Visualizing Russian: Illuminating Corpora, Conjugations, and Classrooms

Steven J. Clancy
Paige Lee

Follow this and additional works at: https://scholarsarchive.byu.edu/rlj

Part of the Slavic Languages and Societies Commons

Recommended Citation
Available at: https://scholarsarchive.byu.edu/rlj/vol72/iss1/2

This Article is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Russian Language Journal by an authorized editor of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Visualizing Russian: Illuminating Corpora, Conjugations, and Classrooms

Steven J. Clancy, Paige Lee

Author Note
The Visualizing Russian project has received support from the Barajas Dean’s Innovation Fund for Digital Arts and Humanities from the office of the Dean of Arts and Humanities at Harvard University, from the Davis Center for Russian and Eurasian Studies at Harvard University, and from the Foreign Language Advisory Group (FLAG) at Harvard University. The project has received developmental support since 2014 from Arts and Humanities Research Computing, particularly from Arthur Barrett, Senior Software Engineer in Academic Technology for the Faculty of Arts and Sciences (FAS), with previous input and development from Christopher Morse and advice from Cole Crawford.

Correspondence concerning this article should be addressed to Steven Clancy, Slavic Languages and Literatures, Harvard University, Barker Center, 12 Quincy Street, Cambridge, MA 02138, United States. Email: sclancy@fas.harvard.edu

1. Introduction: Visual tools for language learners, teachers, and linguists
The Visualizing Russian (Clancy, 2014–2022) project offers a suite of tools benefitting language learners, teachers, and linguists and enabling each of these user groups to access the complex system of the Russian language through visualization methods in order to leverage the powers of compression and expansion of a massive data set. Users can analyze texts for relative difficulty with regard to vocabulary content with Visible Vocabulary, create frequency lists and identify the most commonly used word forms for each lemma in a text or compare sets of target vocabulary to covered forms in a particular text with the Mini-Story Creator, compare the relative frequency of near-synonyms or other items in a semantic domain with the Quick Lemma tool, view the usage frequency of various nouns in particular cases with the Case Distribution tool, and identify
the case governance and preposition usage of Russian verbs in the *Verb Histogram* tool. Additional components provide word-formation analysis by breaking down words into component prefixes-roots-suffixes, gauge imperfective/perfective aspect usage for individual verb pairs, visualize verb forms across various person/number/tense/aspect combinations, and identify the field of morphologically related or semantically related words for any target word.

Along the way, the project has also provided an ideal undergraduate research opportunity outside the usual literature and culture framework for a student majoring in Slavic Languages and Literatures and Computer Science. With input and guidance from the rest of the team, Paige Lee, the undergraduate coauthor of this paper, has designed, prototyped, tested, and deployed tools to the *Visualizing Russian* website in addition to contributing to the maintenance and development of the data set. She has also analyzed Russian language data sources such as the disambiguated morphological standard subcorpus of the Russian National Corpus (RNC subcorpus) (Lyashevskaya et al., 2005) to bring “real world” statistics and examples to the tools to demonstrate hot spots in paradigms based on actual language usage. In the process of developing these tools, she has explored novel web visualization techniques for language data using JavaScript and the D3 library; gained a deeper knowledge of the intricacies of Russian morphology, semantics, syntax, and grammar; applied concepts in statistical language analysis such as word embeddings and principal component analysis; and learned about the role that digital pedagogical tools can play in Russian language teaching and learning. These tools have also contributed to the creation of a new Russian textbook series, *Foundations of Russian* (Clancy et al., in press), which presents a curated 4,000-word beginner-to-intermediate vocabulary target based on the most frequently occurring and communicatively necessary words in Russian. These textbooks are informed by frequency and leverage research from usage-based, cognitive, and constructional approaches to linguistics.¹

¹ The *Foundations of Russian* textbook project shares similar goals with the *Min russiske reise* [My Russian Journey] textbook under development at the University of Trømso, Norway, in conjunction with the SMARTool, the Russian Constructicon, and other tools. See Sokolova et al. (in press) for more information about this open educational resource (OER).
In this paper, we lay out the goals and origins of the *Visualizing Russian* project in Section 2, then turn to the evolution of the vocabulary categories used in the database and provide a brief description of the tools in Section 3. Due to space limitations, detailed descriptions of all tools are available on the project website.\(^2\) In Section 4, we discuss how the project provided a research opportunity for an undergraduate student and expanded the traditional academic path for a major in Slavic Languages and Literatures. In Section 5, we present some usage scenarios for the tools and demonstrate the benefit for learners, teachers, and linguistic researchers.

2. *Visualizing Russian*: Project goals

All of the tools in the *Visualizing Russian* project stem from the goal of finding ways to present the complexity of Russian morphology, lexicon, and grammar in a way that is visually appealing to learners. As teachers balance the use of authentic materials alongside materials created for language learners, they struggle to make authentic materials accessible to students while ensuring that pedagogical materials meet the targets of their curriculum. The *Visualizing Russian* tools assist with both goals. Using authentic materials—perhaps written or audiovisual materials created by and for native speakers for communicative purposes or raw sentence-length examples taken from language corpora—is quite difficult, particularly in languages such as Russian with vocabulary largely unrelated to English and with a high degree of morphological complexity. Several new tools and repositories have begun to address such difficulties in addition to *Visualizing Russian*. The SMARTool\(^3\) based on Janda and Tyers’s (2021) work on word-form frequency utilizes corpus-based examples that mitigate the difficulties learners face when dealing with the lexical and syntactic complexities of sentences created for native speakers in natural contexts. The Russian Constructicon\(^4\) includes constructions and examples for partially idiomatic constructions that are otherwise

\(^2\) The curious reader may try out any of the tools described here at https://visualizingrussian.fas.harvard.edu. Since the tools rely heavily on color, visualization, and dynamic user interaction, the tools are best experienced hands-on at the site with texts and words that individual users are most interested in analyzing.

\(^3\) See the SMARTool at https://smartool.github.io/smartool-rus-eng/.

\(^4\) See the Russian Constructicon at https://constructicon.github.io/russian/.
Visualizing Russian: Illuminating Corpora, Conjugations, and Classrooms
Clancy, Lee

difficult for learners to analyze or acquire. The CoCoCo\textsuperscript{5} tool provides corpus-based information on collocations. The Textometr/Текстометр\textsuperscript{6} tool analyzes texts and correlates the vocabulary content with the standardized lists of “lexical minimums” compiled for the Test of Russian as a Foreign Language (TORFL)/Тест по русскому языку как иностранному (ТРКИ) and thereby rates a given text on the Common European Framework of Reference for Languages (CEFR) and TORFL scales. Resources such as these open doors for the use of quantitative tools and data sources in the language classroom and in the creation of teaching materials. Researchers in second language acquisition and foreign language pedagogy can continue to measure the efficacy of such tools as they determine adequate percentages of known and unknown lexical items for effective reading in extensive reading and instructional contexts. Meanwhile, teachers can elaborate on the uses of the tools in the classroom and in the creation of learner materials.

The backbone behind most tools on the Visualizing Russian website is the project database, which Clancy began compiling around 2009 by hand-entering vocabulary items from various textbooks used to teach Russian (Live from Russia [2008–2009], Making Progress in Russian [1997], Leaping into Russian/С места в карьер, Начало [1995]). The lexical items in that database were later combined with static lexeme and frequency information from Sharoff’s (2008) frequency lists based on the RNC (lemma frequency list, form frequency list), and the database was also expanded to include more than 30,000 lexemes\textsuperscript{7} with forms obtained in 2017 from the Russian version of Wiktionary (https://ru.wiktionary.org), along with subsequent manual entry. Our database was initially informed by the lexical selection and frequency information contained in these sources but at present is an amalgam of many sources of information about the basic lexical and morphological facts of Russian, and we have incorporated frequency information from the RNC itself. The current database features over 300 fields for nearly 33,000 entries of lemmas, multiword expressions (e.g., [, потому что] [because], [после того, как] [after], [как раз] [just]), and highly frequent collocations (e.g., день рождения [birthday], домашнее

\textsuperscript{5}See the CoCoCo tool at https://cococo.cosyco.ru.
\textsuperscript{6}See the Textometr/Текстометр tool at https://textometr.ru.
\textsuperscript{7}Before the publication of the dictionary (Sharoff et al., 2013), Sharoff had made various frequency lists based on the RNC available on his website (http://corpus.leeds.ac.uk/serge/frqlist/), but the exact files no longer seem to be available.
задание [homework]), along with inflected word forms and lexeme and form frequency information.

As the database was becoming increasingly comprehensive for Russian, Clancy turned to officially launching the Visualizing Russian project in 2014 with the goal of making the complexity of the Russian language and this vast assortment of quantitative information available to students, teachers, and researchers in a visually compelling, relevant, and useful format. As the project began to come together in 2013–2014, Clancy was motivated by work in big data visualization, including new implementations of HTML5, CSS, and JavaScript, particularly the D3.js library, to turn these powerful visualization tools to the analysis of language. Even a decade later, a deep-dive New York Times article on the effects of an avalanche, “Snow Fall,” 8 remains an impressive and technologically inspiring piece that exemplifies this new kind of webpage experience. Information and plentiful examples of possibilities can be found at Bostock’s site for D3.9

Inspiring language applications include online interlinear readers10 and dictionaries11 as well as text analysis and concordancing tools such as Anthony’s AntWordProfiler12 and the work and play of “internet polyglot entrepreneurs,” particularly Kaufmann’s LingQ.13 LingQ has powerful features for keeping track of individual learners’ known words and progress in vocabulary acquisition, especially for learners attracted to Krashen’s comprehensible input approach to language acquisition (see Krashen [2003, pp. 1–14] for a summary of the basic tenets of this “just listen and read” approach). However, LingQ’s treatment of “words” as word forms rather than as lemmas is a nonstarter for morphologically complex languages, in which a learner would need to indicate manually that they knew a term such as Эбола [Ebola] and in which related case forms such as Эболы, Эболи would not be considered parts of the same word when occurring later in that text or in a subsequent text. Given that Russian verbs alone conservatively present 33 word forms (not counting all the case forms of the participles)

10 See https://interlinearbooks.com.
12 See http://www.laurenceanthony.net/software/antwordprofiler/.
13 See https://www.lingq.com/.
when considered as aspectual pairs, compared to the typical English
verb with 4–5 word forms, this is clearly not the most effective approach
for measuring words known in a grammatically and morphologically
complex language. Visible Vocabulary, described in more detail in the
subsequent sections, currently takes a levels approach to evaluating
textual difficulty and learner-appropriateness, but it is also possible
to set vocabulary knowledge and learning goals at an individual level,
which resources such as LingQ and Livingston’s Hedera project14 already
excel at. Visible Vocabulary and Hedera have a collaborative relationship
among digital humanities projects at Harvard, and in future iterations,
we hope to compile targeted vocabulary for levels as well as a measure
of individual user achievement and learning goals.

3. Visualizing Russian: Description of the suite of tools
Visualizing Russian has developed gradually with iterative improvements
and refinements to the basic tools, expansions of the database, and the
addition of new tools and features as new ideas have come to us. The
ability to break the project down into smaller parts with shorter-term
goals has made the project particularly well suited for including an
undergraduate researcher on the team and has also helped the project
weather the departure of multiple professional programmers who
worked on the project for varying periods of time. Our general process
is to propose an idea for a visualization or an analytical feature, either
as a refinement to an existing tool or as an idea for a new tool. Then,
we experiment with various statistical methods, ideas for visualization,
and new sources of linguistic data. As a new tool comes online, we work
through stages of development before deploying the tools into service on
the website’s public page.

3.1. Vocabulary levels for learners of Russian
As we develop the project, we keep the needs of learners, teachers, and
researchers in mind and hope that all tools will benefit each of these

14 Hedera (Livingston, 2022) sees itself more as a compendium of curated texts appropriate
for beginning and intermediate language students (focusing mostly on learners of Latin)
and as a tool for keeping track of learned and targeted vocabulary items. For more
information, see https://atg.fas.harvard.edu/hedera and https://hederaproject.org. Visible
Vocabulary showcases frequency-based and pedagogically motivated vocabulary goals
and more easily allows for the incorporation of authentic materials of variable linguistic
and lexical difficulty.
users. The central tool in *Visualizing Russian* is the text parser, *Visible Vocabulary*. This is the “Swiss Army knife” of the website, and we will likely incorporate features that are developed in the subtools into the text parser over time.

*Visible Vocabulary* began as a means of identifying the relative difficulty of texts and their appropriateness for use in courses for students at various levels of Russian proficiency. Simultaneously, Clancy had begun work on the *Foundations of Russian* textbook project, beginning with the intermediate level of Russian but planning all along to also complete the elementary and advanced levels. *Foundations of Russian* is intended to be informed by frequency of vocabulary and by language use as exhibited in corpora and authentic materials.

With these goals in mind, Clancy worked to identify the critical vocabulary items for university students of Russian. The *Visible Vocabulary* tool currently utilizes four levels: Core (green words), Foundations (blue words), Expansions (purple words), and Specializations (orange words) (Table 1). Words that are absent from the database remain unanalyzed (black), and thus an implicit fifth level emerges from among these unanalyzed words, mostly proper nouns, neologisms, slang, more recent borrowings, exceedingly rare words, the occasional misspelling, or a lexeme for which the word forms are missing in the database.

<table>
<thead>
<tr>
<th>Table 1. Vocabulary Levels in the Visible Vocabulary Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core (green)</strong></td>
</tr>
<tr>
<td><strong>Foundations (blue)</strong></td>
</tr>
<tr>
<td><strong>Expansions (purple)</strong></td>
</tr>
<tr>
<td><strong>Specializations (orange)</strong></td>
</tr>
<tr>
<td><strong>Unanalyzed words (black)</strong></td>
</tr>
</tbody>
</table>
The Core and Foundations levels together represent roughly 4,000 of the most frequent and communicatively necessary words in the Russian language. Frequency and ranks have been obtained from the Sharoff (2008) frequency lists and directly from the RNC. Word inclusion has been informed by frequency, but pedagogical and communicative concerns have motivated our choices as well. Core vocabulary accounts for roughly 1,500 lexical items (currently reflected as 1,879 entries in the database including aspectual pairs and adjective/adverb forms together as a single lexeme). We do admit items into the Core level that do not appear in the first 1,500 words, when frequency alone is considered, to include aspectual counterparts, adjectives and their adverbs, and items in necessary domains for language learners (e.g., various foods and drinks, months, days of the week, etc.). If frequency were the only concern, various domains would regularly have gaps, and related words would not be learned for quite some time.  

With the Core group consisting of the most frequent and necessary words and the Foundations level comprising a relatively genre-free belt of words needed in all domains, the remaining two levels are almost purely determined by frequency and comprise the remainder of lexical items in the Russian language as represented in the database. The Expansions level (purple words) includes the next most frequent 15,600 words in the language. The Specializations level (orange words) includes the remaining 12,500 words in the database. Sharoff’s (2008) lists included a lemma-frequency list and a form-frequency list. If we look at the levels of form frequency according to the lemmas they belong to, an interesting pattern emerges reflecting the character of these four levels. Figure 1 considers the total number of lemmas per level to which ranges of word forms belong in 6,000-word belts observed throughout 48,000 word forms.

The following characteristics emerge for each level: Core (green) words show Zipfian descent (Zipf, 1935) among the 1,500 most frequent lemmas in the language, yet the continuing solid line shows that even among the most frequent lemmas, not all forms in declension and conjugation are highly frequent, and some less-frequent lemmas are

15 Days of the week appear from ranks 1,547 (среда [Wednesday]) to 8,834 (вторник [Tuesday]). Nationality terms even for a single country/group are broadly scattered in frequency, for example, words related to France/French appear from 1,490 to 17,419 yet are included in any beginner’s textbook.
included among Core (green) words for pedagogical purposes. The Foundations (blue) words represent a relatively stable descent from 25% to 15% of the forms in their belts. This is characteristic of the need for these words as a foundation for vocabulary as these items appear across genres and in any and all domains. The Expansions (purple) words represent the new words a speaker steadily encounters in ongoing life experience with the language. Specializations (orange) words (and unanalyzed forms missing from the database) are reliably rare words. These major levels ripple through the other tools in the Visualizing Russian suite of tools described in the sections that follow. Due to space limitations, we describe only some of the major tools and their features. Full details are available for all tools on the project website.

Figure 1. Form frequency and level characteristics (6,000-word belts).

3.2. Visible Vocabulary tool
The Visible Vocabulary tool (Figure 2) visualizes the relative difficulty of a Russian text based on Core, Foundations, Expansions, and Specializations vocabulary. Words are colorized according to their level, and summary vocabulary-level statistics are shown for the text as a whole, including a bar chart and pie graph showing the percentage by level in the text. These measures provide the user with an understanding of the text’s general difficulty, whether for the purpose of instruction or individual reading. Given the importance of building
Core and Foundations for beginning and intermediate students, these two levels can be combined into one segment in the charts for a better estimate of what an intermediate or early advanced student should find approachable in the text. The colored text below the summary statistics provides a visual map of familiar and potentially unfamiliar words for the reader, a helpful feature for both language learners (which words to focus on learning) and teachers (which words to gloss in materials). Beginning and intermediate students know their attention is best served by mastering green and blue words, while more advanced students can direct their attention to acquiring the vocabulary of specific domains and greater nuance represented by purple words.

Figure 2. Word-level breakdown and coloring on the first part of Anton Chekhov’s “Lady with the Lapdog” in the Visible Vocabulary tool.

Clicking on any word in the text brings up a word information panel to the side of the text, providing additional grammatical information. Clicking on a verb shows grammatical information about the verb form as well as a gauge visualization showing the verb’s relative aspect ratio of imperfective to perfective. Stress patterns are shown for nouns, and words at the Core and Foundations levels have English
glosses. The user can copy-paste any text into the tool, analyze the text, then copy-paste the colorized text into a word processor for printing out and further editing, or the user can read and interact directly with the text in the tool as an online reader. We have solved several problems with processing text, including parsing hyphenated words (кто-то [someone]) and multiword expressions (“, потому что” [because]). The tool scales up well: hundreds of pages of text can be processed at a time; a novel the size of The Brothers Karamazov can be analyzed in four parts.

3.3. Verb Histogram tool

The Verb Histogram tool is designed to provide information about verb constructions to help learners better understand the prepositions and cases used with verbs and thereby increase their command of the language. The tool uses data from the disambiguated subcorpus of the RNC to reveal verb constructions, word orderings, and construction frequencies throughout the corpus (Figure 3).

Figure 3. Graphical breakdown of обращать [turn, direct] across case constructions in the Verb Histogram tool. “-/+1” indicates that the construction begins one word before or after the target verb.
The tool displays a histogram of cases and prepositions plus cases surrounding every verb in the corpus. For example, if the user looks up обра́щать [turn, direct], the resulting histogram reveals that some common constructions with обра́щать are на + accusative, bare accusative, and bare genitive. The summary construction at the top of the tool immediately tells us that the most common string of constructions is [NOM + обра́щать + ACC + (на + ACC)]. The related-words section of the tool offers additional information on words that frequently appear with the verb in question. The data from the corpus reveals these suggested constructions as well as the common collocation обра́щать внима́ние [pay attention].

The histogram is interactive and includes corpus sentences representing each construction component. Perhaps the user is curious about the use of обра́щать with the genitive case. Clicking on the GEN label in the histogram legend reveals 11 example sentences with GEN in the position immediately following the verb, all of which involve не обра́щать внима́ние with GEN of negation. The tool makes case usage visible and accessible and conveys verb constructions in an accessible, standardized way.

3.4. Case Distribution tool
The Case Distribution tool generates radar charts for a noun’s case usage profile based on the RNC subcorpus. The chart shows how often a word occurs in its different case forms. Radar charts are a good way of visualizing relative frequency data across categorical items. In the tool’s basic view, the categories are always the six Russian cases (nominative, accusative, genitive, locative, dative, and instrumental16). In the detailed view, case usage is broken down to include bare case use and preposition + case. For example, “genitive” might break down into “для + genitive,” “от + genitive,” and so on in the detailed graph. Clicking on the categories around the perimeter of the radar chart displays the actual sentences from the RNC subcorpus.

For example, a user may input a noun like квартира to get a better idea of its case breakdown. In the basic chart (Figure 4), we see forms of квартира singular (yellow) appearing most frequently in the accusative case (96 instances) and locative case (93). The most frequent case among

16 This ordering of the cases most efficiently deals with overlaps in case endings (syncretisms), and the hexagonal format of the radar chart allows for easy comparison across cases (direct/oblique, etc.).
the plural forms (blue) is the genitive case (19). The tool tells us the total counts: there are 300 singular and 57 plural instances of the word квартира in the RNC subcorpus.

![Figure 4. Basic chart view for noun cases (квартира).](image)

From the detailed view (Figure 5), the most common preposition + case collocation is в квартире, which makes up 86/300 of the singular examples. The user can see that в [in] is used much more frequently than на [on, at], revealing information about preposition usage. Clicking on the label provides the user with example sentences.

![Figure 5. Detailed chart view for noun cases (квартира).](image)
Words that are better represented in the abridged corpus present more interesting and, we believe, reliable visualizations than do words with relatively few occurrences. In the future, we would like to determine the occurrence thresholds necessary for a stable, illustrative case graph. Our attempts to create truly representative case graphs from raw corpus data have not yet proven fruitful due to case syncretism. We are also developing analogous tools for verbs and adjectives, but the greater number of morphological forms in verbs make the visualization more cumbersome.

3.5. Additional and future tools
Space precludes detailed discussion of four additional tools: Mini-Story Creator, Quick Lemma, Wordburst, and Similarity, but we encourage readers to go to the website to try out all of these tools and find further information.

The Mini-Story Creator allows for the creation of a new text or analysis of an existing text for lemma frequency and word form counts in comparison with a set of target vocabulary items. Students can create texts using target vocabulary, and teachers can verify the presence of target vocabulary in their materials.

The Quick Lemma tool provides information about the frequency of all of the various forms of a particular Russian word or about the frequency of a set of lemmas as represented in the project database, the RNC, and the Google Ngram Viewer17 (Figure 6).

Figure 6. Near-synonym comparison in the Quick Lemma tool (manufacturing locations: завод, фабрика, комбинат).

The Quick Lemma tool provides information about the frequency of all of the various forms of a particular Russian word or about the frequency of a set of lemmas as represented in the project database, the RNC, and the Google Ngram Viewer17 (Figure 6).

---

17 See Michel et al. (2010) and https://books.google.com/ngrams.
The *Wordburst: Word Formation* tool breaks down words into prefixes, roots, and suffixes and provides a dynamic graph of words related by root.

![Figure 7: Word Formation tool for root lay/put.](image)

![Figure 8. PCA graph for 40 words related to ocean [ocean].](image)

For example, the user may be interested in words with the root associated with lay/put (Figure 7). Items closer to the center have higher frequency than items at the periphery, and colors represent vocabulary levels.
The *Similarity* tool incorporates techniques from natural language processing of vector spaces to explore sets of related words with semantic and contextual connections (see Pennington, et al. 2014). Figure 8 shows the principal component analysis (PCA) graph for words related to *океан* [ocean].

We have far from exhausted the work we would like to do with the database itself and are only beginning to implement vector embeddings and other big data and deep learning methods.

4. *Visualizing Russian and undergraduate research opportunities*

Research opportunities for undergraduate students are heavily weighted toward science and engineering in general and are reflected in the physical space and laboratory resources on a typical university campus. Support for undergraduate research in the humanities, even with regard to digital humanities, is of relatively recent provenance. However, creative thinking can extend the notion of humanities research and create opportunities for undergraduates in projects that benefit humanistic fields of study.

The undergraduate degree in Slavic Languages and Literature, or the “Typical Russian Major,” is a program of study that generally includes coursework in language, literature, and culture. The coursework often includes extensive language study, exposure to the Slavic literary canon in English translation with some readings in the target language, and discussions of Slavic cultures in linguistic and literary contexts. In this framework, the research opportunities for undergraduates working toward a Slavic degree may involve information gathering and analysis of specific authors, works, literary movements, literary theories, historical moments, and linguistic developments, or some combination of these topics.

*Visualizing Russian* extends beyond the bounds of these “normal” Slavic research opportunities for undergraduates or even the typical pursuits of graduate students. At its core, our project is an interactive interface intended for any level of scholar, from undergraduates learning Russian to linguistic researchers interested in testing their hypotheses on corpus-based visualizations of language data. In comparison to an...

---

18 These tools are powered by the embeddings of the Natasha/Navec project (Kukushkin, 2022).
article or book that takes on a static end form at publication, our work is continually modified and improved based on evaluation and feedback. Rather than solely benefiting the research community, the project is also available to nonresearchers, including students and teachers. Another difference between Visualizing Russian and other humanities projects is the explicitly digital nature of our project. The end product of our work is currently available on the web at no charge and takes advantage of many of the affordances of the digital display, such as user interactivity, dynamic figures, and vast database storage. The integration of “big” data into the website allows for massive flexibility and customizability for users to find information on the words and sentences they are interested in.

Visualizing Russian is unique in general as a digital humanities research project, but it has been particularly valuable because it coincidentally aligns well with the specific interests of the undergraduate researcher and coauthor of this paper, Paige Lee. Paige is pursuing a joint concentration (Harvard-speak for a double major) in Computer Science and Slavic Languages and Literatures. Having already matriculated with advanced knowledge of Russian, Paige is interested in the union of these fields, and she hoped to conduct research on a topic that combined Slavic language and culture with computational tools and frameworks. The Visualizing Russian project combined these interests, allowed her to build on existing language knowledge and computational skills, and additionally introduced her to new fields and interests, such as digital pedagogical tools and corpus linguistics.

This “out-of-classroom” research experience has augmented Paige’s professional and personal development. She has learned new technical skills in web development and linguistic data analysis from working on a live, dynamic website in collaboration with Arthur Barrett, a professional academic technologist at Harvard. Her work on the project sparked her interest in the ever-expanding field of natural language processing, in which computational tools are used to analyze human language. The project also has significantly refined her skills in designing and implementing web-based data visualizations.

From the faculty perspective, individual professors may be hesitant to work with undergraduate researchers because of the temporary nature of their availability. Paige has already worked on Visualizing Russian for
two years, including highly productive work during a leave of absence during the pandemic, and still has a year not only to contribute to the project but also to allow experience with the project to form her plans for a senior thesis in Slavic and Computer Science. The modular nature of the project can also add value to the contributions even of short-term participants, as they can work on specific small-scale tools that are integrated as components of the project as a whole. Paige’s longevity with the project over the course of her undergraduate years as well as the “suite of tools” nature of *Visualizing Russian* contribute to forming a productive research opportunity.

Finally, this has been Paige’s first experience working on any kind of project with a team of academic researchers. Exposure to the academic research and publishing process is invaluable for any student interested in pursuing higher education or academic careers. Meanwhile, Paige learned how to develop a website in a team setting, a task requiring a specific set of skills in version control and web development. The professional diversity of the team (a Slavic linguist, a technology professional and his team, and an undergraduate student) allowed for cross-disciplinary insights, contributing to a well-rounded final product. As Paige is a current language student, her perspective mirrored that of a potential end user of the tool, which is especially helpful during the feedback process. The benefits of this collaboration will be assets for Paige’s personal and professional growth going forward.

5. Usage scenarios
The *Visualizing Russian* tools have already been used in numerous teaching and learning contexts. The *Mini-Story Creator* allows creators of teaching materials to verify that their dialogues, texts, and examples match with target vocabulary and ensure that target vocabulary is being recycled in those materials. *Visible Vocabulary* allows teachers to select the most appropriate texts by level for their students and allows independent learners to focus their study on more frequent vocabulary (green and blue words) at earlier levels, while more advanced students can focus on developing genre-specific vocabulary and professional jargon in texts by paying more attention to less-frequent vocabulary (purple and orange words). The *Verb Histogram* tool allows for quick classroom demonstration of verbal case governance along with relevant
examples from the RNC. The tools are thus equally applicable to authentic materials and pedagogical materials, both of which have their place in teaching and learning Russian as an L2.  

5.1. Learning materials and textbook development

Authentic texts and pedagogical texts created by native-speaking teachers often unnecessarily complicate matters for students. While these more complicated texts have a certain native flair, they can distract and confuse learners at a time when their attention would be better focused on obtaining a broad general vocabulary across a number of semantic domains. In the preparation of *Foundations of Russian* (Clancy et al., in press), such examples have been spotlighted using the Visual Vocabulary tool:

«Каждое слово имеет значение, — говорит она. — Убеждать — значит правильно подбирать выражения.
[“Each word is significant,” she says. “Persuading is a matter of correctly choosing expressions.”]

Сстраю дома из кубиков маленький Боря скучал, зевал, а потом и совсем уснул.
[While building a house out of blocks, little Borya got bored, yawned, and then totally fell asleep.]

Вы не подскажете, как добраться до вокзала?
[Could you tell me how to get to the train station?]

Какими масками можно напугать людей на Хеллоуин?
[What masks can you use to scare people with on Halloween?]

For an intermediate student of Russian, encountering rarer lexical items (purple and orange words) in their textbook is not only unnecessary and confusing but also misses the opportunity to showcase and repeat the target vocabulary for the level (Table 2).

---

19 However, authentic corpus examples are often long, complicated, and difficult to extract from the broader narrative context they were originally used in. As one reviewer of this article noted, we are still very much in need of short, dialogue-style utterances, even if we can maintain a corpus-based but simplified approach.
Table 2: Verbs Found in Some Native-Speaker Texts That Could Be Replaced with Target Vocabulary

<table>
<thead>
<tr>
<th>Form used by native speaker...</th>
<th>Target vocabulary for intermediate students...</th>
</tr>
</thead>
<tbody>
<tr>
<td>подбирать</td>
<td>подобрать  (orange</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>выбрать</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>заходить</td>
<td>зайти  (blue</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>входить</td>
<td>войти  (green</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>подсказать</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>сказать</td>
<td>(</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>заваривать</td>
<td>заварить чай  (purple</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>варить</td>
<td>сварить  (blue</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>нануть</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>путать</td>
<td>испугать  (blue</td>
</tr>
</tbody>
</table>

Frequency plays into these usages in various ways. While the specific construction заваривать | заварить чай is a more frequent collocation than is варить | сварить чай, the latter verbal pair is overall more frequently used\textsuperscript{20} and thus more urgent for students to know so that they are aware of cooking by boiling water as a general concept among cooking verbs in Russian before they add a more nuanced verb like заваривать | заварить to their vocabularies. Likewise, in the preceding examples, a basic level of expression is established in the green and blue words that needs to be acquired by learners before they turn their attention to near-synonyms and more specialized items.

5.2. Guided acquisition of lexical items

Students studying Russian as an L2 in the classroom with a teacher have a guide on hand for their learning, but when they study on their own or read independently, they lack direction about which lexical items are statistically more frequent. For instance, a learner reading a selected

\textsuperscript{20} Forms of заваривать | заварить with чай occur roughly 3.7 times as often in the main corpus of the RNC as варить | сварить with чай, but overall варить | сварить [12,664|4,678 = 17,342] are more frequent lemmas than are заваривать | заварить [770|1654 = 2424] in terms of occurrences in the RNC. An anonymous reviewer pointed out that заваривать | заварить чай has especially taken off since around 1980. However, “coffee” tells another story, in which the verbs are reversed with варить | сварить, with кофе around 5.1 times more frequent as заваривать | заварить.
passage from Gogol’s *Dead Souls* might encounter a number of items from 19th-century realia (*приданое* [dowry, *purple*], *помещик* [landowner, *purple*], *кучер* [coachman, *purple*], *форейтор* [postilion, *purple*], *экипаж* [carriage, *purple*]); diminutive or nonstandard forms (*тысячонок* [a thousand, unanalyzed]), *внутренно* [inwardly, unanalyzed]); and high-level vocabulary (*досадовать* [be vexed by, *purple*], *разведать* [reconnoiter, *purple*], *лакомый* [tasty, *orange*]). Reading with guidance from *Visible Vocabulary*, the independent student can gauge the relative frequency of words and also make better judgments about what words are likely to be archaic or less-than-ideal candidates for inclusion in vocabulary lists. This is the sort of advice an instructor might give. A beginner-intermediate student would know to focus in general on green and blue words and to leave purple words for later acquisition, whereas even an advanced student would have a better sense of what purple words would be good to pay attention to while casting aside the least-frequent orange and unparsed words except for momentarily understanding their use in the passage at hand. For teachers preparing handouts for their students, it instantly becomes obvious which words are most in need of being glossed to make reading such a passage more approachable.

### 5.3. Case governance in the classroom

The *Verb Histograms* tool can be used in the classroom to demonstrate what cases and prepositions Russian verbs tend to be used with and what the main constructions are for each verb. Teachers and textbooks regularly explain that *помогать*/*помочь* [help] is used with a dative receiver of the assistance or that “answer a question” is expressed by *отвечать*/*ответить* plus the preposition *на* with the accusative case. With the *Verb Histogram* tool, these constructions can be shown along with broader case and preposition usage and with copious examples from a corpus. For the preceding verbs, the results of the tool show that “help” is also used with the infinitive and that an “answer” can be directed to a dative receiver or to a question (*на вопрос*) or that one can “answer for” someone or something (*за* + accusative). However, potential drawbacks here include poor construction representation due to lack of data for some verbs and the fact that authentic corpus examples are often difficult to understand and may thus fail to exhibit the essential point of the basic case governance for our students.
6. Conclusion
A large-scale project like *Visualizing Russian* would not have been possible without time, dedication, and financial support, but most essentially, it would not have been possible without collaboration. Given the intersection of linguistic analysis, pedagogical expertise, technical and programming skills, and visual design, it is unlikely that a single researcher could make much progress on their own. It is simply too difficult to keep one’s feet in all of these different areas and to keep up with changes in programming languages, new tools, and contemporary web design and data visualization. However, a large-scale, long-running project such as this proved to be an excellent project for collaboration among individuals from a variety of backgrounds, gaining attention and support from initiatives in digital humanities and serving as an opportunity not only for undergraduate research but also for thinking about ways we can expand the traditional understanding of a Slavic Languages and Literatures undergraduate major.

References


**Past and present Russian language textbooks mentioned**


