Towards Intelligent Correction of Collocational Errors in Russian L2 Academic Texts in the CAT&kittens Writing Support Platform

ALEKSANDR KLILOV, OLESYA KISSELEV, MIKHAIL KOPOTEV

1. Introduction
The study of academic language is driven to a large extent by the need to teach second language (L2) writers about established practices and patterns found across different genres and registers common in academic written discourse. Over the span of the past few decades, the area of academic language research has been hugely influenced by two interconnected digital approaches: computer-assisted language learning (CALL) and computational linguistics, including corpus linguistics approaches and tools.

CALL has enjoyed a long history in the field of language education. Following the cautious success of the first computer-assisted instruction systems, such as PLATO (Van Meer 2003), scores of CALL services have been created to address various instructional needs of language learners. Today, multiple language-learning digital platforms, such as Duolingo, Rosetta Stone, and others (Heil et al. 2016), provide language learners with language activities, flashcards, and communicative tasks; the majority of these platforms target the most commonly taught languages, such as English and Spanish. Fewer CALL tools and platforms have been developed for the study of less-resourced and less commonly taught languages, with some notable exceptions, such as Revita (Katinskaia, Nouri, and Yangarber 2018), a platform for self-study of such endangered languages as Erzya, Komi-Zyrian, and North Saami, as well as some more commonly studied languages, such as Swedish, German, and Russian.

Other CALL resources are designed for native speakers and advanced L2 learners; these tools check texts for possible grammatical and stylistic deviations and are envisioned as a writing scaffold or an editorial assistant. One of the most widely used such resource is Grammarly (Nova

---

1 The authors are grateful for the assistance of all students who worked on various stages of this project. Our special thanks to Yves Wienecke from the Portland State University. We also very much appreciate suggestions made by the reviewers of the manuscript and the editors of the Special Issue.
Towards Intelligent Correction of Collocational Errors in Russian L2 Academic Texts
Klimov, Kisselev, Kopotev

2018), which allows English language writers to check spelling, grammar, and punctuation errors and to evaluate their vocabulary usage (specifically, overuse of particular lexical items, hedging devices, passive voice, etc.) and writing style.

Corpus linguistics methods have similarly made great strides in the area of instructional language resources (Crossley et al. 2017, 51). Corpus-based tools, for instance, have immensely contributed to the analysis of English academic discourse and have added to the repertoire of resources for teaching English for Academic Purposes in the classroom and beyond (Ackermann and Chen 2013; Biber, Conrad, and Cortes 2004; Durrant and Mathews-Aydınlı 2011; Gray and Biber 2013).

Despite the fact that corpus-based and computational approaches to the study of the Russian language have enjoyed brisk development in the past decade, practical CALL and corpus-based applications for Russian are few. Some promising work has begun to address this lack of resources. Visualizing Russian (Clancy et al. 2019), for example, provides an analysis and visualization of the relative difficulty of a Russian text and level of difficulty of words in the text, and the above-mentioned Revita platform (Katinskaia, Nouri, and Yangarber 2018) provides opportunities for learners to practice languages, including Russian. Another online resource, Collocations, Colligations, and Corpora (Kopotev et al. 2015), allows teachers and learners to actively examine phraseological aspects of Russian, providing information of the collocational and colligational behaviors of searched items.

The project that the current paper reports on, titled CAT&kittens and available online at catandkittens.linghub.ru, aims to contribute to the repertoire of resources, by creating a large systematically collected Corpus of Academic Texts (CAT) and a platform for the evaluation of features of non-native texts (a.k.a. “kittens”) against the features of academic Russian attested in the corpus.

The rest of this paper is structured as follows: in Section 2, we provide a brief overview of previous research on academic Russian and present a case for the necessity of corpus-based research in this area. Section 3 describes the compilation of the CAT and the future writing support platform, discussing some pertinent methodological considerations. Here, we also provide a brief introduction to the tools and services developed based on the analyses of the corpus. We then introduce the algorithm we developed for the detection and correction of miscollocations in L2 academic texts. We conclude the article by discussing the potential of CALL and corpus-based tools.
2. Russian for academic purposes in the corpus-based era

The study of the Russian academic genre has flourished in the 1960s and 1970s. The majority of studies concerning Russian academic writing belong to two major traditions: the prescriptivist tradition and the descriptivist tradition (often pre-corpus). The prescriptivists are represented by several key studies conducted in the late 1970s through the early 1990s (cf. Vasilyeva 1976; Mitrofanova 1985; Kozhina 1993) and implement introspective methods of analysis. These works are based on limited sets of data, often inadequate for generalizations, with examples of usage often based on the author's own experience and intuition and not based on extensive data analyses. The problems associated with the prescriptivist approach persist in the most recent books on academic stylistics, for example, Golubeva and Maksimov (2020), which are based on outdated introspection and recommendations that do not reflect recent developments in contemporary academic Russian.

The descriptivist tradition in the study of the Russian academic genre, dating back to the 1960s, has always emphasized the use of larger amounts of data to provide generalizations regarding language use. A number of seminal studies, such as Piotrovsky (1971) and Piotrovsky, Bektayev, and Piotrovskaya (1977), were conducted using a corpus-analytic framework and investigated academic style from various perspectives: from analyzing word lengths to phraseology and vocabulary. The obvious limitation of early corpus-based research was the size of corpora that researchers had access to at the time; the biggest corpus of the era, Zasorina (1977), comprised only one million tokens. Early statistical approaches do not meet the standards of modern corpus linguistics. Most importantly, the observations gleaned from those early studies have reduced relevance today, given that the Russian academic style is changing rapidly as it is being increasingly influenced by the Western academic tradition. In light of these observations, there is an evident need for a new corpus-based exploration of the Russian academic genre.

It is important to mention that the field has made some strides in the direction of corpus-based exploration. For example, the Russian National Corpus (RNC; www.ruscorpora.ru), one of the largest text corpora of the Russian language, contains an academic genre sub-corpus (ca. 27.4 million tokens). The academic sub-corpus represents various text types and sub-genres, such as article annotations, dissertations, research articles, and educational texts. The academic sub-corpus also contains a large number of texts representing the popular science sub-genre. However, the
academic sub-corpus provides no filter to sort the different sub-genres, which results in a considerable obstacle in conducting research with these data. Another issue with the RNC academic data is that it is unbalanced in terms of the size of the academic disciplines represented in the corpus. For example, while the Philology sub-corpus consists of ~2.5 million tokens pulled from 412 documents of the diverse sub-genres and History has ~6.6 million tokens from 1,372 texts, Sociology is only made up of ~500 thousand tokens from 102 texts, thus making the balance weak—not only in terms of the number of tokens but also in terms of the balance of text types, consequently decreasing style diversity and representativeness. The RNC does not provide fine-grained classification: it does not allow searching only for Linguistics texts, for instance, instead including them in the Philology sub-corpus. Finally, and most importantly, the RNC is not available for downloading and additional processing, making its further utilization in applied research impossible.

Another digital Russian project that provides some access to academic Russian is the Russian academic sub-corpus RU-AC of the larger IntelliText project led by Sharoff, http://corpus.leeds.ac.uk/itweb. IntelliText allows its users to search by words, collocations, and grammatical tags; however, RU-AC is relatively small (five million tokens) and consists of mixed-quality student papers and thus cannot be considered representative.

The purpose of the CAT&kittens project is twofold: we hope to contribute to the study of the contemporary Russian academic language through the corpus-based study of various linguistic features and to create a writing support platform that provides a set of language-analyzing and editing services. We describe corpus construction and the tools based on it.

## 3. Corpus construction

### 3.1. Corpus description

The CAT is envisioned as a large corpus of academic Russian language spanning multiple disciplinary domains. It currently includes 3600 research articles sourced from high-impact peer-reviewed Russian academic journals; the articles represent six general disciplinary fields: Economics, Education and Psychology, Legal texts, Linguistics, History, and Sociology. To ensure that the corpus reflects the contemporary state of the Russian academic language, the corpus consists of articles published between
2010 and 2018. The current size of the corpus is 14,010,588 tokens, which is rather small given the linguistic nature of the task being attempted; a bigger corpus, CyberCAT, based on the largest Russian scientific online library CyberLeninka, https://cyberleninka.org, is now under construction. The numbers of texts and tokens per discipline in the CAT are presented in Table 1. Upon completion, the corpus will be available for online and offline use.

Table 1. Distribution of texts in CAT

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Texts</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economics</td>
<td>600</td>
<td>2,494,422</td>
</tr>
<tr>
<td>Education and Psychology</td>
<td>600</td>
<td>1,880,004</td>
</tr>
<tr>
<td>Legal texts</td>
<td>600</td>
<td>2,636,290</td>
</tr>
<tr>
<td>Linguistics</td>
<td>600</td>
<td>2,691,363</td>
</tr>
<tr>
<td>History</td>
<td>600</td>
<td>2,808,313</td>
</tr>
<tr>
<td>Sociology</td>
<td>600</td>
<td>1,500,196</td>
</tr>
<tr>
<td>Total</td>
<td>3600</td>
<td>14,010,588</td>
</tr>
</tbody>
</table>

3.2. Data preprocessing

There exist a number of options for encoding Cyrillic characters; in this project, we used the UTF-8 encoding since it is a de facto standard and is required by the UDpipe tagger (Straka and Straková 2017), which is used to annotate the data. UDpipe is a trainable pipeline that combines tokenizing, POS-tagging, lemmatization, and dependency parsing tools for CoNLL-U files (ibid.). Using this pipeline allows us to provide not only basic information such as frequencies of tokens but also more valuable information on grammatical characteristics of the tokens and lemmas as well as syntactic dependencies.

Since Russian is one of the academic lingua francas (Pavlenko 2006), it is reasonable to assume that some papers are authored by non-native speakers. However Russian leading academic journals require proofreading by a native speaker, which improves quality. At the algorithm level, we set a frequency threshold to filter out rare collocations (see Section 4.3.1).

---

2 Since Russian is one of the academic lingua francas (Pavlenko 2006), it is reasonable to assume that some papers are authored by non-native speakers. However Russian leading academic journals require proofreading by a native speaker, which improves quality. At the algorithm level, we set a frequency threshold to filter out rare collocations (see Section 4.3.1).
Towards Intelligent Correction of Collocational Errors in Russian L2 Academic Texts

Klimov, Kisselev, Kopotev

Figure 1. UDpipe markup sample

Before the application of UDpipe, all the texts entered in the corpus were preprocessed: all references, tables, figures, university and author names, and page numbers in the running heads were omitted; the texts were divided into separate sentences marked with periods; all punctuation marks defining the end of the sentence were transformed into periods; and other punctuation marks (i.e., commas, semicolons, etc.) were deleted. All numbers in the texts were replaced with the NUM token (where NUM stands for any number or figure used in the text\(^3\)). Only the main sections of the articles (introductions, literature reviews, methodology and results sections, and discussion and conclusion sections) were entered in the corpus; all tables, figures, equations, and references (in-text citations, as well as bibliographies) were deleted automatically using custom Python scripts available at https://github.com/kopotev/CATandkittens.

3.3. Corpus-based language assessment tools

CAT&kittens is a multifunctional platform under development with different features for learners and teachers to be available when completed. Apart from the collocation substitution algorithm that we present in the paper in greater detail, the platform is designed to also assess other textual aspects.

The first level of checking is a set of readability metrics, or an estimation of the text’s difficulty level. To assess the readability, we use the Flesch Reading Ease and the Flesch-Kincaid Grade level tests, both adapted for use with Russian academic texts (Solovyev et al. 2018). Average sentence length and type/token ratio (TTR) in the L2 text are also compared to the

\(^3\) In a small number of cases, the specific number constitutes an invariable part of a collocation; such collocations are lost through our preprocessing procedure. However, this substitution allows for significant reduction of noise in the data.
same parameters in the CAT to show how the learner’s text compares to the general level of difficulty accepted in the academic discourse.

Another set of metrics concerns lexical variation. The set of lexical analysis is closely connected to both collocational and readability measurement checks. It includes a check on the low-frequency items, including *hapax legomena*, a check on overuse/underuse of specific lexemes and terms that are unattested within the discipline domain; in addition to simply identifying deviations, the system suggests alternatives when available.

The third set of metrics is an academic-specific grammar check. Unlike available spell-checkers, this set of tools is focused on detecting deviations that are featured in academic writing—specifically those deviations produced by non-native speakers, such as genitive chains or prdrop. In this paper, we focus on one of these sets, the collocational issues in L2 academic writing, and we give a brief overview of methods and algorithms designed for detection and correction of miscollocations and review preliminary evaluation of these methods.

4. Collocations

4.1. Collocations in learner texts
The definition of the term *collocation* differs depending on the theoretical orientation; in traditional phraseology (Mel’cuk 1995; Men 2018), *collocation* is defined as a lexical unit based on the semantic relationship of the words it comprises. In the statistical approach, collocations are defined as a syntagmatic lexical relation between two or more words that co-occur with relatively high probability, based on statistical measures (Stubbs 2001, 2009; Evert 2008). In this article, we utilize the latter approach, identifying collocations statistically.

The ability to produce standard or native-like collocations—that is, multi-word expressions whose co-occurrence is greater than chance—is a notoriously difficult skill to acquire for L2 speakers of a language (Nesselhauf 2005; Römer 2010). Collocations specific to particular disciplinary fields may present a unique challenge for L2 academic writers, simply because the chances of being exposed to such collocations are small. The following examples of miscollocations (i.e., collocations non-attested in standard Russian) come from samples of L2 data; however, even native speakers of the language may exhibit gaps in collocational knowledge when producing texts outside of their immediate area of expertise or experience.
Towards Intelligent Correction of Collocational Errors in Russian L2 Academic Texts

Klimov, Kisselev, Kopotev

Example 1.

a. *используем точки зрения

use.Prs.1Pl point.Pl.Acc view.Gen.Sg

[we] use the points of view

Cf. Rus.

b. рассматриваем точки зрения

consider.Prs.1Pl point.Acc.Pl view.Gen.Sg

[we] consider the points of view

Example 2.

a. *исследования на сфере

research.Nom.Pl on.Pr sphere.Loc.Sg

research in the sphere

Cf. Rus.

b. исследования в сфере

research.Nom.Pl in.Pr sphere.Loc.Sg

research in the sphere

4.2. Automatic extraction of collocations in learner data

Collocational research has been at the forefront of corpus-based approaches to the study of the standard and learner language varieties. And while identification of standard collocations is methodologically well established, automatic detection of collocational errors is still developing.

The first approaches that were proposed following the innovative work on error detection by Shei and Pain (2000) were rule-based; predictably, these approaches returned highly accurate results but within a very limited environment in which the rules worked best (Eeg-Olofsson and Knutsson 2003). Later, statistically based algorithms became the mainstream in the field, either based on basic statistical techniques (Park et al. 2008) or based on more advanced machine learning algorithms (Brockett, Dolan, and Gamon 2006; Gamon et al. 2008; Wu, Chen, and Chang 2017), which led to increased accuracy in detecting miscollocations, although accuracy in correcting them still lagged behind. Many proposed algorithms have relied on nonspecified internet searches (Guo and Zhang 2007; Bolshakov and Gelbukh 2003), while some alternative approaches have utilized prefabricated error databases consisting of common collocational errors (Futagi et al. 2008). Other algorithms have tested the assumption that collocations in L2 are influenced by a native language by utilizing reverse translations to detect miscollocations (Chang et al. 2008).
An integrated approach to miscollocation detection and correction can be found in Ferraro et al. (2014), who take into account previous findings: specifically, they account for graphically similar collocates, calquing from students’ native languages, collocate association strength, and the context of miscollocations. Even this integration still does not allow reaching a reliable level of accuracy in collocational error detection. The authors conclude that “the potential of contextual features has not been fully explored as yet” (Ferraro et al. 2014, 17); this remains to be explored in further works.

One way to account for contextual features is to apply distributional semantics modeling in order to locate a word or a word combination in its semantic space. This approach has been tested in Kochmar (2016) and Rodríguez Fernández (2018), and the results are very promising.

4.3. Current approach to collocational errors extraction and correction

In order to detect and correct instances of miscollocations in our data (see Section 4.1. for examples), we rely on methods that attest to the strength of collocations in academic texts and methods of distributional semantics to look up the closest semantic neighbors and suggest alternative standard collocations.

A significant advantage in this work is that the academic genre is inherently more clichéd and contains more fixed collocations than many other genres. Example (3) shows the difference between general and academic texts.

Example 3.

a. Автор *[расследует вопрос] об автор.Ном.Сг investigate.Prs.1Sg question.Acc.Sg about.Pr *[установлении закона] establishment.Loc.Sg law.Gen.Sg

The author investigates the question of the establishment of the law

Cf. Rus.

b. Автор исследует вопрос о author.Ном.Сг study.Prs.3Sg question.Acc.Sg about.Pr принятии закона enactment.Loc.Sg law.Gen.Sg

The author studies the question of the enactment of the law
Towards Intelligent Correction of Collocational Errors in Russian L2 Academic Texts

Klimov, Kisselev, Kopotev

In example (3) two nonstandard expressions are used in the same sentence: *расследует вопрос ‘investigates the question’ and *установлении закона ‘establishment of law.’ The first expression is unattested both in the CAT and the RNC. The second expression can be used in science texts with the meaning ‘to discover a [physics] law’; however, this expression is unattested in the corpus of humanities disciplines (to which the learner text belongs as well). Thus, the platform identifies such a phrase as a candidate for correction and attempts to find a corresponding replacement in our academic domains.

Identifying a miscollocation that is not present in the reference corpus is not a challenging task computationally. Its replacement with a semantically similar collocation, however, is challenging. There are two general methods for substitution of unattested collocations in the learner text: (1) straightforward frequency-based substitution and (2) substitution based on distributional semantics modeling.

4.3.1. Finding an unattested collocation
Each text can be presented as a set of words, that is, a syntagmatic n-gram or a linear string, where the number of words is equal to \( n \). Thus, each text can be presented as a long list of 2-grams (i.e., strings of two words), then 3-grams, 4-grams, etc. But the resulting strings are not collocations, neither in the statistical sense nor in the traditional phraseological sense, since according to our definition, a collocation should be statistically significant.

A simple way to identify a recurrent string of words is to count frequencies of every n-gram and to consider the most frequent n-grams that surpass some threshold—for example, five instances in the reference corpus—as legitimate collocations. The task of collocation replacement in this case is straightforward and follows these steps: first, splitting a learner’s text into 2-, 3-, 4-grams, etc., and then, comparing the resulting n-grams to the list of most frequent n-grams in a reference corpus. Any n-gram found in the text and not present in the corpus is considered unattested and a candidate for replacement. This method does not always work as intended: simple absolute frequencies elicit many function-word-based collocations, such as *так и ‘as well as’ and *как в ‘as in’ that technically are not collocations but rather discontinuous units; however, it tends to push those rare collocations that are content-word-based to the very bottom of the extracted strings.

To make sure that we extract meaningful and pedagogically useful collocations, probabilistic measures need to be applied. For collocation
extraction, several statistical metrics are generally used (Pivovarova, Kormacheva, and Kopotev 2017):

(1) **T-score** is a widely used test that looks at the mean and variance of the sample and then compares this observed mean with a predicted mean, scaled by the variance. It allows one to understand whether a string of co-occurring words is a collocation or not; it is a measurement of collocation certainty based on the frequency.

(2) **Log-likelihood (LL)** measures the likelihood of two words appearing in a corpus independently of each other.

(3) **Pointwise Mutual Information (MI)** evaluates the possibility of seeing a second collocate given a first and vice versa.

(4) **The Dice score** is like MI, but it is less sensitive to infrequent collocations.

These measures have their pros and cons when applied to the Russian language (for more detail, see the discussion in Pivovarova, Kormacheva, and Kopotev [2017]). Although probabilistic measures can result in strings with function words only as top collocations, they also help extract such strings as смотрите также ‘see as well,’ анонимный рецензент ‘anonymous reviewer,’ and неполным оканье ‘incomplete retention of unstressed [о],’ as well as other collocations that may be useful for L2 writers. While the method of using simple raw frequencies generates many clusters of function words, the collocational metrics listed above have a strong tendency to detect meaningful phrases (Pivovarova, Kormacheva, and Kopotev 2017).

The more computationally challenging task is to automatically correct miscollocations. In what follows we describe the approach CAT&kittens takes to provide possible substitutions for miscollocations found in L2 academic texts.

4.3.2. Distributional semantics in lexical and collocation deviation detection

Distributional semantics is a field of computational linguistics that attempts to understand semantic similarities of linguistic items based on their represented distributional properties. The idea behind distributional semantics is that words that occur in similar contexts tend to have similar meanings (Turney and Pantel 2010, 142–43); this idea is referred to as distributional hypothesis. The context consists of the content words in a window of a maximum of five tokens. The similarity of two words is a consequence of the fact that they are both tuned to predict the same context.
The words then are treated as points in multidimensional space so that distance between them can be calculated. The visualized representation of semantic distances of content words from example (2) can be found in Figure 2. Figure 2 is a two-dimensional visualization of the semantic distances between words автор ‘author,’ расследовать ‘investigate,’ исследовать ‘study,’ вопрос ‘question,’ установление ‘establishment,’ принятие ‘enactment,’ and закон ‘law,’ produced by the RusVectores skip-gram model trained on Taiga corpus (Kutuzov and Kuzmenko 2016). The figure shows that the closer the words are, the more similar are the contexts in which they occur (it can be said that such words are codistributed).

Figure 2. Semantic distances of the words in Example 3.

Models of distributional semantics, such as word2vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), ELMo (Peters et al. 2018), and BERT (Devlin et al. 2018), represent words as vectors of fixed dimensions (word embeddings) that are produced with neural networks. Some older models allowed only for context-free representation of word similarity; the more recent ones, such as ELMo and BERT, allow for differentiating between meanings based on the context. Distributional semantics frameworks can be used in many ways: for paraphrasing suggestions and plagiarism detection, in semantic searches, for word sense disambiguation, and for many other tasks. In our research, we are interested in the implementation of the models for replacing miscollocations.

Our platform uses a combination of statistics with distributional semantics to produce replacements of the miscollocations. The algorithm consists of three main steps:
(1) We retrieve all n-grams from the text submitted for checking. After that, all extracted n-grams are evaluated against the collocations found in the CAT. If an expression is not found in the reference corpus, it is marked as “unattested,” and the algorithm proceeds to the next stage. This step is applied to all 2-, 3-, and 4-grams; however, for the sake of brevity, we describe the following steps using only 2-gram collocations as examples.

(2) The system considers all words in the string for possible replacement; for example, the phrase расследовать вопрос ‘to investigate the question’ is analyzed as two strings “расследовать X_noun” and “XVerb вопрос.” Our algorithm then finds the closest semantic neighbor to the fixed word (that is not X in each string) to replace X, given that the substitutions for X are the same part of speech as the original X. This can be achieved in two ways: taking into consideration the meaning of the given collocate (расследовать ‘to investigate’ or вопрос ‘question’ in the case) or ignoring it. The first approach proved to be more efficient because the given collocate narrows the search scope. Using this method, we find a triangle in the space of embeddings (such as the space represented in Figure 2), where two dots are the original words and the third dot is a neighbor closest to both. This yields a list of several possible replacements for a given expression; for instance, исследовать вопрос ‘to study the question,’ where исследовать ‘to study’ is the closest semantic neighbor to both расследовать ‘to investigate’ and вопрос ‘question.’ Two approaches to semantic addition are generally used to locate a miscollocation and its substitutions in semantic space: additive and multiplicative (Mitchell and Lapata 2008). We utilize the additive model, which allows the algorithm to join several vectors and locate the nearest point in the semantic space.

(3) The resulting list of replacements may contain collocations that are non-attested in the academic genre and are, thus, infelicitous to use in an academic paper. To this end, a filter is used, which checks whether the suggested collocation is attested in the CAT. As a result, only those collocations that are semantically close to the original and are represented in the CAT are offered to the platform user for consideration.
The location of individual words within embedding spaces is by no means precise, and simple proximity within the embedding space is by no means a guarantee of well-formed collocations. Two assumptions are made in order to eliminate this issue. The first is based on the nature of collocations, which, as opposed to grammatical rules, are syntagmatic and linearly restricted. Therefore, applying the statistically based filters presented in Section 4.3.1 to the list of collocations reduces the number of occasional lapses. The second assumption is based on POS-filtering, which takes into account proper syntax and allows for filtering semantically close, but syntactically ill-formed word combinations. Filtering based on these two assumptions eliminates many of possibilities from consideration, but some still seem to be inappropriate word substitutions.

The algorithm we utilize also has two limitations: it provides collocations of the same length as the original one, and the output is lemma-based. The first limitation means that replacements from a bigram *делать исследование ‘to do research’ is only possible to a n-gram of the same length (e.g., проводить исследование ‘to conduct research’); the replacement to a unigram (e.g., исследовать ‘to study’) is not yet possible with our algorithm. The second consideration means that collocations from a learner text must be lemmatized before being processed. After all the possible replacements are found, the algorithm provides the user with the suggested lemma collocations. As a result, we get a collocation, which a writer should put in the right form.

To illustrate collocation substitution in action, we have taken an authentic learner text and run the replacement algorithm for two unattested bigrams. Some of the replacements are found in a general Russian corpus; however, not all are suitable for use in academic discourse. There are three main ways that the algorithm presents the results to the user. First, it provides the tokens for replacement, which are the collocations in the same form as the original one—for example, *расследовал вопросы ‘investigated the questions’ → исследовал вопросы ‘studied the questions.’ Second, if a token collocation is not found in the CAT, the algorithm suggests the closest lemma collocations, which the platform asks the user to grammatically adjust to the context: *расследовал вопросы ‘investigated the questions’ → ИССЛЕДОВАТЬ ВОПРОС ‘to study question’. Third, in cases when the algorithm can find neither tokens nor lemmas, it underlines the miscollocations and proposes that they are possibly nonstandard/nonacademic, without providing a possible candidate for substitution.

4 Capitalized letters designate lemma, not token.
Below, we will illustrate each step of the algorithm with two of the analyzed collocations to show token replacements and lemma replacements.

Case 1: *Большая важность ‘big significance.’* In the first step, the algorithm finds that the given collocation is not attested in the CAT. The collocation is then lemmatized and split into two sets: ‘X_{adjective} ВАЖНОСТЬ’ and ‘БОЛЬШОЙ X_{noun}.’ In the next step, the algorithm looks for semantically closest neighbors to both lemmas, noun and adjective, respectively (see Table 2).

Table 2. Cosine neighbors and their distances for БОЛЬШОЙ ‘big’ and ВАЖНОСТЬ ‘importance’

<table>
<thead>
<tr>
<th>БОЛЬШОЙ ‘big’</th>
<th>ВАЖНОСТЬ ‘significance’</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemma</td>
<td>cosine distance</td>
</tr>
<tr>
<td>ОГРОМНЫЙ ‘giant’</td>
<td>0.685</td>
</tr>
<tr>
<td>ГРОМАДНЫЙ ‘enormous’</td>
<td>0.648</td>
</tr>
<tr>
<td>ВАЖНЫЙ ‘important’</td>
<td>0.588</td>
</tr>
<tr>
<td>БОЛЬШИЙ ‘bigger’</td>
<td>0.586</td>
</tr>
<tr>
<td>КОЛОССАЛЬНЫЙ ‘colossal’</td>
<td>0.570</td>
</tr>
</tbody>
</table>

In the next stage, the algorithm looks up token word combinations in the CAT, which has the same tagset as was in the learner’s text; for example, the tagset for the word важность ‘significance’ is \(\text{NOUN Animacy=nan|Case=Nom|Gender=Fem|Number=Sing}\). It returns a list of all possible combinations, whether they are actual collocations or not—for example, огромная важность ‘giant significance,’ громадная важность ‘enormous significance,’ colossal significance, больная значимость ‘big significance,’ большое значение ‘great importance,’ and the like. To discriminate between these candidates, the last step is applied; the CAT filter, which looks for real, statistically significant collocations in
the reference corpus. The output is two well-formed collocations, больная значимость ‘big significance’ and большое значение ‘great importance,’ which both are semantically close to the original and yet are attested in academic discourse.

Case 2: *Расследую вопрос ‘I investigate a question’ is a semantic deviation where the learner likely confuses two words, расследую ‘I investigate’ and исследую ‘I study.’ After the algorithm identified the original collocation as unattested in CAT, it then suggested several semantically close replacements, outlined in table 3.

Table 3. Replacements and their semantic distances for *расследую вопрос ‘I investigate a question’

<table>
<thead>
<tr>
<th>РАССЛЕДОВАТЬ ‘investigate’</th>
<th>ВОПРОС ‘question’</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemma</td>
<td>cosine distance</td>
</tr>
<tr>
<td>ВЫЯСНЯТЬ ‘look into’</td>
<td>0.577</td>
</tr>
<tr>
<td>ОБСУЖДАТЬ ‘discuss’</td>
<td>0.550</td>
</tr>
<tr>
<td>РАЗБИРАТЬ ‘see into’</td>
<td>0.536</td>
</tr>
<tr>
<td>ОБЪЯСНЯТЬ ‘explain’</td>
<td>0.519</td>
</tr>
<tr>
<td>ИССЛЕДОВАТЬ ‘study’</td>
<td>0.387</td>
</tr>
</tbody>
</table>

The subsequent tagset-based token collocation search generated the following token phrases, both felicitous and infelicitous: выясняю вопрос ‘I look into a question,’ обсуждаю вопрос ‘I discuss a question,’ разбираю вопрос ‘I sort out a question,’ объясняю вопрос ‘I explain a question,’ исследую вопрос ‘I study a question,’ расследую расследование ‘I investigate an investigation,’ расследую проблему ‘I investigate a problem,’ расследую выяснение ‘I investigate an identification,’ расследую обвинение ‘I investigate an accusation,’ and расследую обсуждение ‘I investigate a discussion.’ Since only исследую вопрос ‘I study a question’ is attested in
the CAT above the frequency threshold ≥ 3, the platform suggests this collocation as the only possible replacement.

4.4. Evaluation of the procedures

Typically, a computational system is evaluated against two basic metrics: precision, or how many of the detected items are correct, and recall, or how many of all relevant items are detected. The latter is a challenging task, when it comes to automatic collocation substitution, since a list of substitution candidates can be, at least theoretically, unlimited if the non-attested collocation has no clear attested analogue. Hence, we evaluate only the metric of precision, to measure how many of the produced collocations are adequate substitutions for a detected miscollocation.

To evaluate the precision of our algorithm, before we test it on authentic learner data, we run it in a “controlled environment.” To this end, we randomly selected one hundred high-frequency 3-grams collocations from the CAT and created a set of four hundred phrases imitating learner errors. For this, we used the word2vec semantic model based on the Wikipedia and Russian National Corpus (Kutuzov and Kuzmenko 2016) and changed one or more elements of the original collocations to imitate learner miscollcations. All miscollcations were found to be unattested in the CAT. Example (4) provides the source token (a), its lemmatized version (b), and the produced lemmatized non-collocations (c–f).

Example 4.

a. Добыча полезных ископаемых
   extraction.Nom.Sg naturalGen.Pl resourcesGen.Pl
   ‘extraction of natural resources’

b. ДОБЫЧА ПОЛЕЗНЫЙ ИСКОПАЕМОЕ
   ‘extraction natural resource’

c. *ВЫПЛАВКА ПОЛЕЗНЫЙ ИСКОПАЕМОЕ
   ‘melting natural resource’

d. *ДОБЫЧА НЕОБХОДИМЫЙ ИСКОПАЕМОЕ
   ‘extraction necessary resource’

e. *ДОБЫЧА ПОЛЕЗНЫЙ МЕСТОРОЖДЕНИЕ
   ‘extraction natural deposit’

f. *ДОБЫЧА ПОЛЕЗНЫЙ РУДА
   ‘extraction natural ore’

5 We reiterate here that only lemmatized forms are provided; producing correct tokens is a separate task, which lies outside the scope of the present paper.
Based on this model, we located all miscollocations, obtained a list of substitutions, and checked whether the substitutions contain an original and, thus, correct collocation. In doing so, we ignored the fact that there might be other acceptable substitutions, instead we only controlled for the presence of one correct collocation. When looking at the first 1000 closest semantic neighbors and combining all possible word combinations for 400 miscollocations, the algorithm is able to recover as many as 375 correct collocations, which means that its precision is 93.75 percent. However, this relaxed approach slows down the time processing of the algorithm. If searching is limited to the first 100 semantic neighbors, the number of recovered correct collocations is downgraded to 290, which gives a precision of 72.5 percent.

Although our evaluation procedures prove the algorithm to work well, some further steps are needed. We considered how the algorithm is capable of returning the original form only under limited conditions—that is, by feeding it the miscollocations that imitate learner errors; a full-scale evaluation based on authentic learner data is needed (for more detail, see Kisselev & Furniss 2020).

Further refinement of the algorithm is required. First, we need to examine the difference between lemma- and token-based collocations, since many collocations are, in fact, token-specific. Second, word-embedding spaces produced by the models we have used are prone to mislocating homonyms like ключ ‘spring’ and ‘key.’ Extending our approach beyond the word2vec and fasttext models toward context-based models is a logical step to improve suggested replacements. We expect that the usage of two recently developed models, ELMo (Peters et al. 2018) and BERT (Devlin et al. 2018), will greatly improve the general quality of the algorithm output.

5. Conclusion
This paper reported the development of a large corpus of academic texts in Russian (namely, the CAT) and the writing support platform CAT&kittens that utilizes the corpus. The primary focus of the paper was to introduce one of the core features of the CAT&kittens platform, that of miscollocation detection and correction.

Automatic extraction and correction of learner miscollocations has been at the forefront of current research in how corpus-based analyses can be utilized in building writing support tools. Given that collocational knowledge is one of the most difficult skills to acquire in the L2, providing language support tools that help learners recognize and correct
miscollocations would be a valuable contribution to the field. Although the robustness of the proposed analysis and the implementation of the algorithm require further testing, the potential of the service under report is significant.

In addition to online detection and correction of misuse of academic language, the CAT&kittens platform has the potential to become a source and a resource for various investigations of this variety of the Russian language. Most currently available research on academic Russian—as well as most widely used pedagogical and reference materials for teaching and learning academic Russian—have rarely been based on or validated by contemporary corpus-based research, with some welcomed exceptions, such as Levinzon, Dzhakupova, and Pliseckaia (2014) and Talalakina, Stukal, and Kamrotov (2020). Our goal is to provide learners of Russian with modern corpus-based tools that will allow for a productive and thorough exploration of contemporary Russian academic genre.

References
Towards Intelligent Correction of Collocational Errors in Russian L2 Academic Texts

Klimov, Kisselev, Kopotev


Guo, Shesen, and Ganzhou Zhang. 2007. “Building a Customised Google-
Based Collocation Collector to Enhance Language Learning.” *British Journal of Educational Technology* 38, no. 4: 747–50.


Towards Intelligent Correction of Collocational Errors in Russian L2 Academic Texts
Klimov, Kisselev, Kopotev


Towards Intelligent Correction of Collocational Errors in Russian L2 Academic Texts

Klimov, Kisselev, Kopotev

