***SERIES:*** *LINGUISTICS*

*Ekaterina A. Vlasova[[1]](#footnote-1), Pavel O. Albitskiy[[2]](#footnote-2), Aleksandra A. Bogoiavlenskaia[[3]](#footnote-3),*

*Inna V. Glebkina[[4]](#footnote-4), Natalia A. Zevakhina[[5]](#footnote-5), Maria A. Ponkratova[[6]](#footnote-6),*

*Sergey A. Slastnikov[[7]](#footnote-7), Nikita R. Sidorov[[8]](#footnote-8), Alina Y. Sherstyuk[[9]](#footnote-9)*

**BUILDING THE RUSSIAN ACADEMIC PHRASE BANK:   
A DEEP LEARNING APPROACH**

**Summary**

This study contributes to pedagogically-oriented linguistic research aimed at building academic phrase banks, which lists formulaic expressions and meaningful lexical units aligned with their communicative functions. We explore how deep learning techniques can be applied to pragmatic annotation of academic discourse. The study reports the results of the corpus-based experiment on combining a rule-based annotation and a deep learning model. We developed an annotation scheme, connecting classes, derived from the previous research, with their linguistic realizations, recommended in the literature and extended by the means of similar words in SketchEngine. Having a delineated view, supported by the key word lists, annotators collected a corpus of textbook on political science, containing 850, 000 tokens after automatic preprocessing, and then prepared two datasets for training and testing a deep learning model. This approach allowed to achieve the general kappa of inter-annotator agreement of 0.82 and F1-score of 0.87 for the LSTM model with lemmatized ELMo embeddings. The error-analysis and linguistic examination of results demonstrated that the proposed model operates well enough on recognizing key phrases and correct predictions of key words in extended context. However, the model is sensitive to words with multiple derivates, such as *действовать* and *определять*, frequent in Russian academic discourse and causing wrong predictions. The paper discussed the ways to eliminate the observed noise for ELMo embeddings by adding more fine-graded techniques to the annotation procedure.

JEL Classification: Z

Keywords: academic Russian, phrase bank, SketchEngine, deep learning, LSTM, ELMo embedding, neural network, construction, academic vocabulary

1. **Introduction**

This paper reports the results of the educational and research project, prepared within the framework of the Academic Fund Program at the HSE University in 2021 [grant № 21-04-059] and aimed at automatic detection of linguistic expressions, centered around recurrent discourse categories of Russian academic texts. In particular, this study focuses on identifying a series of communicative functions emerging from research practice, and then matching them with a range of linguistic expressions of different grammatical structures, including phraseology, phrasal verbs, formulaic constructions, etc. Lists of the most frequent expressions, classified by their discursive and textual classes, provide useful insights for language teaching, automatic writing assistant services, style recommendations and editing practice. The computational technique, designed in our study, contributes to the emerging cross-disciplinary methodology, intended to combine corpus-linguistic approaches with machine learning methods for an automatic analysis of text corpora and large collections of texts, assembled with linguistic annotation and other labels with relevant information.

The analyzed data includes modern Russian academic texts, which remains a rather understudied phenomenon, sparsely covered in modern computational linguistics. There are several reasons for research interest in Russian academic discourse. First, this is an emerging area of linguistic studies, facilitated by increasing pedagogical demand for courses on academic writing and other tools [Klimov et al. 2018]. Second, an existing series of monographs on Russian academic style primarily provide description of scholar papers, written in the Soviet and Post-Soviet period [Митрофанова 1985, Клобукова 1987]. The described text collections and the reported observations do not reflect recent methodological trends in style, language and sociocultural circumstances, such as the recent rise of multi-disciplinary research, changes in post-soviet paradigms in the humanities and soft sciences, growing popularity of evidence-based approaches in many fields beyond the fundamental science. Due to the recent sociocultural and methodological developments, there is a demand for empirical linguistic explorations of Russian academic discourse, which contributes to applied linguistics and theoretical studies on genre, style, complexity and discipline-specific variation.

Modern computational linguistics provides an elaborated methodology for quantitative Russian studies [Kopotev 2017 et al.] and a large set of techniques [Gritsenko et al. 2021] for automatic processing of Russian academic texts, including a detection of contextual categories, compiling lists of n-grams and relevant expressions [Pivovarova et al. 2017], and other different tools for form-function matchings. Besides, the principles of open science guarantee favorable settings for building an academic corpus, as they facilitate academic publishing online with flexible types of the creative common license. Due to that, big collections of academic texts become available for constructing a corpus and its automatic processing. Machine learning also contributes significantly to linguistic analysis, empowering it with new approaches to classifications. However, training an effective algorithm requires preparation of adequate high-quality datasets, based on technically clean text collections with reliable linguistic annotation, eliminated data noise and transparent taxonomy of categories, regularly observed in the data.

This paper reports the design and construction of Russian academic corpus, focused on social disciplines, and development of an extended annotation scheme, based on pedagogically-oriented classifications of discourse categories, proposed in the literature. The prepared text collection was annotated manually and served a dataset for adjusting and training algorithms of machine learning. Our study serves as a testcase to discuss an application of neural network techniques to match recurrent linguistic expressions to the discursive research-oriented categories of Russian academic texts. The central research questions of this article are as follows:

1. To what degree different algorithms can learn to assign pragmatic categories of a complex annotation scheme to collocations rather than to sentences?
2. Is our annotation procedure robust enough to be learned by networks?
3. What elements of annotation are captured by neural network and what are omitted?

The structure of the paper is as follows. The section 2 reviews the related works and summarizes the findings, influential classifications and computational techniques, impacted the ongoing research of academic discourse. The section 3 describes the corpus and the dataset. The section 4 discusses the design of annotation scheme and reports the metrics of inter-annotator agreement. The section 5 describes the designed deep learning architecture, the overall accuracy metrics of the model and error analysis. The section 6 examines linguistic phenomena underlying the incorrect predictions. The conclusion summarizes the observations on the model performance and discusses annotation techniques to eliminate the observed noise.

1. **Related work**

The universal principleы of academic discourse have been a subject of many linguistic explorations, including theoretical monographs [Hyland 2008, Biber&Gray 2010, Flowerdew 2014, Biber&Gray 2016] and applied studies, primarily determined by pedagogical interest to university language [Biber 2006] and academic writing [Hyland 2008]. Three lines of research are observable in the literature: (i) content-based analysis and systematization of different rhetorical categories (ii) explorations of recurrent linguistic patterns (iii) connecting textual categories and their linguistic realizations.

The content-based studies of academic texts examine compositional sections of professional scholar papers (articles and dissertations) and develop taxonomies of reoccurring categories. The most elaborated and exhaustive classification has been designed for introductions [Swales 1981, Swales 1990, Swales 2004]. Swales’ model, called Create a Research Space, captures three regular rhetorical categories withing three thematic classes, such as “establishing a research territory”, “establishing a niche”, “presenting research”. Within content-based line of research, another influential concept is *generic move*, justified by [Bunton 2002] and later developed as a taxonomy of *rhetorical moves and steps* [Pho 2008]. Later, a few studies proposed inventory of regular rhetorical categories for other compositional sections of research papers, such as conclusions [Bunton 2005] and literature review [Flowerdew&Forest 2009].

The book-length study [Biber 2007] demonstrated how corpus-based methods can be used for discourse analysis, applied to the description of academic discourse organization. Due to development of corpus-based methods, the recent decades have seen much interest in recurrent linguistic patterns, their frequencies and distributions across academic disciplines. A large body of corpus-based studies focuses on lexical bundles [Biber et al. 2004], which are multi-word expressions (or collocations), frequently used in academic texts and automatically identified in corpora, e.g. *in terms of the*, *it was found that*, *as shown in fig*., *in the case of*, etc. Importantly, lexical bundles are discourse units (or “building blocks of discourse” [Biber et al. 2004], rather than grammatical or idiomatic units. As the bulk of the literature has been concentrated on English lexical bundles, they are typically 4-word expressions that include content and functional words. Investigating lexical bundles leads to text corpora development. Text corpora include, on the one hand, novice academic texts that are student texts of various disciplines and, on the other hand, academic journal texts. To illustrate, Biber and Barbieri [2007] makes a comparison between classroom vs. written academic discourses, whereas Cortes [2004] investigates professional vs. student written academic discourses.

These academic corpora, containing educational and professional texts, enable linguists to study various aspects of academic discourses and conduct comparative studies on register variation and diachronic changes between disciplines [Hyland&Jiang 2018, 2021], different sections of scientific articles [Haotong et al. 2020], novice and professional writers [Qin 2014]. An emerging line of research focuses on automatic extractions of lexical bundles that constitute semantically congruent formulaic expressions [Iwatsuki&Aizawa 2018, Iwatsuki et al. 2022].

A growing body of research is intended at matching expressions, extracted from a corpus, and rhetorical moves. Cortes [[2013](https://onlinelibrary.wiley.com/doi/full/10.1111/lang.12250?casa_token=rNbNI_GIuzwAAAAA%3Atda8oZSkjT31pQ4lPQYacR7TYStjQRbwhC7HzbsuN7HmcPhwhL_21gVWEnVscY69K_dneDjNMP1zO7NB#lang12250-bib-0030)] recently suggested an approach to “connecting lexical bundles with moves according to their functions” in research article introductions, and this idea was further extended on practical applications, such as language pedagogy [Lu et al. 2020] and online writing tools [Mitzumoto et al. 2017]. Earlier, Durrant and Mathews-Aydinli [2011] proposed a function-first approach, which implies preliminary annotation of a text by communicative functions and then automatic extractions of recurrent linguistic patterns within each function. As it was recently reported [Bender et al. 2020], deep learning models provide a new toolkit for automatic extraction of formulaic expressions and meaningful lexical bundles, classified by their textual categories. The results of machine learning predictions, however, are so far lower than expected.

In our paper, we report the results of the experiment, aimed at development of a deep learning model for automatic detection of semantically congruent academic expressions and their functions, as they defined in pedagogically-oriented taxonomy of communicative acts, typical for Russian scholarly papers. The settings include compilation of our own dataset, an extended ruled-based approach to annotation and a deep learning model, adjusted to the training dataset.

1. **The data**

The Russian academic discourse is well-represented online and available for linguistic explorations. A large body of recently published research articles are available with favorable creative commons licenses, legalizing the use of scholar papers with mentioned authorship. These legal regulations allow computational linguists to download and process large collections of academic papers for corpus-based investigationsata, used for the datasets, are drawn from a text collection, representing the academic discourse of textbooks on political science. To follow the principle of representativeness, the choice of authors was not arbitrary. We analyzed the literature recommended for the courses on political science in leading Russian universities, qualifying specialists in politics and international relations. After comparing the books, mentioned in all literature lists, against the most popular books, ordered in electronic libraries, six popular textbooks were chosen. The Table 1 presents the detailed information about the content and size of the initial corpora.

|  |  |  |  |
| --- | --- | --- | --- |
| Tab. 1. The corpus of the Russian textbooks on political science | | | |
| Author | Title | Size, tokens | Coverage |
| Селютин В.И. | Теория и практика политической науки. Воронеж, 2009. | 82,457 | 12% |
| Соловьев А.И.,  Пугачев В.П. | Введение в политологию. М., 2000. | 91,079 | 13% |
| Баранов Н.А. | Политические отношения и политический процесс в современной России: Курс лекций. СПб: БГТУ, 2004. | 115,466 | 16% |
| Соловьев А.И. | Политология: Политическая теория, политические технологии: Учебник для студентов вузов. М, 2006. | 116,145 | 16% |
| Макарин А.В. | Теория и история политических институтов. СПб, 2008; | 103,061 | 14% |
| Туровский Р.Ф. | Политическая регионалистика.  М.: Издательство ГУ-ВШЭ, 2006. | 207,785 | 29% |

While preprocessing the collection for further computational analysis, we removed pictures, figures, schemas and tables, deleted notes, footnotes, multiple repetition of the title in footers, page numbers, etc. A separate issue was the reference sections, because authors**’** names and book titles in different languages and alphabets establish data noise, crucial for small-sized exploratory datasets. For further machine learning tasks, we split the corpus of political science into two datasets for training the model and for its testing, as it is recommended in the literature [Pustejovsky&Stubbs 2013]. Table 2 represents quantitative characteristics of the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Tab. 2. The datasets for the machine learning experiment | | | |
|  | training dataset | testing dataset | collection size |
| sentences | 19246 | 22539 | 41785 |
| words | 323298 | 399734 | 723032 |
| punctuation | 57118 | 68790 | 125908 |
| total | 382972 | 473153 | 856125 |

Both datasets were lemmatized by the Spacy pipeline for the automatic linguistic annotation. Every word was assembled with its lemma, part-of-speech and syntactic features.

1. **Annotation design**

For this study, we designed an extended annotation scheme, derived from the state of the art in research on academic discourse [Hyland 2009, Biber et al. 2007], corpus-based approaches to compiling phrase lists [Pivovarova et al. 2017, Kopotev et al. 2021, Lu et al. 2021], function-first approach to identifying formulaic language in academic texts [Durrant&Mathews-Aydinli 2011] and functionalism framework in pedagogically-oriented studies of Russian for professional use [Величко 2009].

* 1. **The annotation schema**

The pedagogically-oriented studies, summarized in the monograph [Величко 2009], proposed a prototype of the Russian Academic phrase bank as a list of recurrent expressions, grouped in research-oriented regular categories, called communicative functions. The phrases included in the lists reflect teachers’ intuition, supported by close-reading examinations, and thus represent a bottom-up investigation of academic texts. This classification establishes a comprehensive taxonomy and serves as a starting point for computational explorations. This prototype for the Phrase Bank covers 14 communicative functions of academic texts and the relevant phrases and grammatical structures centered around a verb or predicative adjective.

|  |  |
| --- | --- |
| Tab. 3. Annotation scheme. The communicative functions of Russian academic text. | |
| Category and its description | Recommended phrases |
| EQUIVALENCE  “X is Y” for relations of equivalence | *X – это*, *X – это Y*, *X есть Y*, *X является чем,* *X представляет собой* |
| SUMMARY  “X is Y”, X and Y are abstract nouns such as *причина* **«**reason**»** | *X заключается в чём/том*, *X состоит в чём/том*, *X сводится к чему / тому*, *X выражается в чём/том* |
| DEFINITION  “X is called X”  for definitions | *X называется чем, X принято называть чем, X называют чем, X получило название*, *X носит название*, *X известно под названием чего.* |
| RESEARCH  “X analyzes” research activities and multiple syntactic derivates | *X изучает Y*, *X исследует Y*, *X анализирует Y*, *X cтавит задачей Y*, *X является объектом Y*, *X составляет задачу Y*, *X является задачей Y* |
| CLASSIFICATION  “X classifies Y as” for division of examined phenomena into groups | *X классифицирует / делит*, *X группирует / объединяет*, *X X выделяет Y*, *X относит Y к чему*, *кто включает что во что* |
| PART-TO-WHOLE  “X consists of Y” for the division of a phenomenon into elements | *X состоит из чего*, *X включает что*, *X содержит что*, *X составляет что*, *X образует что*, *X входит во что / в состав чего*, *X относится / принадлежит к чему*, *X содержится в чем, X выделяется в чем*, *X делится на что*, *X подразделяется на что*, *X распадается / разлагается на что*, *X членится* |
| FUNCTION  “X functions as” for indicating a role or destination of an object | *X служит чем, X выступает чем / в качестве / как что, X выполняет что* |
| APPLICATION  “X is used as” for use and implementation | *X служит чем*, *X используется для чего / где*, *X употребляется для чего / где*, *X применяется для чего / где* |
| EXISTENCE  “X exists or is observable in” for posing a phenomenon | *существует X, имеется X, где есть X, где наблюдается X, где встречается X, где бывает X, где распространено X, где содержится X, где нет X, где отсутствует X, кто лишен X* |
| FEATURE  “X typical for Y” | *X имеет что*, *X обладает чем*, *X характеризуется чем*, *X отличается чем*, *для X характерно / типично / свойственно / присуще что*, *X наделен чем* |
| OBSERVATION  “X is observable in” | *X охватывает что*, *X распространяется на кого / на что*, *X известно гдe*, *X распространено где / среди кого* |
| RANGE  “X covers Y” | *X охватывает что*, *X распространяется на кого / на что*, *X известно гдe*, *X распространено где / среди кого* |
| CAUSE “X causes or determines Y” for causal nexus | *X обусловливает что*, *X определяет что*, *X зависит от чего*, *X связано с чем*, *X находится в связи с чем*, *X находится в зависимости от*, *X определяется чем*, *X обусловливается чем*, *X обусловлено чем*, *X испытывает влияние / воздействие чего*, *X подвергается воздействию / влиянию чего*, *X вызывает что*, *X ведет к*, *X* *сказывается на что*, *X отражается на что*, *X помогает чему*, *X способствует чему*, *X мешает чему*, *X препятствует чему* |
| CHANGE  “X changes, becoming Y” | *становится чем / каким*, *X делается каким*, *X делает что каким*, *X меняет / видоизменяет что*, *X изменяется*, *X подвергается чему*, *X превращается во что*, *X переходит во что*, *X сближается с чем*, *X совпадает с чем*, *X получает / приобретает*, *X теряет / утрачивает что*, *X сохраняется / сохраняет что*, *X увеличивает / уменьшает что*, *X повышает / понижает что*. |

The taxonomy, presented in Table 3, covers the most common communicative acts reoccurring in academic texts, regardless of the genre and rubrics. However, implementing this classification to the computational analyses reveals a few problematic cases. First, the classes are unbalanced regarding a number of lexemes listed for different categories: the class OBSERVATION only contains three words with the same syntactic structure, while the class EXISTENCE with a similar meaning includes ten words. Second, the same linguistic expressions belong to different classes. This problem concerns phrases with verbs *X распространено где* (EXISTENCE and RANGE), *X служит чем* (FUNCTION and APPLICATION), *выражается в* (SUMMARY and EXISTENCE). Annotating the same word sequences as different classes establishes noise and impairs the quality of the model’s predictions. Third, the manual trial annotation based on the listed expressions, revealed a systematic issue of polysemy and other ambiguities emerging in an extended context.

**4.2. The annotation procedure and the rule-based approach**

The academic texts of soft science are of specific nature, as the humanities often operate on abstract categories and classes, and many phenomena establish a continuum with vague borders. It is also common for pragmatic annotation, like labeling communicative acts, to tackle phenomena with multiple interpretations [Archer et al. 2008]. Given that academic texts are generally more challenging in regards to lexical, grammatical, and information complexity [Biber&Gray 2016], it is an expensive and not reliable task to set up a classification and annotation procedure for students based only on the definitions and their intuitive understanding of research categories in academic texts. To make the close-reading process and decision less time-consuming and classes clearer, we gave annotators a delineated view on the category based on the claim that regular communicative categories of research-related texts manifest themselves linguistically in two ways: i. high-frequent lexemes and recurrent linguistic structures; ii. wide synonymic fields of predicate lexemes, constructions, and idioms, describing the same phenomena or nuances. These assumptions allow us to develop a rule-based approach to annotation. More specifically, it is possible to compile a wordlist of lexemes and structures, signifying the communicative function to annotate.

The starting point for compiling such a list was the form-function classification, proposed for the Russian academic language and analyzed in Table 3. For the experiment setting, we chose the four broadest classes: EXIST for identifying phenomena, their existence, and their manifestations, CAUSE for marking causal relations, CHANGE for descriptions of observable changes, SUMUP for clarifying concepts and ideas. Then we revised the list of the relevant linguistic expressions for each category and searched for similar words using the SketchEngine. More specifically, we used the function “Thesaurus” for each predicate lexeme from the list , compiling synonyms and similar words. We used the Russian Web (RuTenTen11) corpus for this analysis, embedded in the SketchEngine. Then we verified the resulting wordlists against the words, compiled from our collection of political science textbooks, used for the experimental setup.

For our study, aimed at developing an annotation procedure and testing an appropriate deep learning model, we chose four classes. Table 4 contains the extended list of keyword lexemes compiled for the rule-based annotation. The sign \* indicates a prefix.

|  |
| --- |
| Table 4. The rule-based approach to the annotation: the extended keyword list |
| tag EXIST  for identifying phenomena, their existence, and their manifestations |
| существовать, бывать, наблюдаться, проявляться / проявиться, выявляться / выявиться, обнаруживаться / обнаружиться, фиксироваться / зафиксирован, отражаться в / отразиться, выражаться в / выразиться, скрываться / скрыться, возникать / возникнуть, появиться / появляться, \*родиться / \*рождаться, создаваться / создаться, формироваться / сформироваться, \*страиваться / строиться, складываться / сложиться, воспроизводиться / воспроизвестись, воссоздаваться / воссоздаться, реальным - \*делать реальным, становиться и стать реальным, воплощаться / воплотиться, реализоваться / реализован, утвердиться / утверждаться, закрепляться / закрепиться, установиться / устанавливаться, актуализоваться, актуализироваться, быть / становиться актуальным, \*кристаллироваться, состояться, осуществиться / осуществляться, стоять перед, поставить перед, встать перед, происходить / произойти, действовать, получить отражение в, находить / найти отражение, быть проявлением, при обилии, при многообразии, при изобилии, в/при отсутствии / отсутствие, в / при наличии, присутствовать, отсутствовать, наступать / наступить, знаменовать / ознаменовать, изобиловать, иметь место, иметься, находиться, исчезать / исчезнуть, пропасть / пропадать, раскрыться / раскрываться, вскрыться / вскрываться, материализоваться, \*зреть / \*зревать, восходить к, активизироваться, активизирован |
| tag CHANGE  for descriptions of observable changes |
| \*меняться / \*мениться, трансформироваться, сужаться / сузиться, расширяться / расшириться, усугубиться / усугубляться, обостряться / обостриться, \*расти, повышаться / повыситься, претерпевать / претерпеть изменения, подвергаться | подвергуться изменению, делаться / сделаться, понижаться / понизиться, превращаться / превратиться, развиваться |
| tag CAUSE  for marking causal relations |
| влиять на, воздействовать на, определять, оказывать эффект, влияние, воздействие, приводить к / привести к, влечь / повлечь за собой, вызывать / вызвать, обусловливать / обуславливать / обусловить, отражаться / отразиться на, сказаться / сказываться на, определяться, испытывает воздействие / влияние, помогать чему-то, способствовать чему-то, стимулировать, стимул, в зависимости от, зависеть от, провоцировать, породить / порождать, вытекать, ввиду, вследствие, по причине, в результате, отсюда следует, происходить из, толчок к, подтолкнуть, стимулироваться, вызываться |
| SUMUP for clarifying and specifying concepts and ideas |
| заключаться в том, что; состоять в том, что; предполагать; подразумевать; суть; по сути дела; сводиться к / свестись к; сводить к / свести к |

An annotator is given the predefined list of keywords, which are predicate lexemes of different morphological structure, including verbs and verbal forms, linking verbs, short adjectives, predicative adverbs, some phraseological expressions (при наличии ‘while someone has’, по сути дела ‘in fact, essentially’).

Two approaches are usually adopted for machine learning tasks to classify word sequences [Wilcock 2009]. The most common technique is to split a text into sentences and then label a sentence based on the key words and features it contains. The sentence-based annotation does not precisely suit our goal since Russian academic texts usually contain long subordinate syntactic units, and one sentence includes phrases belonging to different classes. Another approach concerns an annotation of phrases, namely key words with extended context, wherein an annotator labels a sequence of words significant for the task. The annotation unit is a lexis-grammar sequence with a predicative key word in extended context, containing lexical clues, disambiguating the possible polysemy of categories. To provide the model enough context, we formulated recommendations to include in the annotated sequence words of three types: functional words, such as a) pronouns, prepositions and conjunctions, b] general academic words like тип ‘type’, признак ‘feature’, зависимость ‘causal nexus’, взаимодействие ‘interconnection’ c) specific high-frequency lexemes of academic texts определенный «certain», своеобразный «kind of», etc. d) some high-frequent general vocabulary, needed for disambiguation of polysemy, such as подвергаться изменению ‘undergo changes’ (tag CHANGE) – подвергаться воздействию ‘to be affected by’ (tag CAUSE).

To collect recurrent phrases, classified by their research-related communicative function, we split the text into tokens, which are word forms and punctuation marks. Thus, a tagged unit is a relevant sequence of tokens. An annotator is supposed to find a keyword from the list, define a semantically congruent context and label the relevant word sequence, as Table 5 represents.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tab. 5. The sample of the annotated datasets | | | | | |
| no | word | lemma | annotator1 | annotator2 | pos |
| 1 | В | в | CAUSE | CAUSE | ADP |
| 2 | результате | результат | CAUSE | CAUSE | NOUN |
| 3 | совершения | совершение |  | CAUSE | NOUN |
| 4 | подобных | подобный |  |  | ADJ |
| 5 | ожидаемых | ожидать |  |  | VERB |
| 6 | действий | действие |  |  | NOUN |
| 7 | создаются | создаваться | CAUSE | EXIST | VERB |
| 8 | устойчивые | устойчивый | CAUSE |  | ADJ |
| 9 | связи | связь | CAUSE |  | NOUN |
| 10 | и | и | CAUSE |  | CCONJ |
| 11 | взаимоотношения | взаимоотношение | CAUSE |  | NOUN |

The annotators were asked to take a moderate approach and avoid labeling too extensive word sequences. Each annotator had an instruction with the guidelines, providing annotation principles , a definition for each category, a wordlist, and recommendations on the annotated phrase length.

**4.4. Inter-annotator agreement**

Although the designed rule-based approach and the guidelines, containing a wordlist of lexemes, establish a rigid framework for an annotator, there is still a certain margin in terms of the length of a relevant word sequence or relevant context. The key factor is subjectivity in interpreting the relevant words to annotate when the context is full of details. To control the annotators` performance and to what extent they were congruent in their classification, we calculated the inter-annotator agreement using coefficients of Cohen’s kappa, which expresses the relationship between the expected coincidence and the actual observed match. We calculated the number of tokens with the same label, using the formula recommended in the literature [Ide, Pustejovsky 2017]. Table 6 presents the results for each tag and general kappa.

|  |  |
| --- | --- |
| Tab. 6. Inter-annotator agreement, Cohen’s kappa | |
| general kappa | K = 0,82 |
| CAUSE | K = 0,96 |
| EXIST | K = 0,85 |
| CHANGE | K = 0,39 |
| SUMUP | K = 0,70 |

Cohen’s kappa indicates very high agreement for the tags CAUSE (k=0.96). Annotators were also congruent while defining the category EXIST with a kappa coefficient of 0.85. The calculations show enough agreement for the tag SUMUP 0.70. The most problematic tag was CHANGE, with only 0.39 percent of agreement. The close-reading investigations show that the reason for such a low performance was a technical issue with some of the files. Further, we will show that the model still was able to annotate the text by categories at a sufficient level.

The Cohen’s kappa coefficients with a range of 0.7−0.96 and general kappa of 0.82 are reported as good or even very good agreement in the literature. These numbers indicate that the rule-based annotation of expressions and constructions focusing on the keywords is reliable for classifying phrases and constructions with enough agreement and objectivity.

1. **Model architecture**

This section formalizes our linguistic objective as a machine learning task. Then we describe the key parameters of the input and the model architecture. Further, we motivate the choice of the applied neural network and report our experimental setup. Situating the initial objective in machine learning methodology, we want a neural network to find extended key phrases and predict their communicative function to resolve the homonymy and polysemy of verbs with abstract meaning. The algorithm is supposed to perform natural entity recognition to detect the key phrases and their length, and then the classification task to choose a correct semantic tag.

**5.1. Input**

The starting point of the NLP task is a conversion of text into numeric data to make words and signs computer-readable and perform mathematical operations required by neural networks. There are various approaches to conduct such a transformation, depending on the importance of contextual clues and semantics.

This is just the initial conversion of words into numbers so that they can be submitted to the machine learning model, while information about the semantics and context of the word is gradually taken into account by the model in the process of training. In many machine learning methods, each word receives a new vector representation, containing all the necessary information about the word and its context clues. Such a representation is called an embedding, and a qualitative model words with similar meanings have similar embeddings in vector space.

For the transformation of natural words, modern neural network algorithms mostly exploit pre-trained embeddings, which are lexical vector representations formed inside the model while training it on a huge data collection with high computing power and a long training time. Obtaining the pre-trained embeddings is a computationally expensive task; that is why they are published online as repositories of vectors and computing code, open for developing algorithms to solve more specific tasks. Models designed on the pre-trained embeddings demonstrate enough quality and reliability in finding dependencies in data.

There are several pre-trained embeddings commonly used for NLP tasks. One of the earlier popular models is *word2vec* [Mikolov 2013], which effectively detects [synonymous](https://en.wikipedia.org/wiki/Synonym) lexemes or suggests similar words. The recently released more sophisticated contextualized embeddings, such as ELMo and Bert [Peters et al. 2018], consider the context of the word while vectorization and deal effectively with polysemy, homonymy, as well as syntactic and semantic dependencies in the analyzed textual data.

This study utilized pre-trained ELMo embeddings, released in 2019 and trained on the lemmatized text corpus of 989 billion words from the Russian-language Wikipedia and the Russian National Corpus. Using this version of ELMo, we replaced every word in our training dataset with its embedding and then submitted their sequences as input units to our model.

**5.2. Model**

In our task, the unit of computational analysis is an annotated key phrase, which is a sequence of words in the model input. This is why we applied a recurrent neural network, following a classical approach to processing a sequence of data. This model iteratively processes the input sequence of vectors, preserving some information about preceding elements. At the first iteration, the first element of the sequence enters the model to be processed, and the resulting number is called the hidden state, which proceeds to the second iteration in the same model along with the second sequence vector. Together they are processed, giving a new hidden state, which is passed to the next iteration along with the next element of the input sequence. The described procedure continues until all the elements of the sequence are processed. Based on the hidden states obtained for each element, one can build a classifier by passing them through a fully connected layer with the number of neurons equal to the number of classes. Thus, a recurrent neural network receives a sequence of elements (vectors) as input and generates a sequence of responses for these elements.

It is generally accepted that the best results are often achieved by a type of recurrent network called LSTM (Long Short-Term Memory). This type of network has been specially designed to capture long-term dependencies in the data sequence, and its architecture is well-suited and typically used for processing phrases and sentences rather than separate words.

In addition to hidden states, the LSTM builds an extra vector transmitted between iterations. Its information is controlled by three special filters and does not change as intensively as in hidden states. The sequence can be processed in the forward order and in reverse. Models that process data in both directions and then concatenate the results for each element are called bidirectional recurrent networks.

Data sequences are sentences consisting of vectorized tokens in the problem under consideration. Each sentence is fed into the bidirectional LSTM model and receives a sequence of tags as a result – one tag for each element. So, the model detects dependencies between all the words of the sentence.

We used the LSTM for classifications of phrases, and the final outputs are predictions of labels, specifying a communicative function of a word sequence based on a given input unit. For all given functions, the model learns properties of representation of phrases, labeled with the same tag, and how they look, and then makes predictions for unseen phrases. While evaluating the quality of the model, we make predictions and compare them to the actual human annotations that we had in our datasets.

**5.3. Experimental setup**

Our experiments are based on the dataset constructed from the collection of university textbooks on political science. The total data size covers 41,785 sentences and 856,125 tokens, including 723,032 words and 125,908 punctuation marks. The total data was split into two parts. The training dataset, annotated by the more experienced master degree students, contained 22,539 sentences and 473,153 tokens, including 399,734 words and 68,790 punctuation marks. Another dataset used for the testing was of comparable size, covering 19,246 sentences and 382,972 tokens, or 323,298 words and 57,118 punctuation marks. The testing dataset was annotated by bachelor degree trainees, instructed for the annotation process, but less experienced in linguistic analysis and identification of communicative function. The experiment setup included only five broadest annotation categories: EXIST for identifying the existence and their manifestations of phenomena, CAUSE for marking causal relations, CHANGE for descriptions of observable changes, SUMUP for clarifying concepts and ideas, tag O for not-specified sequences of tokens.

**5.4. The results of the model experiment**

Table 7 shows the results of fitting the LSTM-model on the testing dataset with five annotated categories and the scores for Precision and Recall, weighted F1 scores.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tab. 7. The results over different categories | | | | |
|  | precision | recall | F1-score | support |
| CAUSE | 0.90 | 0.85 | 0.88 | 644 |
| CHANGE | 0.90 | 0.77 | 0.83 | 342 |
| EXIST | 0.91 | 0.80 | 0.85 | 962 |
| SUMUP | 0.93 | 0.70 | 0.80 | 969 |
| O | 1.00 | 1.00 | 1.00 | 140257 |
| accuracy |  |  | 0.99 | 143174 |
| macro avg | 0.93 | 0.82 | 0.87 | 143174 |
| weighted avg | 0.99 | 0.99 | 0.99 | 143174 |

Table 7 contains three common metrics to estimate the quality of the model. For each given class, precision shows how many objects labeled by the model truly belong to this class. Conversely, recall illustrates the portion of objects, which the model correctly recognized. F1-score is a harmonic mean of precision and recall. Additionally, support indicates how many times instances of this class were found in the sample.

At the bottom of the table, there is information about accuracy (the portion of correct answers), averaging of precision, recall and F1 (line “macro avg”), and their averaging with weights equal to the frequency of each class. The issue with interpreting the validity of these metrics is that they are biased by class frequency and take into account the distribution of classes among the dataset. Our classes are unbalanced. The objects of the class “O” make up about 98% of the entire sample; that is why such metrics as accuracy and weighted precision, recall and F1 are not reliable since they make sense only with balanced samples. In our case, even for a trivial model that always predicts the class "O", these metrics would be 0.98 or 1. Thus, in many aspects, the most reliable quality metric is averaged F1-score, which is 0.87.

Further, we will focus on the individual metrics for each class, as they are trustworthy, regardless of the balance of the sample, and reflect the quality of the model quite well. The results demonstrate that the model performs well enough for all the proposed classes and the quality for each class is not lower than 0.8. Among meaningful tags indicating communicative functions, “CAUSE” has the highest quality or recognition (f1 = 0.88), while “SUMUP” has the lowest quality (f1 = 0.8).

The next step of analysis concerns a confusion matrix, which demonstrates how often the examined model confuses one class with another. This confusion matrix is presented in Figure 1.

|  |
| --- |
|  |
| Fig. 1. Confusion matrix of the LSTM-model |

For clarity, the words with the tag "O", which the model classified as "O", have been removed from consideration. The matrix shows that the model more often confuses meaningful classes with “O” than with each other. In other words, the model does not recognize some phrases and their parts as communicative functions. We performed a close-reading analysis of the model’s errors to understand their linguistic nature for more detailed explanations of this phenomenon.

1. **Linguistic validation and interpretation of results**

In this section, we report the results of close-reading linguistic analysis of incorrect predictions made by the model. More specifically, we extracted the list of words, wherein the manually annotated tags are different from automatically assigned labels. Then we explored the most common linguistic factors hindering computational analysis due to various asymmetries inherent to language data, such as homonymy and polysemy, derivation and prefixed verbs in Russian, Zipfian patterns in word frequencies, free word order with dependencies encompassing distant positions of phrases and elements. Although more sophisticated word embeddings, such as the ELMo, are designed to eliminate the linguistic noise and capture the syntactic patterns, the resulting performance of the model depends on the specific task. Our exploratory linguistic analysis is intended to detect what grammatical and lexical qualities of Russian words cause errors in the automatic annotation.

We extracted the words with different tags from the validation dataset to explore the differences between a real annotator and the model performance. As the confusion matrix showed, the most common errors are confusions with “O” class, which includes not-classified words. When the model assigns the label “O” to a word with the other tag, this indicates underperformance of the algorithm. Simply saying, the model misses significant words for our task. When the model assigns a tag to a word, classified by an annotator as “O”, that means overperformance, which usually happens for several reasons. First, the annotator eliminates the homonymy and excludes the construction with different meanings, not relevant to the class. If the model still labels the omitted word, this indicates that the algorithm is ineffective in dealing with homonymy. Second, manual annotation is not a perfectly exact procedure. A human may miss a tag for technical reasons or the wrong interpretation of the guidelines. When the model still finds and classifies words correctly, this indicates that the algorithm solves the problem better than an annotator.

A proficient expert annotated the words with different tags assigned by a human and the model to validate the models` results, using two interpretations. The tag overperformed concerns the cases with correct model predictions for words missed by an annotator. The tag underperformed indicates incorrect model predictions when the model does not recognize significant keywords.

The preliminary analysis of the validation dataset also showed that the most common reason for different predictions is that an annotator usually marks extended contexts of the keywords to give lexical clues to the model for resolving homonymy, while the model assigns the label only to keywords, omitting the extended context and labelling the variable part of the context as not-classified. For this reason, we added an interpretation” element” to mark the words, establishing variable context. Table 8 demonstrates the results of the expert’s evaluations.

|  |  |  |
| --- | --- | --- |
| Table 8. The results of the expert’s evaluation of the model | | |
| **evaluation** | **count** | **%** |
| element | 670 | 63% |
| underperformed | 270 | 25% |
| overperformed | 131 | 12% |
| total | 1072 | 100% |

It can be seen that 63% of confusions concern variable words from extended context. The model captures short constructions, such as *лежат в основе*, but fails to annotate a longer expression *лежат в основе изменений*, which is also a semantically congruent and relevant phrase for our task.

At the next stage, we extracted the frequency list of lexemes with confusions and analyzed the words from the top, see Table 9.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 9. The most frequent lexemes, causing model confusions | | | | |
| **lemma** | **element** | **underperformed** | **overperformed** | **total** |
| действовать | 1 | 28 | 8 | 37 |
| определять | 1 | 22 | 7 | 30 |
| основа | 0 | 12 | 13 | 25 |
| определяться | 0 | 11 | 0 | 11 |
| основание | 0 | 6 | 5 | 11 |

Table 9 shows that many confusions are associated with specific words. The model regular leaves unclassified words *действовать* и *определять*, but, in some cases, recognizes them correctly and better than an annotator. The exploration of the frequency list reveals an interesting feature of these lexemes. The verb *действовать* ‘act’ has multiple prefixed verbs, fairly frequent in academic texts. Table 10 represents the list of the prefixed verbs with their frequency, measured as instances per million.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 10. The derivates of the verb *действовать* | | | |
| **lemma** | **translation** | **political science, ipm** | **the RNC academic subcorpus** |
| действовать | act | 796.88 | 313.34 |
| взаимодействовать | interact | 64.65 | 25.27 |
| воздействовать | affect | 49.19 | 20.85 |
| содействовать | promote | 16.87 | 28.04. |
| противодействовать | counteract | 11.02 | 5.83. |
| задействовать | involve | 12.65 | 5.59. |
| бездействовать | not act | 4.22 | 1.88. |

Compared to the other verbs, labeled as keywords, *действовать* has multiple prefixed derivates, and this factor potentially may contribute to confusions, as the model employs contextualized embeddings. Another issue concerns the verbs *определять* and *определяться*, which often remain unclassified by the model. The issue about this pair is that it they have a few high-frequent derivates, such *определять* ‘define as; affect’, *определяться* ‘be affected by’, *определенный* **‘**certain; defined**’**, *неопределенный* ‘uncertain, undefined’, *определение* **‘**definition**’**.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 11. The derivates of the verb определять | | | |
| **lemma** | **translation** | **political science** | **the RNC academic subcorpus** |
| определенный | certain, defined | 1405.44 | 422.80 |
| определять | affect | 663.37 | 232.64 |
| определяться | be affected | 484.88 | 157.41 |
| определение | definition | 520.01 | 319.33 |
| неопределенность | uncertainty | 43.56 | 28.85 |
| неопределенный | uncertain, undefined | 18.27 | 38.97 |

The frequency analysis shows that the reason for the confusion may be the situation wherein a word has several high-frequent derivates with only some of them annotated as meaningful. The analyzed wordlist contains another derivates *основа* and *основание*, causing confusions in automatic annotations. This is another piece of evidence proving the impact of high-frequent derivates on the model**’**s accuracy.

**Conclusion**

In our paper, we report the results of a deep learning experiment intended at training a network to detect and classify recurrent phrases, representing linguistic realizations of research-specific communicative functions. We adopted the list of communicative functions proposed for Russian academic texts in pedagogically-oriented studies for this task. Although the initial classification contains 14 classes, for the experimental setup, we chose the four biggest and the most elaborated categories, such as a «phenomenon being exhibited», «causal nexus», «change», «summarization». We also established the zero-class “O”, including non-specified lexical units. In deep learning tasks, a unit of classification is usually a sentence or an extensive text excerpt. Training a neural network model to identify recurrent constructions requires adjusting the annotation process to neural network input and architecture. To provide an annotator with a delineated view on such a vague unit as a construction we designed the guidelines, providing the list of keywords, establishing a construction, and recommendations on extended context highlighting the communicative function and resolving homonymy. The starting point for building a list of keywords were lexemes, proposed in the classification, and the similar words, recommended by the SketchEngine. We also added a few discipline-specific lexemes found during the bottom-up annotation procedure. The recommended list of the key lexemes for each category provided a rigid framework and a high level of inter-annotator agreement, while the rule of marking semantically congruent extended context left a margin for intuitive understanding of construction, covering recurrent key lexemes and variable words. Cohen’s Kappa, indicating the general inter-annotator agreement, was 0.82, which is considered a sufficient and even good level for machine learning annotation of extended textual units with variable context.

A few neural network models were trained based on the manually annotated dataset for the automatic classification task. The best F1-score metrics of 0.80−0.90, depending on the class, were obtained for the model, which deployed the contextualized and lemmatized ELMo embeddings and the LSTM architecture. The linguistic exploration of the automatically annotated data revealed that the model distinguishes very well the key phrases in an extended context, such as зависеть от, отражаться на, etc. The model also rarely confuses meaningful classes and classifies correctly formally ambiguous constructions, for example *отражаться на* as causal nexus and *отражаться в* as the manifestation of a phenomenon.

The error analysis showed that the most confusions made by the model concern the zero class “O”, covering non-classified contexts. This means that the model does not recognize significant units in some cases. The close-reading analysis revealed that the issue concerns specific lexemes (*определять*) with multiple frequent derivates because they establish linguistic noise for the contextualized ELMo embeddings. By adding more classes and elaborating fine-graded annotation techniques , it seems possible to eliminate the data noise, as more lexemes and contexts will be classified, and thus the model will be provided with more precise contextual clues.

The experiment demonstrated that deep learning methods provide an efficient instrument for automatic detection and classification of academic phrases with an overall accuracy of 87%. However, the task demands a fine-graded annotation based on rule-based techniques and predefined lists of lexemes to guarantee a high level of inter-annotator agreement, homonymy and elimination of linguistic noise.

**Reference**

Archer, D., Culpeper, J., Davies, M., 2008, Pragmatic annotation In Lüdeling, A. & Kytö, M. (eds), Corpus Linguistics: An International Handbook. De Gruyter, pp. 613-614.

Becker, M., Bender, M., & Müller, M., 2020, „Classifying heuristic textual practices in academic discourse: A deep learning approach to pragmatics‟, International Journal of Corpus Linguistics, vol. 25, no. 4, pp. 426-460.

Biber, D, 2006, University language: A corpus-based study of spoken and written registers, Studies in Corpus Linguistics 23, Amsterdam/Philadelphia, John Benjamins Publishing Company.

Biber, D., & Barbieri, F., 2007, Lexical bundles in university spoken and written registers In English for specific purposes, 26(3), pp. 263-286.

Biber, D., Conrad, S., & Cortes, V., 2004, „If you look at…: Lexical bundles in university teaching and textbooks‟, Applied linguistics, vol. 25, no. 3, pp. 371-405.

Biber, D., Connor, U., & Upton, T. A., 2007, Discourse on the Move: Using corpus analysis to describe discourse structure, Amsterdam: John Benjamins.

Biber, D., & Gray, B., 2010, „Challenging stereotypes about academic writing: Complexity, elaboration, explicitness‟, Journal of English for Academic Purposes, 9(1), pp. 2-20.

Biber, D., & Gray, B., 2016, Grammatical complexity in academic English: Linguistic change in writing, Cambridge University Press.

Bunton, D., 2002. Generic moves in Ph.D. thesis introductions In Flowerdew J. (Ed.) Academic discourse, Longman, pp. 57-75.

Bunton, D., 2005. „The structure of PhD conclusion chapters‟. Journal of English for academic purposes, 4(3), pp. 207-224.

Cortes, V. 2004, „Lexical bundles in published and student disciplinary writing: Examples from history and biology‟. English for specific purposes, 23(4), pp. 397-423.

Cortes, V., 2013, „The purpose of this study is to: Connecting lexical bundles and moves in research article introductions‟, Journal of English for Academic Purposes, vol. 12, pp. 33-43.

Durrant, P. & Mathews-Aydinli, J., 2011, „A function-first approach to identifying formulaic language in academic writing‟, English for Specific Purposes, vol. 30, pp. 58–72.

Flowerdew, J., & Forest, R, 2009, Schematic structure and lexico-grammatical realization in corpus-based genre analysis: The case of research in the PhD literature review. In Academic writing: At the interface of corpus and discourse, 15-36.

Flowerdew, J., 2014, Academic discourse, Routledge.

Gritsenko, D., Wijermars, M., & Kopotev, M. 2021, The Palgrave Handbook of Digital Russia Studies, Springer Nature.

Haotong, W., Lepage, Y., Chooi-Ling, G., 2020, „Typicality of Lexical Bundles in Different Sections of Scientific Articles‟, 2nd Symposium on Signal Processing Systems, China, pp. 56-60.

Hyland, K., 2008, „Genre and academic writing in the disciplines‟, Language Teaching, vol. 41, no. 4, pp. 543-562.

Hyland, K., 2009, Academic discourse: English in a global context, A&C Black.

Hyland, K., & Jiang, FK, 2018, „“In this paper we suggest”: Changing patterns of disciplinary metadiscourse‟. English for Specific Purposes, 51, 18-30.

Ide, N., & Pustejovsky, J. (eds.), 2017, Handbook of linguistic annotation, vol. 1, Berlin, Springer.  
Iwatsuki, K. & Aizawa, A., 2018, „Using Formulaic Expressions in Writing Assistance Systems‟, The 27th International Conference on Computational Linguistics, USA, pp. 2678–2689.

Iwatsuki, K., Boudin, F., Aizawa, A., 2020, „An Evaluation Dataset for Identifying Communicative Functions of Sentences in English Scholarly Papers‟, Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, 1712–1720.

Iwatsuki, K., Boudin, F., & Aizawa, A. 2022. Extraction and evaluation of formulaic expressions used in scholarly papers. Expert Systems with Applications, 187, pp. 115840.

Jiang, F. K. & Hyland, K., 2021, ‘The goal of this analysis…’: Changing patterns of metadiscursive nouns in disciplinary writing. Lingua, 252, pp. 103017.

Klimov, A., Toldova, S., Kopotev, M., Zevakhina, N., Dmitrieva, A., Kisselev, O., Baranchikova, A., Fedorova, M., 2018, „CAT&kittens: a corpus-based text-analytic tool for Russian academic writing‟, SlaviCorp 2018 Book of Abstracts, Charles University, Prague, pp. 22-25.

Kopotev, M., Lyashevskaya, O., & Mustajoki, A. (eds.), 2017, Quantitative Approaches to the Russian Language, London, Routledge, pp. 137-157.

Kopotev M., Mustajoki A., & Bonch-Osmolovskaya A., 2021, „Corpora in Text-Based Russian Studies‟, in Gritsenko, D., Wijermars, M, & Kopotev, M. [eds.], The Palgrave Handbook of Digital Russia Studies, Palgrave Macmillan, Cham, pp. 299-317.

Lu, X., Yoon, J., & Kisselev, O., 2021, „Matching phrase-frames to rhetorical moves in social science research article introductions‟, English for Specific Purposes, vol. 61, pp. 63-83.

Mikolov, T., Chen, K., Corrado G.S., Dean J., 2013, „Efficient estimation of word representations in vector space‟, CoRR.

Mizumoto, A., Hamatani, S., & Imao, Y., 2017, „Applying the bundle–move connection approach to the development of an online writing support tool for research articles‟, Language Learning, vol. 67, no. 4, pp. 885–921.

## Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L, 2018, „Deep contextualized word representations‟, North American Chapter of the Association for Computational Linguistics.

Pho, P. D., 2008, Research article abstracts in applied linguistics and educational technology: A study of linguistic realizations of rhetorical structure and authorial stance. In Discourse studies, 10 (2), pp. 231-250.

Pivovarova, L., Kormacheva, D., & Kopotev, M., 2017, „Evaluation of collocation extraction methods for the Russian language, in Kopotev, M., Lyashevskaya, O., & Mustajoki, A. (eds.), Quantitative Approaches to the Russian Language, London, Routledge, pp. 137-157.

Pustejovsky, J. & Stubbs, A., 2013, Natural language annotation for machine learning, A guide to corpus-building for applications, O’Reilly.

Qin, J., 2014, „Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics‟, System, vol. 42, pp. 220-231.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019, „Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter‟, 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019.

Swales, J. M., 1981, Aspects of Article Introductions, The University of Michigan Press.

Swales, J. M., 1990. Genre analysis: English in academic and research settings. Cambridge University Press, Cambridge.

Swales, J. M., 2004, Research genres: Explorations and applications, Cambridge University Press, Cambridge.

Wilcock, G., 2009, Introduction to linguistic annotation and text analytics, Synthesis lectures on human language technologies, Morgan and Claypool Publishers.

Величко, А. В. (ред.), 2009, Книга о грамматике, Русский язык как иностранный, Москва, Издательство МГУ.

Клобукова Л.П., 1987, Обучение языку специальности, Москва, Издательство Московского университета.

Митрофанова О.Д., 1985, Научный стиль речи: проблемы обучения, 2-е издание, переработанное и дополненное.

**Digital Recourses**

Deep Pavlov <https://deeppavlov.ai/>

RNC – The Russian National Corpus https://ruscorpora.ru/new/index.html (accessed 05.12.2021)

SketchEngine <https://www.sketchengine.eu/>

Spacy <https://spacy.io/>

**Acknowledgements**

This paper reports the results of the educational and research project, prepared within the framework of the Academic Fund Program at the HSE University in 2021 [grant № 21-04-059]

**Contact details**

Ekaterina A. Vlasova

HSE University (Moscow, Russia). Faculty of Humanities, School of Linguistics. E-mail: evlasova@hse.ru, tel. +7 926 7877708

**Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE University.**

**© Vlasova, Albitskiy, Bogoiavlenskaia, Glebkina, Zevakhina, Ponkratova, Sidorov, Slastnikov, Sherstyuk, 2021**

1. HSE University, Faculty of Humanities, School of Linguistics, Associate Professor [evlasova@hse.ru](mailto:evlasova@hse.ru) [↑](#footnote-ref-1)
2. HSE University, Faculty of Humanities, School of Linguistics, master’s degree student, [poalbitskiy@edu.hse.ru](mailto:poalbitskiy@edu.hse.ru) [↑](#footnote-ref-2)
3. HSE University, Faculty of Humanities, School of Linguistics, master’s degree graduate 2021 [aleksandra.bogoiavlenskaia@uit.no](mailto:aleksandra.bogoiavlenskaia@uit.no) [↑](#footnote-ref-3)
4. HSE University, Faculty of Humanities, School of Linguistics, master’s degree graduate 2021 [inna.glebkina@gmail.com](mailto:inna.glebkina@gmail.com) [↑](#footnote-ref-4)
5. HSE University, Faculty of Humanities, School of Linguistics, [nzevakhina@hse.ru](mailto:nzevakhina@hse.ru) [↑](#footnote-ref-5)
6. HSE University, Faculty of Humanities, School of Linguistics, dekalisto18@yandex.ru [↑](#footnote-ref-6)
7. [HSE Tikhonov Moscow Institute of Electronics and Mathematics (MIEM HSE)](https://miem.hse.ru/en/) /[School of Applied Mathematics](https://miem.hse.ru/en/edu/pm/), Associate Professor, sslastnikov@hse.ru [↑](#footnote-ref-7)
8. [HSE Tikhonov Moscow Institute of Electronics and Mathematics (MIEM HSE)](https://miem.hse.ru/en/) /[School of Applied Mathematics](https://miem.hse.ru/en/edu/pm/), master’s degree student, mister.loonatik@yandex.ru [↑](#footnote-ref-8)
9. HSE University, Faculty of Humanities, School of Linguistics, postgraduate student [aline.savelieva@](mailto:aline.savelieva@)gmail.com [↑](#footnote-ref-9)