often do you use your Russian? (Circle all that apply and provide an estimation of time you are engaged in this activity each week)

Classes: 4+ hours
Home: everyday hours
Place of worship: 0 hours
Friends: 2 hours
Russian Immersion Dorm: 0 hours

Tutor: 1–2 hours
Extra-curricular activities: 1 hours
Other (please specify): N/A

References:

American Council on the Teaching of Foreign Languages, ACTFL, http://www.actfl.org/.

Barlow, Michael. Concordancing and Corpus Analysis Using MP 2.2. Houston: Athelstan, 2003.

Conrad, Susan. "Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century?" TESOL Quarterly, 34, no. 3 (2000): 548-560.

Dagneaux, Estelle, Sharon Denness, and Sylviane Granger. "Computer-Aided Error Analysis." System, 26 (1998): 163-174.

Dobrushina N.R. "Kak ispol'zovat' Natsional'nyy korpus russkogo yazyka v obrazovanii?" In Natsional'nyy korpus russkogo yazyka: 2003 – 2005. Rezul'taty i perspektivy, foreword by V.A. Plungyan, 308-329. Moskva: Indrik, 2005.

Dobrushina N.R. "Korpusnye metodiki obucheniya russkomu yazyku." In Natsional'nyy korpus russkogo yazyka: 2006 – 2008. Novye rezul'taty i perspektivy, edited by V.A. Plungyan, 335-352. SPb.: Nestor-Istoriya, 2009.

Ellis, Rod. "The Methodology of Task-Based Teaching." Asian EFL Journal 8, no 3 (2006): 19-45.

Granger, Sylviane. "From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora." In Languages in contrast, edited by Karin Aijmer, Bengt Altenberg, and Mats Johansson, 37-51. Lund, Sweden: Lund University Press, 1996.

Granger, Sylviane. "Use of Tenses by Advanced EFL Learners: Evidence from an Error-tagged Computer Corpus." In Out of Corpora. Studies in Honour of Stig Johansson, edited by Hilde Hasselgard and Signe Oksefjell, 191-202. Amsterdam: Rodopi, 1999.

Granger, Sylviane. "The International Corpus of Learner English: A new resource for Foreign Language learning and teaching and Second Language Acquisition research." TESOL Quarterly, 37 (2003): 538-546.

Granger, Sylviane. "Computer Learner Corpus Research: Current Status and Future Prospects." In Applied Corpus Linguistics: a

Multidimensional Perspective, edited by Ulla Connor and Thomas Upton. 123-145. Amsterdam & Atlanta: Rodopi, 2004.

Hana, Jirka, Svatava Škodova, Alexandr Rosen, and Barbora Štindlova. "Error-tagged Learner Corpus of Czech." In Proceedings of the Fourth Linguistics Annotation Workshop, ACL 2010, edited by Nianwen Xue and Massimo Poesio, 11-19. Uppsala, Sweden, 2010.

Hasko, Victoria. "Unidirectional and Multidirectional Motion Events in the Speech of L12 Learners of Russian." In Linguistic Relativity in L2 Acquisition: Evidence of L1 Thinking for Speaking, edited by Han ZhaoHong and Teresa Cadierno, 34-58. Clevedon: Multilingual Matters, 2010.

Hasko, Victoria. "Motion Domains in Russian and English: Corpus-based Analysis." In New Approaches to Slavic Verbs of Motion, edited by Victoria Hasko and Renee Perelmutter, 197-224. Amsterdam/ Philadelphia: John Benjamins, 2010.

Hinkel, Eli. (2001). "Matters of Cohesion in L2 Academic Texts." Applied Language Learning, 12. no. 2 (2001): 111-132.

Janda, Laura. "Studenty – pol'zovateli Natsional'nogo korpusa russkogo yazyka." In Natsional'nyj korpus russkogo yazyka i problemy gumanitarnogo obrazovaniya, edited by N.R. Dobrushina, 60-73. Moskva: Teis, 2007.

Kopotev M.V. and Arto Mustajoki. "Sovremennaya korpusnaya rusistika" In Instrumentariy rusistiki: korpusnye podkhody, edited by A. Mustajoki, M.V. Kopotev, L.A. Biryulina, E. Yu. Protasova, 7-24. Khel'sinki, 2008.

Leech, Geoffrey. "Teaching and Language Corpora: A Convergence." In Teaching and Language Corpora, edited by Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 1-23. London: Longman, 1997.

Leech, Geoffrey. "Learner corpora: What They Are and What Can Be Done with Them." In Learn English on Computer edited by Sylviane Granger, xiv – xx. London,1998.

Levinzon A.I. "Ispol'zovanie Natsional'nogo korpusa russkogo yazyka v obuchenii russkomu yazyku angloyazychnykh studentov." Russkiy yazyk za rubezhom, 4 (2007): 64 -73.

McCarthy, Michael, Jeanne McCarten, and Helen Sandiford. Touchstone 1-4: From Corpus to Course. Cambridge UK: Cambridge University Press, 2005.

Stritar, Mojca. "Slovene as a Foreign Language: The Pilot Learner Corpus Perspective." Slovenski jezik – Slovene Linguistic Studies 7 (2009): 135-152.

Pavlenko, Aneta and Victoria Driagina. "Russian Emotion Vocabulary in American Learners' Narratives." The Modern Language Journal 91, no 2 (2007), 213-234.

Pravec, Norma. "Survey of Learner Corpora." ICAME Journal: Computers in English Linguistics, 26 (2002). http://icame.uib.no/ij26/.

Savchuk S.O. and D.V. Sichinava. "Obuchayushchiy korpus russkogo yazyka i ego ispol'zovanie v prepodavatel'skoy praktike." In Natsional'nyy korpus russkogo yazyka: 2006 – 2008. Novye rezul'taty i perspektivy, edited by V.A. Plungyan, 317-334. SPb.: Nestor-Istoriya, 2009.

Scott, Mike. WordSmith Tools. Version 5.0 (2010). http://www.lexically.net/wordsmith/.

Tenfjord, Kari, Paul Meurer, and Knut Hofland. "The ASK Corpus – a Language Learner Corpus of Norwegian as a Second Language." Paper presented at The TALC 2004 Conference, Granada Spain, 69 July 2004. 1821-1824, http://gandalf.aksis.uib.no/lrec2006/pdf/573_pdf.pdf.

Tono, Yukoi. "Learner Corpora: Design, Development and Applications." In Proceedings of the Corpus Linguistics 2003 Conference (CL 2003). Technical Papers 16, edited by Dawn Archer, Tony McEnery, Paul Rayson, Andrew Hardie. Lancaster University: University Centre for Computer Corpus Research on Language. (2003): 800-809. http://ucrel.lancs.ac.uk/publications/CL2003/papers/tono.pdf.

Zolotova G.A. Sintaksicheskiy slovar': Repertuar elementarnykh edinits russkogo sintsksisa. Moskva: Nauka, 1988.

Yazyk russkogo zarubezh'ya: Obshchie protsessy i rechevye portrety: Kollektivnaya monografiya. Editor E.A. Zemskaya. Moskva; Vena: Yazyki slavyanskoy kul'tury: Venskyy slavisticheskyy al'manakh, 2001.

Links to corpora (mentioned in the article)

Centre for English Corpus Linguistics: "Learner corpora around the world," last modified July 9 2012, http://www.uclouvain.be/en-cecl-lcWorld.html.

IntelliText: The Intelligent Tools for Creating and Analysing Electronic Text Corpora for Humanities Research, last modified March 2011, accessed August 12, 2012. http://corpus.leeds.ac.uk/it/.

Norwegian as second language corpus: "Norsk andrespråkskorpus (ASK)," last modified November 30 2009, http://ask.uib.no/index.page.

Russian National Corpus, http://www.ruscorpora.ru/index.html.

Russian National Corpus: "Educational" sub-corpus, http://www.ruscorpora.ru/search-school.html.

The British National Corpus, http://www.natcorp.ox.ac.uk/.

The Helsinki Annotated Corpus of Russian Texts (HANCO), http://www.ling.helsinki.fi/projects/hanco/.

The National Corpus of Polish, http://nkjp.pl/.