D.O.I. http://dx.doi.org/10.1590/2176-4573e55747

ARTICLES

Russian Learner *Corpora* Research: State of the Art and Call for Action / *Pesquisa com* corpora *de aprendizes de russo: estado da arte e apelo à ação*

Olesya Kisselev[*]

ABSTRACT
With the increase in availability and user-friendliness of Russian language *corpora* and *corpus*-analytic tools, the field of Russian language education has recently begun to employ *corpus* linguistics as an approach to understanding the dynamic of language development in users of Russian as a second and heritage language. The paper provides a brief overview of the current state of learner *corpus* research as a field and explores the benefits of application of *corpus* linguistics methods and instruments to the study of Russian. The paper reviews pertinent issues in *corpora* design, compilation, and annotation; offers an overview of the existing Russian language *corpora* and reports on the currently available *corpus*-based studies of Russian as a second/heritage language. The paper concludes with a call to the field to explore the benefit of *corpus*-based approaches to the study of Russian.
KEYWORDS: *Corpus* linguistics; Learner *corpus* research; *Corpus*-based research; Russian language *corpora*; Second language acquisition; Heritage language acquisition

RESUMO
Com o aumento da disponibilidade e facilidade de uso de corpora *de língua russa e ferramentas de análise de* corpus*, o campo do ensino da língua russa começou recentemente a empregar a linguística de* corpus *como uma abordagem para entender a dinâmica de desenvolvimento de russo como segunda língua e língua de herança. O artigo fornece uma breve visão geral do estado atual da pesquisa na área de* corpora *de aprendizes e explora os benefícios da aplicação de métodos e instrumentos de linguística de* corpus *para o estudo do russo. O artigo revisa questões pertinentes na área de* design*, compilação e anotação de* corpora*; oferece uma visão geral dos* corpora *de língua russa existentes e descreve os estudos de russo como a segunda língua/língua de herança baseados em análise de* corpus*, atualmente disponíveis. O artigo conclui com um chamado aos especialistas na área para explorar o benefício de abordagens baseadas em* corpus *para o estudo do russo.*
PALAVRAS-CHAVE: Linguística de corpus*; Pesquisa de* corpus *de aprendizes;* Corpora *de língua russa; Aquisição de segunda língua; Aquisição de língua de herança*

[*] University of Texas at San Antonio, Department of Bicultural Bilingual Studies, San Antonio, Texas, USA; https://orcid.org/0000-0003-2514-3107; olesya.kisselev@utsa.edu

Introduction

The wide-spread advancement of computer technology that gathered speed in the 1990s has resulted in significant changes in many social disciplines, including linguistics and applied language studies, which saw the increased prominence of the new discipline of *corpus linguistics* that focuses primarily on data-driven (rather than theory-driven) explorations of large and principally-organized language databases known as *language corpora*. Described as a methodology and a method (Gries, 2009; MCenery; Hardy, 2012), a practice and a "philosophical approach" (Leech, 1992), *corpus* linguistics utilizes the methods and instruments of computer-assisted analyses of language that allow researchers to analyze large quantities of authentic linguistic data to search for patterns, regularities and idiosyncrasies of language structure and language use across language modalities, varieties, registers, genres, and groups of speakers. The impact of *corpus* linguistics on the field of language studies has been significant, and is described by many linguists as nothing short of revolutionary (Hunston, 2002; Kopotev; Mustajoki, 2008; Gries, 2011, *inter alia*), contributing to every linguistic subfield.

The area of language pedagogy has, arguably, been one of the greatest benefactors of *corpus* linguistic approaches. Briefly, the convergence between the fields of language education and *corpus* linguistics has followed two major directions (Leech, 2014). One focuses on applying the knowledge culled from investigations of standard *corpora* to better serve pedagogical needs of language teachers and learners. This approach, for instance, has produced an array of modern-day evidence-based reference grammars, frequency dictionaries, phrasal lists, textbooks and other teaching/learning materials based on *corpus* data (Conrad; Biber, 2009; Biber; Conrad, 2010; Kopotev; Mustajoki, 2008; Lu et al., 2018; Lebedeva, 2020). In addition, language educators have been developing pedagogical methods and techniques for data-driven learning, an approach that allows for independent and semi-independent exploration of *corpus* data by language learners (Boulton, 2017).

The other *locus* of the convergence is in the application of *corpus* linguistic methods and tools to the study of *learner* language, that is the language produced by learners at different levels of linguistic proficiency, with an eye toward better understanding of the

developmental trajectories of linguistic behaviors, lacunas and abilities of those learning a language as a second (L2), foreign (FL), or heritage language (HL) (Granger, 2009; Leech, 2014).

Both directions have developed robustly over the course of the past three decades. Admittedly, the most progress has been made in the area of English as a second/foreign language (ESL/EFL), where the availability of well-developed standard and learner *corpora* and the embrace of *corpus* linguistic methods were early and supported through various institutionalized practices. Recent years, however, have seen some encouraging developments in Russian corpus linguistics, both with regard to standard *corpora* and learner *corpus* linguistics (Kisselev; Furniss, 2020; Lebedeva, 2020).

In the current paper, I provide a review of some of these developments, specifically in the area of *corpus*-based approaches to the study of Russian learner language,[1] and advocate for further advancement in the true convergence between Russian *corpus* linguistics and Russian second language acquisition studies (SLA).

1 Advances in the *Corpus*-Based Study of Russian Learner Language

Since the early 1990s, *corpus* linguists have argued for the value of *learner corpora* in language education. Learner *corpora* represent language produced by speakers whose command of the language has not yet reached maturity (Leech, 2014); these include first language/child language (L1) developmental *corpora*, second language learner *corpora* culled from L2 or FL speakers of the language at different levels of proficiency, and, lately, heritage language *corpora*, comprising language data from HL speakers and/or HL learners of a language. The major purpose of learner *corpora* is to "contribute to a better understanding of the universal, as well as language- and group-specific, patterns of

[1] I gladly refer the reader interested in the direct and indirect applications of *corpora* to many papers and volumes on the topic, including but not limited to: Dobrusina and Levinzon (2006), Mustajoki, Kopotev, Birjulin, and Protasova (2009); Alsufieva, Kisselev, and Freels (2012), Furniss (2013), Kisselev and Furniss (2020), Novikov and Vinogradova (2022), and the special issue on *Corpus* Linguistics in Teaching Russian as a Second Language of Russkij Yazyk za Rubezom (Ed. Lebedeva, 2020).

Second/Foreign language acquisition" (Kisselev, 2021, p.525). As such, learner *corpora* are instrumental to both the theoretical study of language acquisition and the applied purposes of creating better curricula, programs of study, and pedagogical materials for language learners.

Russian SLA, too, has made inroads in the development and investigation of learner *corpora*. The first publicly available *corpus*, the *Russian Learner* Corpus *of Academic Writing* (RULEC), is now over a decade old. A longitudinal *corpus* of advanced-level writing, it contains written texts (homework assignments, essays and research papers) created by Russian language students who were all enrolled in the same sequence of advanced Russian language courses at an American university. The unique feature of RULEC is its balanced distribution of data across language learning backgrounds, with 19 of the 36 authors in the *corpus* representing HL learners and the rest coming from the FL background[2]. This unique feature allows for a systematic comparison of developmental patterns in the language of FL and HL instructed learners. The *corpus* also provides other important types of *metadata*, or information about the texts and the learners who create them, such as *level of language proficiency* (on the ACTFL proficiency scale), *name of the course* for which the paper was written, *text type* (e.g., paragraph, essay, research paper), *function* targeted by the task (e.g., definition, narration, argumentation, etc.), and *time* restriction (timed or untimed writing). These metadata help researchers create sub*corpora* based on learner and text characteristics and compare these sub*corpora* along various linguistic parameters, with the goal of understanding relative effects of proficiency level, genre, topic, and other characteristics of learners and the texts they create on the linguistic features of the texts.

The original RULEC data is *raw*, i.e., the language is not lemmatized, tagged for parts of speech or syntactically annotated. Although all of these procedures have since become easily available (Kisselev, 2021), the first studies based on RULEC data utilized the raw data. In fact, certain research questions could be successfully investigated using only unparsed data with the help of appropriate *corpus*-analytic procedures. Such was the approach in Kisselev and Alsufieva (2017), who set out to analyze the dynamics of the use of complex

[2] For a more detailed description of RULEC design, compilation procedures, and purpose, as well as ideas for pedagogical use of the *corpus*, refer to Alsufieva *et al*. (2012) and Kisselev and Alsufieva (2017).

sentences with conjunctions by Russian learners at Intermediate to Advanced levels of proficiency. Drawing on the RULEC data, the authors created four sub*corpora* which separated the learner texts by level and background (HL Intermediate, HL Advanced, FL Intermediate, and FL Advanced), extracted a word list for each of the sub*corpora*, and then searched the word lists to establish which conjunctions the learners used in their writing. Using the list of extracted conjunctions as a guiding tool, the authors then conducted a comprehensive analysis of concordance lines (i.e., language samples containing all conjunctions in question) extracted from the four sub*corpora*. Having analyzed and categorized the extracted complex sentences, the authors assessed the quantitative changes in the structural and functional use of complex sentence structures, as well as the rates of accuracy and the types of error patterns across the HL and L2 groups at Intermediate and Advanced levels of language proficiency. For example, while there were no numeric changes in the amount of subordinate structures used by the FL students, the frequency of subordinate structures increased among the HL learners. However, the numbers actually converged at the Advanced level for both groups, suggesting that, perhaps, FL learners generally begin to acquire the skill of connecting ideas in writing through the use of various conjunctions earlier, since their exposure to Russian is heavily literacy-based from the beginning levels. HL learners, who tend to begin college-level courses and acquire academic literacy *after* having developed intermediate level oral skills, find themselves working on overt marking of complex syntax at the Intermediate and Advanced levels. Kisselev and Alsufieva also analyzed functional and structural types of sentences with conjunctions and found that the less frequent types and structurally more complex structures were better represented at the Advanced levels for both groups, with the HL learners exhibiting advantage over the FL learners with regard to structures that require structural manipulation of the constituents of the subordinated clause (e.g., *to, chto* ‘that;’ *chtoby* ‘in order to;’ *kotoryj* ‘which’).

A subsequent study (Kisselev; Kopotev; Klimov, forthcoming) addressed largely the same question of development of complex sentence structure but employed a more advanced computational analysis. First, the authors grammatically parsed the raw RULEC data using the trainable NLP application tool UDPipe (Straka; Straková, 2017), which provides

tokenization, lemmatization, and morphological and syntactic parsing of language data. Then, using in-house Python scripts, the researchers analyzed and compared data produced by four learner groups (HL Intermediate, HL Advanced, FL Intermediate, and FL Advanced) along twelve general syntactic complexity indices. These indices included: mean sentence length, proportions of coordinate and subordinate clauses per overall number of clauses, proportion of specific types of subordination (infinitive clauses, adverbial clause modifiers, and relative, gerund and participle clauses), and measures of phrasal "depth" (i.e., maximum and mean nesting depth of a syntactic phrase, as well as the number of phrases with "shallow" nesting depth). The results of the study supported most of the observations of the previous study by Kisselev and Alsufieva (2017); for example, the results of both studies aligned with the conclusions of many previous syntactic complexity studies conducted on L2 *corpus* data and demonstrated overall complexification of syntax in the writing of more advanced learners. And while Alsufieva and Kisselev's (2017) study was largely descriptive, the computational approach of the Kisselev et al. (2021) study also has implications for Russian language assessment showing how specific syntactic features correlate with various proficiency levels in learners of Russian.

The difference between the two studies is not simply the difference between the possibilities of grammatically tagged vs. raw data; the fact of the matter is that different research questions may necessitate different treatment of the data and a different combination of qualitative and computational methods. For example, the focus of Peirce's (2018) study, which also utilized the RULEC data, was on tracking the development of accuracy in nominal morphology involving a genitive case in nouns, adjectives, and determiners. By setting out to analyze this specific type of error (i.e., genitive case errors), the author had to resort to a method that integrated manual coding of errors and the tagging of those errors for subsequent computational analysis using particular software (here oXygen XML Editor). Combining the benefits of human rater analysis with the effectiveness of *corpus*-based procedures allowed Pierce (ibid.) to consider different factors possibly affecting the development of this morphological feature in learners of Russian. The study made use of the meta-data available in the RULEC *corpus*, specifically time constraint in the writing of text (timed or non-timed condition) and language learning background (HL or FL), as independent variables.

Comparing the rate and types of errors by group and by time constraint allowed the author to discuss the results of the study in light of the central role that early/late exposure to language plays in language attainment, both in possible representations of nominal functional features in two groups of learners and in processing constraints that the two groups of learners may be subject to in timed task conditions.

As the studies reviewed above demonstrate, a *corpus* study may be more or less technology-dependent to best address the research foci of the investigators (and perhaps, their level of familiarity with *corpus*-based procedures). However, the potential of error-tagged learner *corpora* cannot be overstated. Systematic error analysis, such as grouping errors by frequency, by group characteristics (such as proficiency levels, age, or parental involvement), and by structural and functional properties can shed light on developmental processes of language development and the factors that influence it. Error analysis can help test hypotheses about the relative effects of L1 interference and L2/HL proficiency, understand the impact of instructional practices and different learning histories, and answer many other important questions that are still largely under-researched in heritage language acquisition.

An ambitious large-scale *corpus* project, the Russian Learner *Corpus* (RLC, http://web-*corpora*.net/RLC) promises to provide the field of Russian language studies with its first fully error-tagged *corpus*. Although the *corpus* is still under construction and the sub*corpora* are not well balanced, the repository currently houses a large collection of texts, oral and written, (appr. three thousand speech samples, Rakhilina et al., 2016) produced by L2 and HL speakers of Russian, representing different levels of language proficiency and a variety of dominant languages (currently over 20 different L1s are listed on the website). The RLC is readily available in raw and POS-tagged forms, and at least a significant part of the *corpus* is set to be error-tagged.

In a recent study, Eremina (2020) has utilized the tagged parts of the RLC *corpus* (indiscriminately, regardless of L1 background) using the error tag “Idiom” that marks an infelicitous multiword expression. The researcher categorized the extracted infelicitous expressions into two main types, structural and semantic, and then analyzed the sub-types further, hypothesizing on the nature of each error. Although the study does not venture to

implement any statistical procedures, it lays the foundation for subsequent statistical analyses of various types of phraseological expressions in the language of L2 learners of Russian. Given the increased attention that the fields of SLA and language pedagogy are paying to L2 learners' ability to successfully use formulaic expressions in their target language, studies that address the development of phraseological complexity in L2 Russian are much needed.

While the work conducted by the RLC team requires manual tagging, the field of computational linguistics is grappling with issues of automatic error detection and correction. A number of research projects have been devoted to the methodological issue of automatic error detection in morphologically rich languages, including Russian (Rosen et al., 2014; Rozovskaya; Roth, 2018). The more learner *corpora* are available to these researchers, the better they can train computational models to recognize specific developmental patterns in language data.

Fortunately, the development of Russian learner *corpora* is on the rise. One such project is the Multilingual Academic *Corpus* of Assignments – Writing and Speech (Macaws, https://sites.google.com/email.arizona.edu/macawswebinar/home), which includes Russian learner data collected through regular classroom activities. The *corpus* is available online; it currently has over a thousand texts produced by 100 students of Russian, mostly from their first and second years of instruction (for more information on the *corpus* see Novikov and Vinokurova, 2022). Two other current learner *corpus* projects are also in development (both are available upon request). The Middlebury Russian *Corpus* of Learner Language (MiRuCLL, Kisselev et al., forthcoming) is a developmental *corpus* that contains data collected from L2 learners of Russian at the beginning and end of an intensive immersive summer program. The unique feature of this *corpus* is the availability of information on the student's proficiency level at the beginning and end of the instructional period. Proficiency assessment is based on the ACTFL proficiency scale, making the data potentially comparable to many other such data samples.

Another Russian learner *corpus* featured in current research is the Russian Essay *Corpus* (Kisselev, 2019; Kisselev et al., forthcoming) The *corpus* is compiled from texts drawn from the annual National Post-Secondary Russian Essay Contest (NPSREC) sponsored by the American Council of Teachers of Russian (ACTR). An annual event, the

NPSREC attracts wide participation from students of Russian across the U.S. Upon completion of the award cycle, fully anonymized learner essays are made available to researchers. So far, at least one year of the NPSREC data has been collected and processed as a stand-alone cross-sectional *corpus* of Russian learner data. While the RULEC *corpus* has a small number of participants who contributed a lot of texts over an extensive period of time, the Essay *corpus* represents a large number of students from a variety of programs across the country. The proficiency levels of the *corpus* contributors are indexed as instructional hour ranges (level 1 includes learners who received less than 100 hours of instruction, level 2 between 100 and 200 hours of instruction and so on); however, a small portion of the *corpus* texts are also rated along the ACTFL proficiency scale. The language learning background distinguishes between HL and FL learners. The Essay contest has the potential to yield results that are readily generalizable across various groups of Russian language learners.

These *corpora* are becoming an important tool for Russian language researchers and Russian language teachers, as investigations of these *corpora* have the potential to significantly enrich our understanding of the developmental paths of Russian language learners, aid in assessment practices, and help evaluate instructional practices. However, in order to bring this promise to full fruition, more and bigger learner *corpora* and many more *corpus*-based research studies are needed in the field. In the following section, I describe some practical considerations and specific steps in creating custom-built learner *corpora* and conducting *corpus*-based studies.

2 Starting with *Corpus* Linguistics Research: A Few Know-Hows

2.1 Data Collection and *Corpus* Compilation

Not every language dataset can be called a *corpus*; in fact, *corpus* compilation requires careful consideration on behalf of the researcher and considerable planning. The

specific principles for collecting and processing data that are entered into a *corpus* bear as much importance in a *corpus*-based study as the computational methods used in the analysis. These principles include the authenticity and size of language data, as well as systematicity of data selection, data representativeness and its balanced-ness.

Authenticity of corpus *data.* One of the most important principles of *corpus* linguistics is its focus on authentic language, i.e., language as it is used by its speakers in authentic communicative contexts; investigating authentic language, rather than language samples created for linguistic experiments, is thought to overcome the potential biases that encroach data collected in experimental settings. Many contemporary *corpora* that are now collected for specific purposes, especially the learner *corpora*, effectively represent elicited data. Nonetheless, these elicited data come in the form of elicited narratives, interviews, and other types of situationally grounded discourses. Authenticity also allows the inclusion of contextual and situational aspects into the analysis by recording them as meta-information.

To ensure that the results of a *corpus* analysis are generalizable, *corpora* are normally large. At the same time, *the size of a* corpus is a relative standard; on the one hand, corpora need to be large enough to allow for the application of statistical analyses and generalizable statistical operations, but they can be smaller if the aim of the *corpus* is narrowed to a specific research question or a local context. Thus, a set of classroom essays and/or oral presentations collected at regular intervals during an academic year or even a semester from the same group of students may become a *corpus* to be used to assess the students' progress or the effectiveness of instructional approaches in this specific instructional context (Biber, Conrad, & Reppen, 2004).

Systematic data selection ensures that data sampling is not random and is clearly relevant for research questions. Notice that even in the case of a small-scale classroom-based *corpus*, a teacher-researcher must pay attention to the systematicity of the data collection, considering the intervals at which the data are collected, the mode of data collection (e.g., at home or in class, hand-written or typed, etc.), and the type of data (e.g., the genre and modality of language production). The systematicity principle is inherently connected to the principle of data *representativeness*, which ensures that the data found in the *corpus* represents a specific mode, variety, or genre of a language or a particular group of speakers

as fully as possible. To illustrate, a large national *corpus* such as the Russian National *Corpus* (RNC, https://rus*corpora*.ru) will amount to many millions of words, and its representativeness is achieved by including texts from multiple modes of communication (written, spoken, multimodal, and intermediary modes such as text messaging), of various genres, and by different authors representing diverse regional and historical varieties. Thus, results of a large-scale investigation done on the data of the RNC could be considered representative of the state of the Russian language today. The learner *corpora* of Russian reviewed in the section above, for example, are representative (to a varying degree) of Russian learners with English as their L1, and thus, certain observations and conclusions based on the study of these *corpora* may not be generalizable across all L1 backgrounds. And, finally, the data entered into the *corpus* has to be *balanced*-out across individual authors, text types, registers, modes, etc. For example, in the case of classroom-based *corpora*, the researcher must ensure that the number of texts authored by the participating students is reasonably equal, that these texts are somewhat similar in length, and/or that no one textual genre or mode is over-represented in the *corpus*; this is done in order to avoid the potential effect of one (or a small sub-set) of the text parameters. Compilers of large-scale *corpora* often must go to great lengths in order to ensure that the *corpus* data are balanced, to ensure the effectiveness and meaningfulness of subsequent analytical procedures.

As one can surmise based on the principles described above, a *corpus* is not just any (large) set of linguistic data; a language *corpus* is a sizable and machine-readable, systematically compiled, balanced collection of authentic texts that are representative of a language or a specific language variety. The following subsection reviews how a researcher can further process and analyze the *corpus* data.

2.2 *Corpus* Annotation

Once the *corpus* is designed well and the data are collected, they must then be systematically described. As mentioned in the previous section, text description is provided in the form of meta-tags, which typically accompany each text or file in the *corpus*. Such

meta-tags can include the name of the text author (or any unique text ID such as a pseudonym or a number), biodata (age, gender, first language), date of creation/occurrence, genre of the text, and any other metadata that may be useful to the purposes of the *corpus*. Metadata descriptors can then be used as variables in analyses of the *corpus* data. The RULEC *corpus*, for instance, records various text and learner characteristics in the "Header Identification Box," as illustrated below (see Illustration 1). Using such information can help the researcher group data along some of these parameters and/or, in general, account for these learner and text parameters as variables potentially affecting the linguistic parameters of the linguistic production.

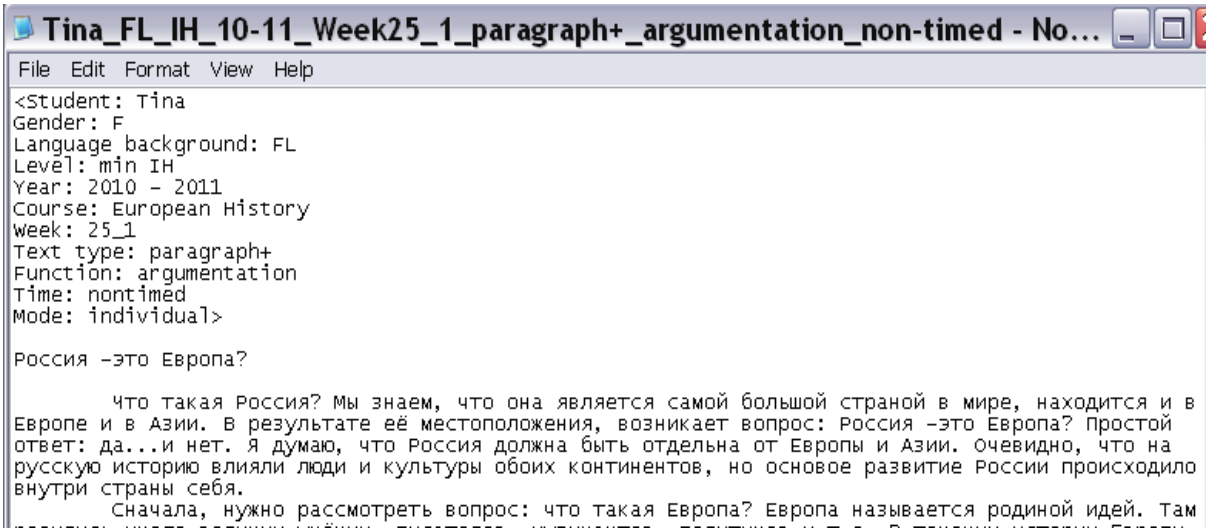

Illustration 1. RULEC *corpus* text header ID. Reprinted with permission from Alsufieva *et al*., 2012

While metadata is a *sine qua non* of *corpus* design and compilation, additional information may also be added to label or *annotate* words, sentences, and any longer or shorter meaningful units of text. Annotation (or mark-up) can provide different information about text-level units and may include morpho-syntactic information (e.g., Parts-of-Speech annotation, as well as person, number, gender, case, voice, tense, aspect), syntactic information (e.g., sentence parsing), semantic information (e.g., word-sense disambiguation, animacy, count/non-count), discoursal information (e.g., speech acts), error-tags, and/or any other information needed for a research-specific *corpus*.

This additional information is “attached” to relevant linguistic units in the form of *tags*, which is why annotation is often referred to as *tagging*. See Table 1 for an example of a learner sentence parsed with the UDPipe parser.

Table 1: A sample of a UDPipe annotation output. Reprinted with permission from Kisselev, Kopotev and Klimov, forthcoming

ID	Token	Lemma	POS	Morphological annotation	Syntactic ID	Syntactic annotation
1	Мы	мы	PRON	Case=Nom|Number=Plur|Person=1	2	nsubj
2	живём	жить	VERB	Aspect=Imp|Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act	0	root
3	в	в	ADP	_	4	case
4	мире	мир	NOUN	Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing	2	obl
5	,	,	PUNCT	_	7	punct
6	где	где	ADV	Degree=Pos	7	advmod
7	ничего	ничего	ADV	Degree=Pos	4	acl:relcl
8	,	,	PUNCT	_	10	punct
9	просто	просто	PART	_	10	advmod
10	чёрная	черный	ADJ	Case=Nom|Degree=Pos|Gender=Fem|Number=Sing	7	conj
11	и	и	CCONJ	_	12	cc
12	белая	белый	ADJ	Case=Nom|Degree=Pos|Gender=Fem|Number=Sing	10	conj
13	.	.	PUNCT	_	2	punct

As discussed in the previous section, the level of annotation needs first and foremost to be necessitated by the research focus of a *corpus* project, and other annotation schemas, either commercially or publicly available or custom-built, may be applied to the data.

2.3 *Corpus* Analytic Tools and Procedures

A well-compiled and well-described *corpus* can be subjected to an array of possible statistical procedures; the majority of these procedures fall under some type of data retrieval, obtaining frequencies, and statistical analysis. These analyses are conducted with the help of *corpus*-analytic and programs software which can be stand-alone (downloadable onto one's personal computer) or web-based. The most commonly used stand-alone programs are the license-based WordSmith Tools (Scott, 2016) and the freely-downloadable AntConc (Anthony, 2019). A host of web-based tools provide similar analytic procedures (see, for example, the suite of tools The NLP Tools for Social Sciences, Kyle & Crossley, 2015, or LancsBox, Brezina et al., 2020). The functionality of these programs may vary, but effectively they are all designed to provide language researchers with tools and ways of quick, automatic, and meaningful ways of *corpus* data sorting, extraction, and analysis. Utilizing such computational tools, a researcher can conduct various analytical procedures. Some of the common procedures that help analyze *corpus* data include the following.

Retrieving descriptive statistics. An array of general descriptive statistical information about the *corpus* data can easily be obtained with the help of even basic *corpus* tools. A researcher can quickly and automatically obtain information on the number of words and word tokens, number of sentences, the number of paragraphs, and, importantly, length of these linguistic units, etc. Multiple studies have shown that length-based measures may successfully index learner language development. For example, an increase in the length of a text produced within a certain window of time, as well as the length of a clause (i.e., a mean number of words per clause) and the length of a sentence (i.e., mean number of words per sentence) may indicate overall development or growing proficiency (Norris; Ortega, 2009; Bulté; Housen, 2012; Polat et al., 2019; Kisselev et al., forthcoming, *inter alia*). Even word length has been shown to grow with proficiency level in Russian (Kisselev et al., forthcoming).

Descriptive statistics also often include information on type/token ratio (TTR), that is, the percentage of unique words (lemmas) or word forms per all words in the *corpus*. TTR dynamics that illustrate the diversity of learners' vocabulary have also shown to index

qualitative differences in language proficiency (Lee; Jang; SEO, 2009; Kisselev et al., forthcoming).

Extracting word lists. With this procedure, a researcher can compile a list of words (lemmas or wordforms) used in the *corpus* (or sub-*corpora*); the list(s) can be sorted either in alphabetical order or by frequency and can subsequently be compared to either standard frequency dictionaries (e.g., *Novyj chastotnyj slovar' russkoj leksiki* [The new frequency dictionary of Russian lexis], Ljashevskaja & Sharov, 2010), learner dictionaries (e.g., *A frequency dictionary of Russian: Core vocabulary for learners*, Sharoff et al., 2014) or lexical minimum lists (e.g., *Leksicheskij minimum po russkomu kak inostrannomu* [Lexical minimum lists for Russian as a foreign langauge], 2013). Comparing learner output with these existing lexical resources and/or words lists based on the *corpora* of students with various proficiencies or backgrounds can help researchers quickly assess the lexical skills of learners.

Word lists are also a useful starting point for many more qualitative inquiries, providing a first glance at learner language and some patterns in lexical knowledge. One can quickly assess over-use/under-use of lexical items, see errors and error patterns, or simply scout the lexical data for further analysis of lexical use patterns. For instance, one of the first Russian learner *corpus* studies, conducted by Pavlenko and Driagina (2007), focused on the acquisition of emotion vocabulary by L2 speakers of Russian (with English L1). The researchers collected three small *corpora* of oral narratives produced by American Russian language students (in Russian), Russian monolinguals (in Russian), and American monolinguals (in English). The authors compared the frequencies and appropriateness of emotion words (e.g. *rasstraivaetsja* 'get upset,' *grustnoe* 'sad') and their stems (e.g. *rasstra/o-* 'upset,' *grus*t- 'sad') among the three groups and found that unlike Russian monolinguals, who showed strong preference for verbs when describing emotion states, the learners preferred adjectival constructions in Russian (similar to monolingual Americans speaking in L1 English); the learners also used a smaller range of emotion words, and often confused or violated conceptual restrictions on the use of emotion vocabulary (e.g., by employing *razozlilas'* 'got mad' instead of *rasstroilas'* 'got upset'). An array of research

questions (mostly with a vocabulary focus) can be conducted using raw *corpora* and these simple procedures.

Retrieving and sorting concordance lines. Concordances are language samples that can be automatically extracted from a *corpus*. Concordances contain the search term (usually a word, a phrase, or even a word stem) that the researcher chooses to investigate. Concordances can be sorted in different ways: alphabetically or by words to the left/right of the search term, etc. Such sorting allows the researcher to identify different patterns in the data. If the *corpus* data is grammatically parsed, one can also extract concordances using grammatical tags as search items.

Retrieving lists of collocates and colligates. Collocations are multi-word units or lexical strings that co-occur with the search term more frequently than would be expected by chance. These formulaic expressions (or *nesvobodnye slovosocetanija* in Russian) are notoriously challenging for language learners and are a prime candidate for SLA research and pedagogical intervention alike. Formulaic expressions in Russian represent an array of structural types and include Adjective+Noun strings (e.g., *sloznaja problema, trudnaya zadaca, krepkij caj, sil'noe lekarstvo*), Adverb+Verb (*krepko zadumat'sja, sil'no tolknut'*), Preposition+Noun (*na rabote, v stole*), etc. A researcher can potentially extract all n-grams (i.e., string of two, three or more words) from a *corpus* and analyze them by hand or employ statistical analyses built into the *corpus*-analytic software to establish a list of recurring patterns.

Relatedly, *colligation* refers to the phenomenon of formulaicity but with grammatical, rather than lexical, constructions. For example, a construction "*igrat' v + accusative*" and "*igrat' na + prepositional*" are colligations. Apresjan (2017) is a fitting illustration of colligation research on L2 and HL Russian. The study investigates Russian possessive constructions with and without the overtly expressed existential verb *est'* using the data from the RLC. The *corpus* search was formulated as "*u* + gen (noun, pronouns) + *est'*" and "*u* + gen (noun, pronouns) + *nom (noun)*" (Apresjan, 2017, p.86). The author analyzed the extracted concordance lines with an eye towards understanding whether specific semantic and pragmatic rules govern the usage of these constructions by HL and L2 learners. The results revealed that HL learners can use the constructions felicitously within all semantic

and pragmatic meanings, while the L2 data contain a number of erroneous instances. Specifically, the L2 learners made twice as many errors with the constructions with unexpressed *est'*, suggesting that L2 learners of Russian may require additional instruction with regard to this structure.

The development of L2 learners' phraseological abilities is an area of increased interest in the field of SLA (Paquot; Granger, 2012). And potential research in this area is now made easier with the development of phrasal dictionaries (e.g., *Slovar' russkoj idiomatiki*, Kustova, n.d.; *Slovar' glagol'noj socetaemosti nepredmetnyx imjon russkogo jazyka*, Biriuk, Gusev, & Kalinina, n.d.) and platforms for investigating collocations and colligations in large standard *corpora* (e.g., *CoCoCo*, Kopotev, 2020), which provide specific information on lexical and grammatical patterns in standard Russian and can serve as a baseline in the analysis of learner data.

The procedures described above are a few of many. The number – and sophistication – of *corpus*-based procedures available today is continuously expanding; however, the main purpose of these tools is to allow a researcher to engage with large quantities of authentic data, and extract and examine multiple samples of linguistic units produced by speakers and writers of the language varieties in focus. By extracting, sorting, and analyzing (statistically or manually) the linguistic structures chosen for analysis, the researcher can look for regularities and patterns of language use that otherwise escape the intuitions of language researchers and language teachers.

Conclusion and Desiderata

By and large, the task of second/heritage language researchers is to understand the mental processes that underlie language production and development in L2 and HL users. Language *corpora* composed of linguistic data produced in authentic settings for communicative purposes have become instrumental in providing language researchers with evidence for the interpretation of these mental processes and the mental representations of knowledge. Coupled with sophisticated computational tools that allow for fast and reliable

extraction and analyses of the data, language *corpora* have proved to be an indispensable tool in linguistic research, and the pedagogical implications of such research are significant for language classroom practices. By embracing *corpus*-based approaches, the fields of Russian SLA and Russian language education stand to benefit tremendously, both through expanding our understanding of the nature of Russian L2 and HL development and through expanding the pedagogical approaches and repertoires of Russian language teaching and learning.

REFERENCES

ALSUFIEVA, A.; KISSELEV, O.; FREELS, S. Results 2012: Using Flagship Data to Develop a Russian Learner *Corpus* of Academic Writing. *Russian Language Journal*, n. 62, pp.79-105, 2012.

ANDRJUSHINA, N.; KOZLOVA, T. *Leksicheskii minimum po russkomu yazyku kak inostrannomu. Bazovyj Uroven'* [*Lexical minimum for Russian as a foreign language. Basic level*]. 5.ed. St. Petersburg: Zlatoust, 2020.

ANTHONY, L. AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available on: http://www.laurenceanthony.net/software, 2019.

APRESJAN, V. YU. Russkie possessivnye konstrukcii s nulevym i vyraženynnym glagolom: pravila i ošibki. *Russkij jazyk v naučnom osvesčenii*, n. 33, pp.86-116, 2017.

BIBER, D.; CONRAD, S.; REPPEN, R. Corpus *Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press, 2004.

BIBER, D.; CONRAD, S. Corpus *Linguistics and Grammar Teaching*. White Plains, NY: Pearson Education, 2010.

BIRIUK, O.; GUSEV, V.; KALININA, YE. *Slovar' glagol'noj sochetaemosti nepredmetnyx imion russkogo yazyka* [Dictionary of Verbal Compatibility of Non-Objective Names of the Russian Language]. Available from http://dict.ruslang.ru/abstr_noun.php.

BOULTON, A. Data-Driven Learning and Language Pedagogy. *In:* THORNE, S., MAY, S. (eds.). *Language, Education and Technology.* Encyclopedia of Language and Education. New York: Springer, Cham, 2017.

BREZINA, V.; WEILL-TESSIER, P.; MCENERY, A. #LancsBox v. 5.x. [software]. 2020. Available from http://*corpora*.lancs.ac.uk/lancsbox.

BULTÉ, B.; HOUSEN, A. Conceptualizing and Measuring Short-Term Changes in L2 Writing Complexity. *Journal of Second Language Writing*, n. 26, pp.42-65, 2014.

CONRAD, S.; BIBER, D. *Real Grammar: A* Corpus-*Based Approach to English.* New York: Pearson/Longman, 2009.

CROSSLEY, S. A.; KYLE, K. Assessing Writing with the Tool for the Automatic Analysis of Lexical Sophistication (TAALES). *Assessing Writing*, n. 38, pp.46-50, 2018.

DONRUSHINA R. N.; LEVINZON, A. I. Informatsionnye tehnologii v gumanitarnom obrazovanii: Natsional'nyj korpus russkogo yazyka [Information Technologies in Humanities Education: National *Corpus* of the Russian Language]. *Voprosy obrazovaniia*, n. 4, 2006.

EREMINA, O. S. Russkie nesvobodnye vyrazhenia v rechi inostrantsev: korpusnyi podhod [Russian Formulaic Expressions in the Speech of Foreigners: *Corpus* Approach]. *Russkii jazyk za rubezhom*, n. 6, pp.29-35, 2020.

FURNISS, E. Using a *Corpus*-Based Approach to Russian as a Foreign Language Materials Development. *Russian Language Journal*, n. 63, pp.195-212, 2013.

GRANGER, S. The Contribution of Learner *Corpora* to Second Language Acquisition and Foreign Language Teaching. *In:* AJMER, K. (ed.). Corpora *and Language Teaching*. Philadelphia/Amsterdam: John Benjamins, 2009, pp.13-32.

GRIES, S. What is *Corpus* Linguistics? *Language and Linguistics Compass*, v. 3, n. 5, pp.1225-1241, 2009.

GRIES, S. Methodological and Interdisciplinary Stance in *Corpus* Linguistics. *In*: BARNBROOK, G.; VIANA, V.; ZYNGIER, S. (eds.). *Perspectives on* Corpus *Linguistics: Connections and Controversies.* Philadelphia/Amsterdam: John Benjamins, 2011, pp.81-98.

HUNSTON, S. Corpora *in Applied Linguistics*. Cambridge: Cambridge UP, 2002.

KISSELEV, O. *Corpus*-Based Methods in the Study of Heritage Languages. *In:* POLINSKY, M.; MONTRUL, S. (eds.). *The Cambridge Handbook on Heritage Languages.* Cambridge University Press, 2021, pp.520-544.

KISSELEV, O. Word Order Patterns in the Writing of Heritage and Second Language Learners of Russian. *Russian Language Journal*, n. 69, pp.149-174, 2019.

KISSELEV, O.; KOPOTEV, M.; KLIMOV, A. Specific Markers of Syntactic Complexity in Academic Russian: A Longitudinal *Corpus* Study. *In:* LEŃKO-SZYMAŃSKA, A.; GÖTZ, S. (eds.). *Complexity, Accuracy & Fluency in Learner* Corpus *Research*. John Benjamins, forthcoming.

KISSELEV, O.; FURNISS, E. *Corpus* Linguistics and Russian Language Pedagogy. *In:* DENGUB, E.; DUBININA, I.; MERILL, J. (eds.). *The Art of Teaching Russian*. Washington: Georgetown University Press, 2020, pp.307-332.

KISSELEV, O.; ALSUFIEVA, A. The Development of Syntactic Complexity in the Writing of Russian Language Learners: A Longitudinal *Corpus* Study. *Russian Language Journal*, n. 67, pp.27-53, 2017.

KOPOTEV, M. Ispol'zovanie èlektronnyx korpusov v prepodavanii russkogo jazyka [The Use of Electronic *Corpora* in Teaching the Russian Language]. *In*: LINDSTEDT J. *et al*. (eds.), *SLAVICA HELSINGIENSIA 35, S ljubov'ju k slovu, Festschrift in honour of Professor Arto Mustajoki on the occasion of his 60th birthday*. Helsinki, 2008, pp.110-118.

KOPOTEV, M. O samom slozhnom: Izuchenie sochetaemosti slov online [About the Most Difficult: Learning the Combination of Words Online]. *Russkij jazyk za rubezhom*, n. 6, pp.36-43, 2020.

KOPOTEV, M.; MUSTAJOKI, A. Sovremennaja korpusnaja rusistika [Modern *Corpus* Russian Studies]. *In:* MUSTAJOKI, A.; KOPOTEV, M.; BIRJULIN, L.; PROTASOVA, YU. (eds.). *Instrumentarij rusistiki: Korpusnye podxody*. Helsinki: Helsinki UP, 2008, pp.7-24.

KUSTOVA, G.I. *Slovar' russkoi idiomatiki. Sochetaniya slov so znacheniyem vysokoi stepeni* [A Dictionary of Russian Idiomology. Word Combinations with the Significance of a High Degree]. Moscow, 2008. http://dict.rislang.ru/magn.php.

KYLE, K.; CROSSLEY, S. A. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, v. 49, n. 4, pp.757-786, 2015.

LEBEDEVA M. YU. Dano mne telo – chto mne delat' s nim? Primenenie korpusnyh tehnologii v lingvodidaktike RKI [I Have Been Given a Body - What Am I to Do with It? Application of *Corpus* Technologies in Linguodidactics of Russian as a Foreign Language.]. *Russkij jazyk za rubezhom*, n. 6, pp.4-13, 2020.

LEECH, G. Corpora and Theories of Linguistic Performance. *In:* SVARTVIK, J. (ed.). *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin, New York: Mouton de Gruyter, 1992, pp.105-122.

LEECH, G. Teaching and Language *Corpora*: A Convergence. *In:* WICHMANN, A. *et al*. (ed.). *Teaching and Language* Corpora. London and New York: Routledge, pp.1-24, 2014.

LEE, S. H.; JANG, S. B.; SEO, S. K. Annotation of Korean Learner *Corpora* for Particle Error Detection. *CALICO Journal*, v. 26, n. 3, pp.529-544, 2009.

LJASHEVSKAJA, O. N.; SHAROV, S.A. *Chastotnyi slovar' sovremennogo russkogo yazyka:* Na materialax Natsional'nogo korpusa russkogo yazyka [Frequency Dictionary of the Modern Russian Language: On the Materials of the National Russian Corpus]. Azbukovnik, 2009.

LU, X.; YOON, J.; KISSELEV, O. Adding to Academic Formula Lists: Phrase-Frames for Research Article Introductions in Social Sciences. *Journal of English for Academic Purposes*, v. 36, pp.76-85, 2018.

MCENERY, T.; HARDIE, A. Corpus *Linguistics: Method, Theory and Practice*. Cambridge: Cambridge UP, 2012.

NORRIS, J.; ORTEGA, L. Measurement for Understanding: An Organic Approach to Investigating Complexity, Accuracy, and Fluency in SLA. *Applied Linguistics*, v. 30, n. 4, pp.555–578, 2009.

NOVIKOV, A.; VINOKUROVA, V. Learner Corpus as a Medium for Tasks. *In:* NUSS, S. V.; WHITEHEAD MARTELLE, W. (eds.). *Task-Based Instruction for Teaching Russian as a Foreign Language*. London and New York: Routledge, 2022.

PAVLENKO, A.; DRIAGINA, V. Russian Emotion Vocabulary in American Learners' Narratives. *The Modern Language Journal*, n. 91, pp.213-234, 2007.

PAQUOT, M.; GRANGER, S. Formulaic Language in Learner *Corpora*. *Annual Review of Applied Linguistics*, v. 32, n. 1, pp.130-149, 2012.

PEIRCE, G. Representational and Processing Constraints on the Acquisition of Case and Gender by Heritage and L2 Learners of Russian: A *Corpus* Study. *Heritage Language Journal*, v. 15, n. 1, pp.95-111, 2018.

POLAT, N.; MAHALINGAPPA, L.; MANCILLA, R. L. Longitudinal Growth Trajectories of Written Syntactic Complexity: The Case of Turkish Learners in an Intensive English Program. *Applied Linguistics*, v. 41, n. 5, pp.688-711, 2020.

RAKHILINA, E.; VYRENKOVA, A.; MUSTAKIMOVA, E.; LADYGINA, A.; SMIRNOV, I. Building a Learner *Corpus* for Russian. *In:* VOLODINA, E. *et al*. (ed.). *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*. Umea, Sweden: LiU Electronic Press, 2016, pp.66-75.

ROSEN, A.; HANA, J.; ŠTINDLOVÁ, B.; FELDMAN, A. Evaluating and Automating the Annotation of a Learner *Corpus*. *Language Resources and Evaluation*, v. 48, n. 1, pp.65-92, 2014.

ROZOVSKAYA, A.; ROTH, D. Building a State-of-the-Art Grammatical Error Correction System. *Transactions of American Computational Linguistics*, v. 2, pp.419-434, 2014.

SCOTT, M. *WordSmith Tools Version 7* [Computer Program]. Stroud: Lexical Analysis Software, 2016.

SHAROFF, S.; UMANSKAYA, E.; WILSON, J. *A Frequency Dictionary of Russian: Core Vocabulary for Learners*. London and New York: Routledge, 2014.

STRAKA, M.; STRAKOVÁ, J. Tokenizing, Pos Tagging, Lemmatizing and Parsing ud 2.0 with Udpipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2017, pp.88-99.

Received September 26, 2021

Accepted September 05, 2022

Reviews

Due to the commitment assumed by *Bakhtiniana*. Revista de Estudos do Discurso [Bakhtiniana. Journal of Discourse Studies] to Open Science, this journal only publishes reviews that have been authorized by all involved.

Review II
This is a very interesting overview of Russian LCR.
The authors should make these two minor changes:
> since the 1990s has resulted in significant changes in many social disciplines, including linguistics and applied language studies, which saw the emergence and increased prominence of the new discipline of corpus linguistics
Computer Corpus Linguistics goes back further, to the 1960s. Please amend this.
> the repository currently houses a large collection of texts
Please specify how large (number of texts and/or words).

Tony Berber Sardinha – Pontifícia Universidade Católica de São Paulo – PUC-SP, São Paulo, São Paulo, Brazil; https://orcid.org/0000-0001-8815-1521; tonycorpuslg@gmail.com

Research Data and Other Materials Availability
The contents underlying the research text are included in the manuscript.