RUSSIAN JOURNAL OF LINGUISTICS

2022 Volume 26 No. 2

Computational Linguistics and Discourse Complexology

Guest Editors
Danielle MCNAMARA, Valery SOLOVYEV and Marina SOLNYSHKINA

Компьютерная лингвистика и дискурсивная комплексология

Приглашенные редакторы
Д.С. МАКНАМАРА, В.Д. СОЛОВЬЕВ, М.И. СОЛНЫШКИНА

Founded in 1997
by the Peoples' Friendship University of Russia (RUDN University)

Научный журнал

Издается с 1997 г.

Издание зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор) **Свидетельство о регистрации** ПИ № ФС 77-76503 от 02.08.2019 г. **Учредитель:** Федеральное государственное автономное образовательное учреждение высшего образования «Российский университет дружбы народов»

DOI: 10.22363/2687-0088-2022-26-2

RUSSIAN JOURNAL OF LINGUISTICS

ISSN 2687-0088 e-ISSN 2686-8024

Publication frequency: quarterly.
Languages: Russian, English.
Indexed/abstracted in Scopus, Web of Science Core Collection (ESCI), RSCI, DOAJ, Ulrich's Periodicals Directory: http://www.ulrichsweb.com, Electronic Journals Library Cyberleninka, Google Scholar, WorldCat.

Aims and Scope

The Russian Journal of Linguistics is a peer-reviewed international academic journal publishing research in Linguistics and related fields. It is international with regard to its editorial board, contributing authors and thematic foci of the publications.

The aims of the journal:

- to promote scholarly exchange and cooperation among Russian and international linguists and specialists in related areas of investigation;
- to disseminate theoretically grounded research and advance knowledge pertaining to the field of Linguistics developed both in Russia and abroad;
- to publish results of original research on a broad range of interdisciplinary issues relating to language, culture, cognition and communication;
- to cover scholarly activities of the Russian and international academia.

As a Russian journal with international character, it aims at discussing relevant intercultural/linguistic themes and exploring general implications of intercultural issues in human interaction in an interdisciplinary perspective. The most common topics include *language and culture, comparative linguistics, sociolinguistics, psycholinguistics, cognitive linguistics, pragmatics, discourse analysis, intercultural communication, and theory and practice of translation*. In addition to research articles, the journal welcomes book reviews, literature overviews, conference reports and research project announcements.

The Journal is published in accordance with the policies of *COPE* (*Committee on Publication Ethics*) **http://publicationethics.org**.

The editors are open to thematic issue initiatives with guest editors.

Further information regarding notes for contributors, subscription, open access and back volumes is available at **http://journals.rudn.ru/linguistics**.

E-mail: **lingj@rudn.ru**

---

4 выпуска в год.
Языки: русский, английский.
Входит в перечень рецензируемых научных изданий ВАК РФ.
Включен в каталог периодических изданий Scopus, Web of Science Core Collection (ESCI), RSCI, DOAJ, Ульрих (Ulrich's Periodicals Directory: http://www.ulrichsweb.com).
Материалы журнала размещаются на платформе РИНЦ Российской научной электронной библиотеки, Electronic Journals Library Cyberleninka, Google Scholar,WorldCat.
Подписной индекс издания в каталоге агентства Роспечать: 36436.

Цели и тематика

Журнал Russian Journal of Linguistics – периодическое международное рецензируемое научное издание в области междисциплинарных лингвистических исследований. Журнал является международным как по составу редакционной коллегии и экспертного совета, так и по авторам и тематике публикаций.

Цели журнала:

- способствовать научному обмену и сотрудничеству между российскими и зарубежными лингвистами, а также специалистами смежных областей;
- знакомить читателей с новейшими направлениями и теориями в области лингвистических исследований, разрабатываемых как в России, так и за рубежом, и их практическим применением;
- публиковать результаты оригинальных научных исследований по широкому кругу актуальных лингвистических проблем междисциплинарного характера, касающихся языка, культуры, сознания и коммуникации;
- освещать научную деятельность как российского, так и международного научного сообщества.

Будучи международным по своей направленности, журнал нацелен на обсуждение теоретических и практических вопросов, касающихся взаимодействия культуры, языка и коммуникации. Особый акцент делается на междисциплинарные исследования. Основные рубрики журнала: *язык и культура, сопоставительное языкознание, социолингвистика, психолингвистика, когнитивная лингвистика, прагматика, анализ дискурса, межкультурная коммуникация, теория и практика перевода*. Кроме научных статей публикуется хроника научной жизни, включающая рецензии, научные обзоры, информацию о конференциях, научных проектах.

Перечень отраслей науки и групп специальностей научных работников в соответствии с номенклатурой ВАК РФ: Отрасль науки: 10.00.00 – филологические науки; Специальности научных работников: 10.02.01 – русский язык, 10.02.04 – германские языки, 10.02.05 – романские языки, 10.02.19 – теория языка, 10.02.20 – сравнительно-историческое, типологическое и сопоставительное языкознание.

Журнал строго придерживается международных стандартов публикационной этики, сформулированных в документе *COPE* (*Committee on Publication Ethics*) **http://publicationethics.org**.

Правила оформления статей, архив и дополнительная информация размещены на сайте: **http://journals.rudn.ru/linguistics**.

Электронный адрес: **lingj@rudn.ru**

Подписано в печать 12.05.2022. Выход в свет 28.06.2022. Формат 70×108/16.
Бумага офсетная. Печать офсетная. Гарнитура «Times New Roman».
Тираж 500 экз. Заказ № 533. Цена свободная.
Отпечатано в типографии ИПК РУДН: 115419, Москва, Россия, ул. Орджоникидзе, 3

Printed at the RUDN Publishing House: 3, Ordzhonikidze str., 115419 Moscow, Russia,
+7 (495) 952-04-41; E-mail: publishing@rudn.ru

Review editor ***Konstantin V. Zenkin***
English language editor ***Julia B. Smirnova***
Computer design ***Natalia A. Yasko***

Editorial office:
10/2 Miklukho-Maklaya str., 117198 Moscow, Russia
Tel.: +7 (495) 434-20-12;
e-mail: lingj@rudn.ru

Литературный редактор ***К.В. Зенкин***
Редактор англоязычных текстов ***Ю.Б. Смирнова***
Компьютерная верстка ***Н.А. Ясько***

Адрес редакции:
115419, Москва, Россия, ул. Орджоникидзе, д. 3
Тел.: (495) 955-07-16; e-mail: publishing@rudn.ru

Почтовый адрес редакции:
117198, Москва, Россия, ул. Миклухо-Маклая, д. 10/2
Тел.: (495) 434-20-12; e-mail: lingj@rudn.ru

Computational Linguistics and Discourse Complexology

CONTENTS

Book reviews

Компьютерная лингвистика и дискурсивная комплексология

СОДЕРЖАНИЕ

Рецензии

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 275–316
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30161

Research article

Computational linguistics and discourse complexology: Paradigms and research methods

Valery SOLOVYEV[1] ✉, Marina SOLNYSHKINA[1] and Danielle MCNAMARA[2]

[1]Kazan (Volga Region) Federal University, Kazan, Russia
[2]Arizona State University, Tempe, USA
✉maki.solovyev@mail.ru

Abstract
The dramatic expansion of modern linguistic research and enhanced accuracy of linguistic analysis have become a reality due to the ability of artificial neural networks not only to learn and adapt, but also carry out automate linguistic analysis, select, modify and compare texts of various types and genres. The purpose of this article and the journal issue as a whole is to present modern areas of research in computational linguistics and linguistic complexology, as well as to define a solid rationale for the new interdisciplinary field, i.e. discourse complexology. The review of trends in computational linguistics focuses on the following aspects of research: applied problems and methods, computational linguistic resources, contribution of theoretical linguistics to computational linguistics, and the use of deep learning neural networks. The special issue also addresses the problem of objective and relative text complexity and its assessment. We focus on the two main approaches to linguistic complexity assessment: "parametric approach" and machine learning. The findings of the studies published in this special issue indicate a major contribution of computational linguistics to discourse complexology, including new algorithms developed to solve discourse complexology problems. The issue outlines the research areas of linguistic complexology and provides a framework to guide its further development including a design of a complexity matrix for texts of various types and genres, refining the list of complexity predictors, validating new complexity criteria, and expanding databases for natural language.

Keywords: *computational linguistics, linguistic complexology, discourse complexology, text complexity, machine learning, natural language processing*

For citation:
Solovyev, Valery, Marina Solnyshkina & Danielle McNamara. 2022. Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics* 26 (2). 275–316. https://doi.org/10.22363/2687-0088-30161

Научная статья

Компьютерная лингвистика и дискурсивная комплексология: парадигмы и методы исследований

В.Д. СОЛОВЬЕВ[1] ✉, М.И. СОЛНЫШКИНА[1], Д.С. МАКНАМАРА[2]

[1]Казанский федеральный университет, Казань, Россия
[2]Университет штата Аризона, Темпе, США
✉maki.solovyev@mail.ru

Аннотация
Важнейшей особенностью современных исследований является значительное расширение научной проблематики и повышение точности расчетов лингвистического анализа за счет способности искусственных нейронных сетей к обучению и возможности не только автоматизировать лингвистический анализ, но и решать задачи отбора, модификации и сопоставления текстов различных типов и жанров. Цель данной статьи, как и выпуска в целом, – представить некоторые направления исследований в области компьютерной лингвистики и лингвистической комплексологии, а также обосновать целесообразность выделения новой междисциплинарной области – дискурсивной комплексологии. В обзоре трендов компьютерной лингвистики делается акцент на следующих аспектах исследований: прикладные задачи, методы, компьютерные лингвистические ресурсы, вклад теоретической лингвистики в компьютерную, применение нейронных сетей глубокого обучения. Особое внимание в спецвыпуске уделено вопросам оценки объективной и относительной сложности текста. Выделяются два основных подхода к решению проблем лингвистической комплексологии: «параметрический подход» и машинное обучение, прежде всего, нейронные сети глубокого обучения. Исследования, публикуемые в специальном выпуске, показали не только высокую значимость методов компьютерной лингвистики для развития дискурсивной комплексологии, но и расширение методологических находок компьютерной лингвистики, используемых для решения новых задач, стоящих перед комплексологами. Они высветили основные проблемы, стоящие перед отечественной лингвистической комплексологией, и наметили направления дальнейших исследований: создание матрицы сложности текстов различных типов и жанров, расширение списка предикторов сложности, валидация новых критериев сложности, расширение баз данных для естественного языка.

Ключевые слова: *компьютерная лингвистика, лингвистическая комплексология, дискурсивная комплексология, сложность текста, машинное обучение, обработка естественного языка*

Для цитирования:
Solovyev V., Solnyshkina M., McNamara D. Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics.* 2022. T. 26. № 2. P. 275–316. https://doi.org/10.22363/2687-0088-30161

1. Introduction

The article addresses modern trends in computational linguistics, language and discourse complexity. It also provides a brief overview of the articles in the issue.

Computational linguistics (hereinafter CL), as the name implies, is an interdisciplinary science at the intersection of linguistics and computer sciences. It explores the problems of automatic processing of linguistic information. Another commonly used name for this discipline, that is synonymous with the term "computational linguistics", is Natural Language Processing (NLP). In a number of research works these concepts are separated, considering that CL is more of a theoretical discipline, and NLP is of a more applied nature. CL began to develop in the early 1950s, almost immediately after the advent of computers. Its first task was development of machine translation, and translation of journals from Russian into English in particular. The initial stage of CL development is comprehensively presented in J. Hutchins (1999). It surely was beyond the capacity of researchers to solve the problems of machine translation very quickly, and the initial optimism turned out to be groundless, although in recent years it has become possible to obtain translations of acceptable quality. However, within 70 years of development, CL has achieved significant success in solving many urgent practical problems, which made it one of the most dynamically developing and important research areas in both linguistics and computer science. In our opinion, the best monographs on CL are (Clark et al., Indurkhya & Damerau 2010). The latest review, including also an analysis of the prospects for its development, can be found in the article by Church and Liberman (2021).

In the review of computational linguistics trends, we focus on the following aspects of research: application-oriented tasks, methods, resources, contribution of theoretical linguistics to computer linguistics, and application of deep learning neural networks. The latter appeared about 10 years ago (Schmidhuber 2015) and revolutionized research of artificial intelligence, including many areas of CL. Artificial neural networks constitute a formal model of biological networks of neurons. Their most important feature is the ability to learn; in case of an error, the neural network is modified in a certain way. Although neural networks were proposed as early as 1943, a breakthrough in their use was made only a few years ago. It is associated with the three following factors: the emergence of new, more advanced 'self-learning', unsupervised training algorithms, improved performance of computers, and Internet database increase. Advances in NLP in the late 2018 were mainly related to BERT (Devlin et al. 2018), a neural network pre-trained on a corpus of texts. Currently, BERT and its enhanced models show better performance on many NLP problems (see Lauriola, Lavelli & Aiolli 2022).

2. Applied problems and methods of computer linguistics

2.1. Application-oriented tasks of Computational Linguistics

In addition to machine translation, the main application-oriented tasks of CL include document processing, computer analysis of social networks, speech analysis and synthesis (including voice assistants), question-answering systems, and recommender systems.

The largest task is document processing including a wide range of subtasks: search, summarization, classification, sentiment analysis, information extraction, etc.

Development of search engines, obviously, is the most well-known and widely used CL task, successfully implemented in Google and Yandex search engines. A detailed introduction to the issue of information retrieval can be found in (Manning et al. 2011). The main type of search queries is a set of keywords. The two main problems of search are as follows: the need to provide fast searches in the vast amount number of texts on the Internet and to ensure that any search takes into account not the query forms only but its semantics. The main idea of a quick search is to preprocess all documents on the Internet with the creation of a so-called search index that indicates location of the query in specific documents. A semantic document search, or a semantic search, is implemented in the well-known concept of Semantic Web (Domongue et al. 2011), based on the idea of ontologies (presented below). E.g., in response to a query "*Beethoven ta ta ta tam*" Google refers to the Wikipedia article about Beethoven's 5th symphony, although the text of the article does not contain the phrase "*ta ta ta tam*". Thus, the Google search engine "understands" that "*ta ta ta tam*" and the 5th symphony are semantically related. A successful search would be simply impossible without linguistic research, which led to the development of algorithms for morphological and syntactic analysis, thesauri and ontologies for the explication of semantic relationships between entities.

The term "information retrieval" is interpreted as a search for information of a certain type in the text, i.e. entities, their relationships, facts, etc. The best developed is the algorithm of extracting named entities (Name Entity Recognition, NER), i.e. persons, organizations, geographical objects, etc. A recent survey of IT professionals from various business areas[1] indicates that the NER task is the most demanded in business applications. Researchers apply various techniques to solve this problem: ready-made dictionaries of people's names and names of geographical objects; linguistic features (use of capital letters), defined patterns of noun phrases; and machine learning methods. An overview of this area can be found in Sharnagat (2014). NER systems based on dictionaries and rules correctly extract about 90% of entities in texts, while BERT-based systems already provide about 94% of correctly extracted entities (Wang 2020), which is comparable to the level of human accuracy and demonstrates benefits of deep learning neural networks.

[1] https://gradientflow.com/2021nlpsurvey/

The task of retrieval of events and facts is challenging. The classic approach here is to create event templates that capture types and roles of the entities participating in the events. For example, the event "June 24, 2021 Microsoft presented Windows 11" is described by the following template: Activity type – sales presentation, Company – Microsoft, Product – Windows 11, Date – June 24, 2021. Templates of this type are created manually, which is labour-intensive. Efficiency of information extraction systems depends on their quality. Typically, such systems extract no more than 60% of facts (Jiang et al. 2016).

In recent years, many studies addressed the problem of text sentiment analysis (cf. Cambria 2017), i.e. identification of the so-called "tone" of texts: whether a text carries a positive or negative attitude towards the text referents. This area is important for companies to evaluate user comments on their products and services. The problem is also being solved with the help of developing specific patterns, dictionaries, and machine learning methods. The Russian dictionary RuSentiLex, (Loukachevitch & Levchik 2016), registers over 12,000 lemmas marked as positive, negative or neutral. The main problem of sentiment analysis of texts is its context-dependency as a word can be positive in certain contexts and negative in others. A possible way of addressing the problem is compiling sentiment lexicon dictionaries for specific subject areas.

Another fundamental problem is not only to assess the tone of the entire text, but define the referential aspect of the sentiment. It is especially important in applied research on customer reviews of products and services (Solovyev & Ivanov 2014). The achieved accuracy in the area, which is about 85%, was effected through BERT technology (Hoang et al. 2019).

Another important task of document processing is text summarization and text skimming (Miranda-Jiménez, Gelbukh & Sidorov 2013). Its practical importance is determined by the gigantic and increasing size of texts on the Internet. There are two approaches to solving this problem: extractive and abstract. The extractive approach –implies assessing the importance score of sentences in the text and selecting a small number of the most significant ones. It requires non-trivial mathematical methods to evaluate informational hierarchy of text parts. The abstract, approach implies a generation of original sentences that summarize the content of the source text. In recent years the task of generating text abstracts was successfully fulfilled with neural networks. An important component of summarization systems are sentence parsing algorithms. A brief overview is provided in Allahyari (2017).

Computer analysis of social networks and social media is another application-oriented task. It can have multiple objectives with monitoring social attitudes, identifying manifestations of extremism and other illegal activities, and even analyzing the spread of epidemics. E.g. at the beginning of the coronavirus pandemic researchers suggested an analysis of social media content, including the spread of misinformation (cf. Cinelli, Quattrociocchi & Galeazzi 2020). Social network analysis implies defining the content of messages and connections between

users, which enables identifying groups of users with common interests. At the same time, heterogeneity of content presents a significant challenge. In recent years, neural networks have become the main tool for social network analysis (cf. Ghani et al. 2019). Batrinca & Treleaven (2015) provide an overview of the research in the area and addresses mostly humanitarians.

Speech analysis and synthesis stand apart in CL, as they require specific software and hardware tools to work with acoustic signals. Speech recognition systems are very diverse and are classified according to many parameters: vocabulary size; speaker type (age, gender); type of speech; purpose; structural types and their selection principles (phrases, words, phonemes, diphones, allophones, etc.). The input speech flow is compared with acoustic and language models, including various features: spectral-temporal, cepstral features, amplitude-frequency, features of nonlinear dynamics. Speech recognition is challenging because words are pronounced differently by different people in different situations. Nevertheless, at the moment there are many commercial speech recognition systems, in particular those built into Windows. One of the best known is "Watson speech to text" developed by IBM (Cruz Valdez 2021).

Speech recognition is the heart of voice assistants becoming increasingly popular worldwide. A voice assistant commonly known in Russia is Alice[2] designed and developed by Yandex. Alice is integrated into the Yandex services: by a voice command it searches for information. E.g. it can find a weather forecast on Yandex.Weather, traffic data in Yandex.Maps, etc. Alice can control smart home systems and even entertain: play riddles with children, tell fairy tales and jokes. Speech recognition in voice assistants is facilitated by their ability to tune in to the voice of a certain person. State of the art review in voice assistants can be found in Nasirian, Ahmadian & Lee (2017), and one of the latest reviews of speech recognition problems is presented in Nassif (2019).

Speech synthesis is being actively used in information and reference systems, in airport, railway and office announcements. They are predominantly used in situations with a limited range of synthesized phrases. The simplest way to synthesize speech is sequencing pre-recorded elements. The quality of the synthesize speech is evaluated based on its similarity with human speech. High-quality speech synthesis systems are still a dream of many researchers and users. The latest overview of speech synthesis is presented in Tan (2021).

We also address recommender systems which are probably familiar to all Internet users. Recommender systems predict which objects (movies, music, books, news, websites) might be interesting to a particular user. For this, they collect information about users, sometimes explicitly, asking them to rate objects of interest, and more often implicitly, collecting information about users' behavior on the Internet. The following idea turned out to be productive: people who similarly estimated some objects in the past are most likely to give similar estimates to other objects in the future (Xiaoyuan & Khoshgoftaar 2009). This particular idea allows

[2] https://dialogs.yandex.ru/store

researchers to effectively extrapolate user behavior. Developing recommender systems depends mostly on linguistic resource. For example, an effective recommender system is based on synonyms dictionaries. Such systems are supposed to "understand" that "children's films" and "films for children" mean the same. For synonymy in recommender systems, see Moon (2019), a general review is presented in Patel & Patel (2020).

Question-answering systems, or QA-systems, are designed to provide answers in natural language, i.e. they have a natural language interface. They search for answers in a textual database that QA systems have. Like search engines, QA systems provide a user with the ability to search for information. However, an important distinguishing feature of QA systems is that they allow a user to find information that might be implicit, e.g., a film that a user might like but it could not be found with a regular search engine. Obviously, the quality of a QA system depends on its database size, i.e. whether it contains an answer to a question at all, as well as on the technologies for processing questions and comparing them with the database information. As for processing a question, it begins with identifying the type of question and the expected response. For example, the question "Who..." suggests that the answer is to contain the name of a person. QA systems apply numerous complex CL methods and, similar to recommender systems, face the issue of synonymy (Sigdel 2020). The latest review of QA systems is published by Ojokoh and Adebisi (2018).

2.2. Methods of Computational Linguistics

All CL methods can be divided into two large classes: a class based on dictionaries and rules (templates) and a class based on machine learning. These two classes are fundamentally different in their approaches. Dictionaries and rules use accumulated knowledge about the language, as well as results of highly professional manual labor, and therefore they are extremely expensive. Machine learning is implemented on a large number of examples, presented in annotated corpora which function as training sets. The algorithm implies analyzing training sets, identifying the existing patterns and then offering solutions to the problems set. Modern machine learning systems vary in their functions and applications, although deep learning neural networks have proved to be the most efficient. At an input node of a neural network, any language data is fed in encoded forms as tokens: letters, bigrams, short high-frequency morphemes, and words.

Application of this approach depends on a large body of annotated texts at a researcher's disposal: the larger the training set, the better the neural network will learn. At the same time, annotation is quite simple and its implementation does not necessarily involve professional linguists as researchers can refer to services of native speakers.

In this article, we will focus on the basic methods of CL and refer readers to the above-mentioned monographs for a detailed review of the area (cf. Clark et al. 2013, Indurkhya & Damerau 2010).

Automatic text analysis usually begins with its pre-processing which includes text segmentation, i.e. segmentation into words and sentences. Though it may seem like a simple task, since words are separated from each other by spaces and sentences begin with a capital letter and end with a period (rarely, exclamation marks, question marks, ellipsis) followed by a space. The most typical example of the rule or pattern is the following: a period – space – capital letter. However, it is not that simple. A period can be in the middle of a sentence after the first initial, followed by a space and then a capitalized second initial. Here, the period does not explicitly indicate the division of the text into sentences. As an example, we can refer to the following sentence: "Lukashevich N.V., Levchik A.V. Creation of a lexicon of evaluative words of the Russian language RuCentilex // Proceedings of the OSTIS-2016 conference. pp. 377–382". Despite all the difficulties, the segmentation problem is considered to be practically solved. In 1989, Riley (1989) managed to achieve a 99.8% accuracy rate for splitting texts into sentences. To achieve this result, the researcher developed a complex system of rules taking into account the following features: length of the word before the dot, length of the word after the dot, presence of a word before the dot in the dictionary of abbreviations, etc.

The next step in the course of text analysis is morphological. Consider, as an example, a language with complex morphology – Russian. For the Russian language, morphological analysis is performed by a number of analyzers: MyStem, Natasha, pymorphy2, SpaCy, etc. In CL, morphological analysis, the purpose of which is to determine the morphological characteristics of a word, is based on a detailed description of inflectional paradigms. For the Russian language, a reference book of this kind is Zaliznyak (1977), which presents paradigm indices of almost 100,000 lemmas of the Russian language. The presence of such a directory made it possible to generate about 3 mln Word forms for the registered lemmas of the Russian language. Automatic text analysis finds a lemma corresponding to any word form and a complete list of morphological characteristics. The main challenge for the existing analyzers is homonymy, which the available parsers have not solved yet. And in situations when users require not all parsing options but one, analyzers produce the variant of morphological parsing of the highest frequency, still ignoring senses of the word in the context.

Another problem is parsing of the so-called "off-list" words, i.e. words not registered in the dictionary. Given that the average number of such words is about 3%, their morphological analysis requires developing special algorithms. The simplest solution foreseen is the following: based on the analysis of its flexion, the off-list word is assigned its morphological paradigm.

Syntactic parsing, or parsing, is much more complex. The result of syntactic parsing of a sentence is a dependency tree that presents a sentence structure either in the formalism of a generative grammar or in the formalism of a dependency grammar (cf. Tesnière 2015). Parsing requires a detailed description of the syntax of the language. The most successful analyzer for the Russian language is ETAP

developed by the Laboratory of Computational Linguistics of the Institute for Information Transmission Problems of the Russian Academy of Sciences as a result of over 40 years of research. Its latest version, ETAP-4, is available at (ENA, June 6, 2020)[3]. ETAP parser is based on the well-known model "Meaning ⇔ Text" (Melchuk 1974), its formalized version is described in the monograph by Apresyan (1989).

In the recent decade, parsing has also been performed by neural networks (cf. Chen & Manning 2014) trained on syntactically annotated corpora. English Penn Treebank (ENA, June 6, 2022)[4] is used for English. For the Russian language, one can use SynTagRus (ENA, June 6, 2022)[5], developed by the Laboratory of Computational Linguistics at the Institute for Information Transmission Problems RAS.

The task of semantic analysis is even more difficult. However, if we want the computer to "understand" the meaning, it is necessary to formalize semantics of words and sentences. The problem is solved in two classical ways. The first was initiated by C. Fillmore (1968), who introduced concepts of semantic cases or roles of noun phrases in a sentence. The correct establishing of semantic roles is an important step towards sentence comprehension. Fillmore's original ideas were realized in FrameNet lexical database (ENA, June 6, 2022)[6].

The second approach was implemented in an electronic thesaurus, or lexical ontology, WordNet (Fellbaum 1998) which was originally designed for the English language. Subsequently its analogues were developed for many languages. There are numerous analogues of WordNet for the Russian language, the most effective and being widely used is RuWordNet thesaurus (ENA, June 6, 2022)[7], (cf. Loukachevitch & Lashevich 2016), comprising over 130,000 words. WordNet-like thesauri explicate semantic relationships between words (concepts) including synonymy, hyponymy, hypernymy, etc., and their systemic parameters partially define their semantics. WordNet has been successfully implemented in a large number of both linguistic and computer research.

The idea of vector representation of semantics, i.e. word embeddings, has been proposed recently. Its core is constituted by the distributive hypothesis: linguistic units occurring in similar contexts have similar meanings (Sahlgren 2008). This hypothesis has been confirmed in numerous studies aimed at defining frequency vectors of words registered in large text corpora. There are multiple refinements and computer implementations of the idea, the most popular of which is word2vec (Mikolov et al. 2013) available in Gensim library (ENA, June 6, 2022)[8]. RusVectores system (Kutuzov & Kuzmenko 2017), available at (ENA, June 6,

[3] http://proling.iitp.ru/ru/etap4
[4] https://catalog.ldc.upenn.edu/LDC99T42
[5] https://universaldependencies.org/treebanks/ru_syntagrus/index.html
[6] https://framenet.icsi.berkeley.edu/fndrupal/
[7] https://ruwordnet.ru/ru
[8] https://github.com/rare-technologies/gensim

2022)[9] identifies vector semantics for Russian words. Specifically, RusVectores evaluates semantic similarity of words.

Obviously, the most important tool for research in CL, as indeed in all modern linguistics, are text corpora. The first corpus compiled in the 1960s was *Brown Corpus* which when released contained one million words. Since then, corpora size requirements have increased dramatically. For the Russian language, the most well known is the National Corpus of the Russian Language (NCRL, ENA, June 6, 2022[10]). Created in 2004, it is being constantly updated and currently includes over 600 mln words. In 2009, Google compiled and uploaded a very interesting multilingual resource, i.e. Google Books Ngram (ENA, June 6, 2022)[11], containing 500 bln words, 67 bln words of which constitute the Russian sub-corpus (cf. Michel 2011).

Another important problem is corpus annotation or tagging, which in difficult cases is done manually. The work is usually carried out by several annotators and their performance consistency is closely monitored (Pons & Aliaga 2021). Despite the fact that corpora have become an integral part of linguistic research, there have been ongoing disputes on their representativeness, balance, differential completeness, subject and genre relatedness, as well as data correctness (cf. Solovyev, Bochkarev & Akhtyamova 2020).

Thus, thanks to CL, researchers fully implement numerous services including information retrieval, automatic error correction, etc. This became possible due to fundamentally important accomplishments not only in computer science, but also in linguistics. CL uses extensive dictionaries and thesauri, detailed syntax models, and giant corpora of texts. Automatic morphological analysis in its modern form would not exist without A. A. Zaliznyak's "Dictionary of the Russian Language Grammar" (1977). Multiple studies in CL are based on manually created WordNet and RuWordNet thesauri. Computer technologies, in turn, contribute to the development of linguistics. Text corpora and statistical methods have already become commonplace; without them serious linguistic research would be impossible.

All key CL technologies are publicly available, e.g. (ENA, June 6, 2022)[12] houses programs to solve numerous basic tasks for numerous languages.

It is not really feasible to cover all the topics of CL, a vast and rapidly developing field of linguistics, in one article. Many important questions have been left beyond. We refer readers interested in the topics of co-reference resolution, disambiguation, topic modeling, etc. to the above-mentioned publications.

[9] https://rusvectores.org/ru/
[10] https://ruscorpora.ru/new/
[11] https://books.google.com/ngrams
[12] https://stanfordnlp.github.io/CoreNLP/

3. Complexity of language and text as a research problem

The core of the special issue is made up of the articles focused on text complexity assessment. At first glance, estimating language complexity based on the number of categories in its system seems to be very logical, and the task itself appears feasible. A good example of the idea can be a phonological inventory of the language, the number of morphophonological rules or verb forms. Obviously, in this case, it becomes possible to compare complexity of different languages and assign them to some objective, absolute complexity (Miestamo, Sinnemäki & Karlsson 2008). Notably, it is the "objective" complexity that is significant when mastering a non-native language. On the other hand, if a language is acquired as a native language, it does not present any difficulty for children, and from this point of view, all languages complexity is absolutely the same. Researchers admit that language and text complexity "resists measurement", and scholars working in this field face conceptual and methodological difficulties.

Significant in the light of the problems under study is the description of the relationship and interdependence of two areas of complexity studies: language, or 'lingue' complexity, i.e. linguistic complexology, on the one hand, and text or discourse, 'parole' complexity, i.e. discursive complexology, on the other.

The interpretation of the very concept of "language (lingue) complexity" changed dramatically in the 19th-20th centuries. In the 19th century, the Humboldtian theory on interdependence between the structure of a language and stage of development of people speaking this language was universally accepted (Humboldt 1999: 37). Acknowledging this concept, researchers actually acknowledge unequal status of languages and peoples. In the XXth century, the Humboldian views asserting inequality of languages and their speakers were replaced by the concept of the so-called single complexity, identical and equal for all languages of the world. The idea received two names: ALEC — "All Languages are Equally Complex" (Deutscher 2009: 243) and linguistic equi-complexity dogma (Kusters 2003: 5). Researchers who support the idea are to prove two hypotheses: (1) language complexity is constituted of sub-complexities of its elements; (2) all sub-complexities in linguistic subsystems are compensated: simplicity in area A is compensated by complexity in area B, and vice versa ("compensatory hypothesis"). Arguing the concept "All languages are equally complex", Ch. Hockett quite boldly stated: "Objective measurement is difficult, but impressionistically it would seem that the total grammatical complexity of any language, counting both the morphology and syntax, is about the same as any other. This is not surprising, since all languages have about equally complex jobs to do: and what is not done morphologically has to be done syntactically" (Hockett 1958: 180–181). Unfortunately, in the works of that period and approach, scholars discussed neither complexity criteria nor its empirical evidence. For a detailed overview of the "linguistic equi-complexity dogma", see the seminal work by J. Sampson, D. Gil, and P. Trudgill, Language Complexity as an Evolving Variable (Sampson et al. 2009).

The twenty-first century opened with a number of critical reviews of ALEC theory, on the one hand (cf. Miestamo et al. 2008), and McWhorter's provocative statement that "Creole grammars are the simplest grammars in the world" (McWhorter 2001). The very idea that all languages are equally complex has been convincingly rejected by sociolinguists, who have shown that language contact can lead to language simplification. This is shown in Afrikaans, Pidgins and Koine. Simplifying a language is possible, hence, before its simplification, the language was more complicated than after. And if a language can be more or less complex at different periods of its history, then some languages can be more complex than others (Trudgill 2012).

In the early 2000s the idea of linguistic complexity and the "dogma of equal complexity" was actively discussed at conferences and seminars (see the seminar "Language complexity as an evolving variable" organized by Max Planck Institute for Evolutionary Anthropology in 2007 in Leipzig ENA, June 6, 2022[13]), in a number of journal articles (cf. Shosted 2006, Trudgill 2004) and monographs (Dahl 2009, Kusters 2003, Miestamo et al. 2008, Sampson et al. 2009).

Publications on language complexity in Russia are predominantly reviews written by foreign scholars, although in recent years interest in the area has visibly grown. The most comprehensive are the studies conducted by A. Berdichevsky (2012) and the review of Peter Trundgill's book "Sociolinguistic Typology", 2011 by Vakhtin (2014). The problems of language complexity were also discussed at the Institute for Linguistic Research of the Russian Academy of Sciences (ILI RAS) in 2018 at the conference "Balkan Languages and Dialects: Corpus and Quantitative Studies".

Local and global complexity

The development of linguistic complexology led to the identification of two types of complexity: global, i.e. the complexity of the language (or dialect) as a whole, and local complexity, i.e. complexity of a particular level of language or domain (Miestamo 2008). And if the assessment of global complexity of a language, according to researchers, is a very ambitious and probably hopeless task which H. Deutscher compares with "chasing wild geese" (Deutscher 2009), then the measurement of local complexity is considered as a feasible task, which imples the compiling of a list and evaluating complexity predictors at various language levels. The list of predictors of *phonological complexity* traditionally includes phoneme inventory, frequency of marked [14] phonemes, tonal differences, suprasegmental patterns, phonotactic restrictions, and maximal consonant clusters (Nichols 2009, Shosted 2006). When evaluating *morphological complexity*, classical "inconvenience factors" (Braunmüller's term 1990: 627) are the size of inflectional morphology of a language (or language variety), specificity of

[13] https://www.eva.mpg.de/fileadmin/content_files/linguistics/pdf/ComplexityWS_Webpage_2007.pdf

[14] Phonemes that are rarely found in the languages of the world are considered marked (Berdichevsky 2012).

allomorph and morphophonemic processes, etc. (Dammel & Kürschner 2008, Kusters 2003). *Syntactic complexity* assessment is based on the accumulated data of syntax rules and follows the principle "the more, the more difficult", as well as language ability to generate recursions and clauses within a syntactic whole (Ortega 2003, Givón 2009, Karlsson 2009). *Semantic and lexical complexity* is estimated based on the number of ambiguous language units, the difference between inclusive and exclusive pronouns, lexical diversity, etc. (Fenk-Oczlon & Fenk 2008, Nichols 2009). *The pragmatic* or "hidden" complexity built on the law of economy is the complexity of inferences necessary to comprehend texts. Latent complexity languages allow for minimalist, very simple surface structures in which grammatical categories inferences are far from being trivial. The idea is exemplified by languages of Southeast Asia, which have achieved a particularly high degree of latent complexity. The latter is observed in the omission of pronouns and consequent multiple co-references in relative clauses, absence of relational markers, "bare" nouns lacking determiners and as such enabling a wide range of interpretations (Bisang 2009).

Research has indicated that high levels of local complexity at one level in a language do not necessarily entail low local complexity at another level, as predicted by the "dogma of equal complexity". For example, the analysis of metrics of morphological and phonological complexity in 34 languages carried out by R. Shosted did not reveal any expected statistically significant correlation (Shosted 2006). And the individual "balancing effects" (trade-offs) between local complexities observed by G. Fenk-Ozlog and A. Fenk, unfortunately, are also insufficient to validate the "dogma of equal complexity" of languages. G. Fenk-Ozlog and A. Fenk, in particular, found that in English the tendency towards phonological complexity and monosyllabicity is associated with a tendency towards homonymy and polysemy, towards a fixed word order and idiomatic speech (Fenk-Oczlon & Fenk 2008: 63). D. Gil has convincingly argued that isolating languages do not necessarily compensate for simple morphology with more complex syntax (Gil 2008).

Factors (or predictors) of language complexity are usually divided into internal and external ones. The number of elements and categories in the language, redundancy and irregularity of language categories are viewed as *the internal factors* of complexity. The modern paradigm developed the so-called "list approach" to assess internal complexity. The latter implies compiling a list of linguistic phenomena, the presence of which in a language increases its complexity. In fact, the lists of intrinsic complexity predictors are lists of the local complexity described above. For example, the complexity predictors list compiled by J. Nichols contains over 18 parameters and includes phonological, morphological, syntactic and lexical features (Nichols 2009). A language is considered more complex if it has more marked phonemes, tones, syntactic rules, grammatically expressed semantic and / or pragmatic differences, morphophonemic rules, more cases of addition, allomorph, agreement, etc. Scholars working in the area are interested, for

example, in the number of grammatical categories in the language (Shosted 2006), the number of phonemic oppositions (McWhorter 2008), the length of the "minimal description" of the language system (Dahl 2009). McWhorter (2001) compares word order, i.e. the position of the verb in the Germanic languages, proving that English syntax has a lower degree of complexity than Swedish and German. The reason for the claim is the loss of the V2 (verb-second) rule in English, according to which the personal verb in Swedish and German takes the second place in the sentence.

Language elements and functions with "duplicate" information or overspecification are viewed as "redundant" internal predictors of complexity, and therefore optional elements in a discourse (McWhorter 2008). P. Trudgill calls such elements "historical baggage" (Trudgill 1999: 149), V. M. Zhirmunsky – "hypercharacterization" (Zhirmunsky 1976), McWhorter – "ornamental elaboration", or "baroque accretion[s]" (McWhorter 2001). Syntagmatic redundancy is exemplified in indirect nomination and "semantic agreement". Language paradigmatic redundancy is manifested in synthetic grammatical categories, such as agreement and obviative markers (see McWhorter 2001).

The irregularity or "opacity" of form and word-formation processes as an internal factor in language complexity (see Mühlhäusler 1974) manifests itself in irregular affixes (prefixes pa- in 'pasynok' (stepson), su- in 'symrak' (twilight), niz- in 'nizvodit' (reduce), suffixes -tash in 'patrontash' (bandolier), -ichok in 'novichok' (novice), -arnik in 'kustarnik' (bush)) (see Kazak 2012).

External factors that determine language complexity are culture, language age and language contacts. Older languages serving well-developed multi-level cultures are considered to be more complex because they accumulated "mature language features" (cf. Dahl 2009, Deutscher 2010, Parkvall 2008). At the same time, intensive contacts between linguistic communities have a significant impact on the complexity of languages. At the beginning of this century, P. Trudgill stated that "small, isolated, low-contact communities with tight social networks" develop more complex languages than high-contact communities (Trudgill 2004: 306). However, in his later work, the researcher clarifies that the dynamics of interacting languages complexity is determined by the duration of contacts and the age of speakers mastering the superstratum: language simplification occurs during short-term contacts of communities, when adults learn a foreign (second) language. Language complication can take place in cases where the contacts are long-term and the second language is mastered not by adults, but children (Trudgill 2011). To prove the influence of language contacts on language complexity, B. Kortman and B. Smrechani (2004) compare the ways of implementing 76 morphosyntactic parameters, including the number of pronouns, noun phrases patterns, tense and aspect, modal verbs, verb morphology, adverbs, ways of expressing negations, agreement, word order, etc. in 46 variants of the English language. Researchers divide all variants of the English language into three large groups: (1) native to their speakers and performing all functions in the language community; (2) languages

that function as the second official language of the state, and (3) creole languages based on English. The study confirmed that the third group of languages, i.e. English-based creoles, are the least complex, native English (first) language varieties are the most complex, and second-language English varieties exhibit intermediate complexity (Kortmann & Szmrecsanyi 2004).

In the most general terms, *analytical methods* for assessing complexity are divided into *absolute* (theoretical-oriented and treated as "objective") and *relative* (user-oriented and thus "subjective[15]") (Crossley et al. 2008.). The absolute approach is popular in linguistic typology and is used to assess language complexity, while sociolinguistics and psycholinguistics use a relative approach. P. Trudgill defines relative difficulty as the difficulty which adults experience while learning a foreign language (Trudgill 2011: 371).

Text complexity as a construct is also modeled in discourse studies, linguistic personology, psycholinguistics and neurolinguistics. The area of these studies also includes relative complexity (or difficulty) of a text for different categories of recipients in different communicative environments, as well as absolute and relative (comparative) complexity of texts generated by different authors (see McNamara et al. 1996, Solnyshkina 2015).

4. Summary of articles in the issue

The current issue contains a detailed review and discussion of the best practices of text and discourse complexity assessment, as well as methods ranging from purely linguistic to complex interdisciplinary including multiple hard- and software tools.

One of the methods, i.e. eye-tracking, is viewed in the area as an objective way of assessing text complexity for different categories of readers. Research implementing eye-tracking techniques to evaluate Russian texts complexity remains spares. The basic task here is to select text parameters and oculomotor activity, as well as to identify methods of measuring text complexity perception. The features typically selected to measure text complexity are average word length and word frequency; as for parameters of oculomotor activity, it is preferably assessed with relative speed of reading a word, duration of fixations, and the number of fixations. Text readability is estimated in the number of words read per minute. Eye-tracking is the focus of articles contributed by Laposhina and co-authors and Bonch-Osmolovskaya and co-authors. Laposhina and co-authors show that the number of fixations on a word correlates with its length, while the duration of fixations correlates with its frequency. The research of Bonch-Osmolovskaya and co-authors is aimed at elementary discursive units (EDU) defined as "the quantum of oral discourse, a minimum element of discourse

[15] Attributing this type of complexity as subjective is not universally accepted, since it is quite objective for all participants in communication. It would be more appropriate to define this type of complexity as "individual".

dynamics" (cf. Podlesskaya, Kibrik 2009: 309). Eye-tracking techniques allow to indicate that the structure of EDE affects text readability.

Methods of neural networks are implemented to assess texts complexity in the articles by Cortalescu et al., Sharoff, Morozov et al., and Ivanov et al. They also share the object of research, i.e. texts for studying Russian as a foreign language. An accurate assessment of their complexity enables better text selection for various educational environments. E.g. implementation of BERT model mentioned above provides a high degree of accuracy of text complexity assessment, i.e. 91–92%.

While using neural networks, researchers face an important research problem, i.e. which text features affect neural network results. A possible approach here is to use neural network to measure correlation coefficients of numerous text features with text complexity. An extensive study of collections of texts of various genres in English and Russian, taking into account dozens of linguistic features, has made it possible to identify a number of non-obvious effects. For example, research shows that more prepositions are used in more complex texts in Russian but in simpler texts in English. Obviously, this is due to the difference in the typological structures of languages. Notably, however, genre has a much larger effect on text complexity across all languages as compared to differences between languages.

A broad review of multiple methods applied in the area is provided in the work of M.I. Solnyshkina and co-authors. The paper covers six historic paradigms of discourse complexology: formative, classical, closed tests, structural-cognitive period, the period of natural language processing, and the period of artificial intelligence.

An important distinguishing feature of the articles in this special issue and its contribution to discourse complexology is constituted by its diverse and extensive data: several hundred linguistic features, different languages, different text corpora, different genres. Text complexity is assessed on several levels: lexical, morphological, syntactic, and discourse. Multifaceted studies prove to explicate the nature of text complexity. The publications in the current issue also provide information on the corpora and dictionaries being compiled.

One of the most important parameters of text complexity is abstractness. The more abstract words a text contains, the more difficult it is. The latter makes it relevant to compile dictionaries of abstract/concrete words and means of estimating text abstractness. English dictionaries of abstract/concrete words were published at the turn of the century, and the Russian language was lately viewed as "under resourced" since no dictionary identifying the degree of words abstractness was available. Solovyov and co-authors present a detailed methodology of composing a dictionary of abstractness for the Russian language. The article also describes the areas of dictionary application.

Linguistic complexity is an interdisciplinary problem, an object of computational linguistics, philosophy, applied linguistics, psychology, neurolinguistics, etc. In the 21st century, complexity studies acquired concepts and terminology, developed and verified a wide range of linguistic parameters of

complexity. The main achievement of the new paradigm was the validation of cognitive predictors of complexity enabling the assessment of discourse complexity. This success, as well as an interdisciplinary approach to the problem, made it possible to integrate studies of discourse complexity into a separate area, i.e. discourse complexology. Complexity issues are not an "end in itself", since the research results are relevant both for linguistic analysis and for predicting comprehension in a wide range of pragmalinguistic situations.

One of these situations is cognitive analysis of mistakes made in a foreign language learning which is the object of research conducted by Lyashevskaya and Yanda and colleagues. Both studies focus on the interrelationship between text complexity of texts and cognitive resources necessary to comprehend a text. Lyashevskaya et al. established that the number of mistakes made by a student is correlated with morphological complexity of his/her discourse. Yanda et al. present a computer system designed to analyze and adequately explain mistakes of a learner of Russian as a foreign language.

5. Conclusion

The recent successes of computational linguistics have largely ensured accomplishments in discourse complexology and allowed scientists not only to automate a number of linguistic analysis operations, but also create user-friendly text profilers. Tools such as ReaderBench, Coh-Metrix, and RuMOR (cf. the current issue) are capable of solving both research and practical tasks: selecting texts for target audiences, editing and shortening texts, analyzing cognitive causes of errors, and even suggesting verbal strategies. The algorithms of automatic text profilers are based on classical and machine learning methods, including deep learning neural networks, one of the latest systems of which is BERT. At present, and this is well shown in a number of articles of the special issue, researchers are successfully combining methods of machine learning and the so-called "parametric approach".

However, the most important feature of modern research is a vast expansion of research problems and accuracy increase resulting from the abilities of artificial neural networks to learn and modify. Artificial intelligence breakthroughs are attributable to the three main factors: new advanced self-learning algorithms, high computer speeds, and a significant increase in training data. Modern databases, as well as dictionaries and tools for the Russian language developed in recent years, allowed the authors of the special issue to address and successfully solve a number of problems of text complexity.

A solid foundation for success in discourse complexity were findings of cognitive scientists at the beginning of our century which completely changed complexology paradigm. If the main achievement of the XXth century complexology was the idea that "different types of texts are complex in different ways", the discourse complexology of the XXIst century proposed and verified complexity predictors for various types of texts and developed toolkits for assessing relative complexity of texts in various communicative situations. With cognitive

methods in its arsenal, complexology acquired two additional variables: linguistic personality of the reader and reading environment.

The new research paradigm of linguistic complexology is manifested in those articles of the special issue which are aimed at defining new criteria for text complexity: expert evaluation, comprehension tests and reading speed tests have been replaced by new methods, which allow scholars to identify discourse units affecting text comprehension.

The studies published in the special issue also highlighted the main problems facing Russian linguistic complexology: creating a complexity matrix for texts of various types and genres, expanding the list of complexity predictors, validating new complexity criteria, and expanding databases for the Russian language.

RU

1. Введение

Статья посвящена современным трендам компьютерной лингвистики и проблематике сложности языка и дискурса. В ней также дается краткий обзор статей выпуска.

Компьютерная лингвистика (далее КЛ) является междисциплинарной наукой на стыке лингвистики и компьютерных наук. Она исследует проблемы автоматической обработки информации в языковой форме. Другое часто используемое название этой дисциплины, фактически синонимичное термину «компьютерная лингвистика», – обработка естественного языка (Natural Language Processing, NLP). Иногда эти понятия разграничивают, считая, что КЛ – в большей степени теоретическая дисциплина, а NLP – более прикладная. КЛ начала развиваться в начале 1950-х гг., почти сразу после появления компьютеров. Первой ее задачей была разработка машинного перевода, в частности перевода научных журналов с русского языка на английский. О начальном этапе развития КЛ можно прочитать в работе (Hutchins 1999). Безусловно, первоначальный оптимизм по поводу быстрого решения проблемы машинного перевода оказался необоснованным, и лишь в последние годы удалось получить переводы приемлемого качества. Однако в КЛ за 70 лет развития достигнуты серьезные успехи в решении многих актуальных практических задач, что сделало ее одним из самых динамично развивающихся и важных разделов как лингвистики, так и компьютерных наук. На наш взгляд, лучшими монографиями по КЛ являются (Clark et al. 2013, Indurkhya & Damerau 2010). Последний обзор, включающий также анализ перспектив ее развития, можно найти в статье (Church & Liberman 2021).

Появившееся примерно 10 лет назад глубокое обучение нейронных сетей (Schmidhuber 2015) обеспечило настоящую революцию в области искусственного интеллекта и в том числе во многих разделах КЛ. Искусственные нейронные сети представляют собой формальную модель биологических се-

тей нейронов. Важнейшей их особенностью является способность к обучению, в случае ошибки нейронная сеть определенным образом модифицируется. Хотя нейронные сети были предложены еще в 1943 г., лишь несколько лет назад был совершен прорыв в их использовании. Он связан с тремя факторами: появлением новых, более совершенных алгоритмов самообучения, повышением быстродействия компьютеров, увеличением накопленного в интернете объема данных для обучения. В области NLP к прорыву привело появление в конце 2018 г. модели BERT (Devlin et al. 2018) – нейронной сети, предобученной на корпусе текстов. В настоящее время BERT и ее усовершенствованные варианты показывают лучшие результаты в решении многих задач NLP (новейший обзор см. (Lauriola et al. 2022)).

В обзоре трендов компьютерной лингвистики делается акцент на следующих аспектах исследований: прикладные задачи, методы, компьютерные лингвистические ресурсы, вклад теоретической лингвистики в компьютерную, применение нейронных сетей глубокого обучения.

2. Прикладные задачи и методы компьютерной лингвистики

2.1. Прикладные задачи компьютерной лингвистики

Кроме машинного перевода можно выделить следующие основные классы прикладных задач, лежащих в русле КЛ: обработка документов, компьютерный анализ социальных сетей, анализ и синтез речи (в том числе голосовые помощники), вопросно-ответные системы, рекомендательные системы. Наиболее объемной является задача обработки документов, включающая в себя большой спектр подзадач: поиск, суммаризация, классификация, анализ тональности, извлечение информации и т.д.

Поиск, очевидно, следует рассматривать как наиболее известную задачу КЛ, успешно реализованную в поисковиках Google, «Яндекс» и повсеместно используемую. Обстоятельное введение в проблематику информационного поиска можно найти в (Маннинг и др. 2011). Основной вид поисковых запросов – набор ключевых слов. Двумя главными проблемами поиска являются: необходимость обеспечить быстрый поиск в гигантском количестве текстов в интернете и обеспечить поиск с учетом семантики запроса, а не просто совпадения слов в запросе и документе. Быстрый поиск предполагает предобработку всех документов в интернете и создание так называемого поискового индекса, указывающего, в каких конкретно документах находится искомое слово. Поиск документов по семантике, или семантический поиск, реализован в рамках хорошо известной концепции Семантической паутины, или Semantic Web (Domingue et al. 2011), в основе которой лежит идея онтологий, о которых речь пойдет ниже. Пример семантического поиска: Google в ответ на запрос *Бетховен та та та там* первой выдает ссылку на статью в «Википедии» о 5-й симфонии Бетховена, хотя в тексте статьи не содержится фраза *та та та там*. Таким образом, поисковик Google «понимает», что *та та та там* и

5-я симфония семантически связаны. Успешный поиск был бы просто невозможен без лингвистических исследований, которые привели к созданию алгоритмов морфологического и синтаксического анализа, тезаурусов и онтологий для экспликации семантических связей между сущностями.

Термин «извлечение информации» трактуется как поиск в тексте информации определенного вида: сущностей, их отношений, фактов и т.д. Наиболее проработанной является задача извлечения именованных сущностей (Name Entity Recognition, NER), т.е. имена персон, организаций, географических объектов и т.д. Недавний опрос IT-профессионалов из различных сфер бизнеса (ENA, June 6, 2022)[16] показал, что задача NER является наиболее востребованной в бизнес-приложениях. Для решения этой задачи применяются различные техники: использования готовых словарей имен людей, названий географических объектов; лингвистических признаков (использование заглавных букв), подготовленных паттернов именных групп; методов машинного обучения. Обзор этой области можно найти в (Sharnagat 2014). Системы NER, основанные на словарях и правилах, правильно извлекают около 90% сущностей в текстах. BERT-основанные системы обеспечивают уже около 94% правильно извлекаемых сущностей (Wang 2020), что сопоставимо с уровнем точности человека и демонстрирует преимущества нейронных сетей с глубоким обучением. Значительно сложнее задача извлечения событий и фактов. Классический подход здесь состоит в создании шаблонов событий, в которых фиксируются типы и роли сущностей, участвующих в событиях. Например, событие «24 июня 2021 г. Майкрософт презентовала Windows 11» описывается следующим шаблоном: Тип активности – коммерческая презентация, Компания – Майкрософт, Продукт – Windows 11, Дата – 24 июня 2021 г. Шаблоны такого вида создаются вручную, что является весьма трудоемким делом. От их качества зависит эффективность системы извлечения информации. Обычно такие системы извлекают лишь около 60% фактов (Jiang et al. 2016).

В последние годы много работ посвящено сентимент-анализу текстов (Cambria 2017). Под этим понимается определение тональности текстов: выражено ли в тексте позитивное или негативное отношение к описываемым объектам. Эта область важна компаниям для оценки комментариев пользователей об их товарах и услугах. Для решения этой задачи также используются паттерны, словари, методы машинного обучения. Для русского языка создан словарь RuSentiLex (Loukachevitch & Levchik 2016), включающий более 12 тыс. слов и словосочетаний, маркированных как позитивные, негативные или нейтральные. Главная проблема сентимент-анализа текстов – это зависимость тональности слова от контекста. Слово в одних контекстах может иметь позитивную окраску, а в других – негативную. Возможным решением данной проблемы можно рассматривать построение словарей сентимент-лексикона для специфических предметных областей. Еще одна фундаментальная проблема – не просто оценить тональность всего текста в целом, а установить, к

[16] https://gradientflow.com/2021nlpsurvey/

какому аспекту ситуации относится оценочное высказывание. Это особенно важно в прикладных исследованиях отзывов пользователей о товарах и услугах (Solovyev & Ivanov 2014). Лучший в настоящее время результат – около 85% по стандартным метрикам точности и полноты – достигнут с применением технологии BERT (Hoang et al. 2019).

Еще одной важнейшей задачей обработки документов является суммаризация или саммаризация текстов (Miranda-Jiménez et al. 2013) – автоматическое построение краткого изложения (абстракта) содержания текста (или текстов). Ее практическая важность определяется гигантским и все возрастающим объемом текстов в интернете. Существует два подхода к решению этой задачи: экстрактивный и абстрактивный. Первый подход – экстрактивный – состоит в оценке информационной значимости предложений в тексте и выделении небольшого числа наиболее значимых. Он требует нетривиальных математических методов оценки информационной значимости фрагментов текста. Второй – абстрактивный – состоит в генерации оригинальных предложений, суммирующих все содержание исходного текста. Для генерации абстрактов, т.е. аннотаций текстов, в последние годы успешно применяются нейронные сети. В качестве одного из наиболее важных компонентов системы суммаризации включают алгоритмы синтаксического анализа предложений. Краткий обзор представлен в (Allahyari 2017).

Следующей задачей, которую мы здесь рассмотрим, является компьютерный анализ социальных сетей (social network, social media). Анализ контента социальных сетей преследует много различных целей. Это и мониторинг настроений в обществе, и выявление проявлений экстремизма и иной противозаконной деятельности, и даже анализ распространения эпидемий. Анализ контента социальных сетей, связанного с пандемией ковида, в том числе с распространением дезинформации, появился уже в начале эпидемии (Cinelli et al. 2020). В ходе анализа социальных сетей определяются как собственно содержание сообщений, так и связи между пользователями, что позволяет выявлять группы пользователей с общими интересами. При этом существенную трудность представляет разнородность контента. В последние годы основным инструментом анализа социальных сетей стали нейронные сети (Ghani et al. 2019). В работе (Batrinca & Treleaven 2015) представлен обзор данной области исследований, специально ориентированный на гуманитариев.

Несколько особняком в КЛ стоят анализ и синтез речи, требующие специфических программно-аппаратных средств работы с акустическими сигналами. Системы распознавания речи очень разнообразны и классифицируются по многим параметрам: размеру словаря; типу (возрасту, полу) диктора; типу речи; назначению; типу структурной единицы и принципам ее выделения (фразы, слова, фонемы, дифоны, аллофоны и др.). Входной речевой поток сопоставляется с акустическими и языковыми моделями, включаю-

щими разнообразные признаки: спектрально-временные, кепстральные, амплитудно-частотные, признаки нелинейной динамики. Распознавание речи признается сложной задачей, поскольку слова произносятся разными людьми и в разных ситуациях по-разному. Тем не менее на настоящий момент существует множество коммерческих систем распознавания речи, в частности встроенных в Windows. Хорошо известна система Watson speech to text, разработанная IBM (Cruz Valdez 2021). На распознавании речи строится работа все более широко используемых голосовых помощников. В России широко известной среди них является разработка «Яндекса» – Алиса (ENA, June 6, 2022)[17]. Алиса интегрирована с сервисами «Яндекса»: по голосовой команде она ищет информацию в одноименном браузере, узнает погоду на Яндекс.Погоде, данные о трафике – в Яндекс.Картах и т.д. Алиса может управлять системами умного дома и даже развлекать: играть с детьми в загадки, рассказывать сказки и анекдоты. Распознавание речи в голосовых помощниках облегчается тем, что им достаточно настроиться на голос определенного человека. Обзор современного состояния проблематики голосовых помощников можно найти в (Nasirian et al. 2017), а по общим проблемам распознавания речи – в (Nassif 2019).

Синтез речи уже активно применяется в информационно-справочных системах, в объявлениях об отправлении поездов, в приглашениях к стойке в аэропортах, к определенному окну в госучреждениях и т.д. Во всех случаях это ситуации с ограниченным спектром синтезируемых фраз. Наиболее простым способом синтеза речи является ее компоновка из заранее записанных фрагментов. Качество синтеза оценивается по сходству синтезированной речи с речью человека. В целом к настоящему времени не удалось создать высококачественные системы синтеза речи. Новейший обзор по синтезу речи представлен в (Tan 2021).

Перейдем к рекомендательным системам, с которыми сталкивалось, вероятно, большинство пользователей интернета. Рекомендательные системы предсказывают, какие объекты (фильмы, музыка, книги, новости, веб-сайты) будут интересны конкретному пользователю. Для этого они собирают информацию о пользователях, иногда в явном виде, просят их дать оценку объектам интереса, а чаще – в неявном виде, собирая информацию о поведении пользователей в интернете. Продуктивной оказалась следующая идея: люди, одинаково оценивавшие какие-либо объекты в прошлом, вероятнее всего, будут давать похожие оценки другим объектам и в будущем (Xiaoyuan & Khoshgoftaar 2009). Именно эта идея позволяет эффективно экстраполировать поведение пользователей. При разработке рекомендательных систем возникают чисто лингвистические проблемы, например учет синонимии. Такие системы должны понимать, что «детский фильм» и «фильмы для детей» – это одно и то же. По проблеме синонимии в рекомендательных системах см. работу (Moon 2019), а общий обзор представлен в (Patel & Patel 2020).

[17] https://dialogs.yandex.ru/store

Вопросно-ответные системы, или QA-системы, призваны обеспечивать ответы на естественном языке на вопросы пользователей, т.е. обладать естественно-языковым интерфейсом. Речь идет о поиске ответов в текстовой базе данных, которой располагают QA-системы. QA-системы, как и поисковики, предоставляют пользователю возможность искать информацию. Однако важным отличительным свойством QA-систем является то, что они позволяют найти такую информацию, о которой пользователь мог и не подозревать, например, соответствующие его вкусам, но не известные ему фильмы, которые он бы не смог найти с помощью поисковика. Очевидно, что качество QA-системы зависит от того, насколько полна база данных, т.е. есть ли в ней вообще ответ на поставленный вопрос, а также от технологий обработки вопросов и сопоставления их с информацией в базе данных. Обработка вопроса начинается с определения типа вопроса и ожидаемого ответа. Например, вопрос «Кто …» предполагает, что в ответе должно быть имя человека. Далее применяются сложные методы КЛ. QA-системы, аналогично рекомендательным системам, также сталкиваются с проблемой синонимии (Sigdel 2020). Обзор проблематики QA-систем можно найти в (Ojokoh & Adebisi 2018).

2.2. Методы компьютерной лингвистики

Все методы КЛ можно разделить на два больших класса: основанные на словарях и правилах (шаблонах) и основанные на машинном обучении. Эти два класса принципиально различаются по подходам. В основе словарей и правил лежат знания о языке, аккумулированные лингвистами. Это высокопрофессиональный ручной труд и поэтому весьма дорогостоящий. Машинное обучение предполагает наличие большого числа примеров, обычно в виде размеченных корпусов (обучающего множества), проанализировав которые и выявив их закономерности, компьютер сможет находить решение и при анализе новых данных. Существуют различные способы машинного обучения, однако наибольшие успехи в последнее время демонстрируют нейронные сети глубокого обучения. Языковые данные подаются на вход нейронной сети в закодированном виде в формате токенов: букв, биграмм, коротких высокочастотных морфем и слов. Сложностью в применении этого подхода является необходимость разметки большого корпуса текстов под решаемую задачу: чем больше обучающее множество, тем лучше обучится нейронная сеть. При этом разметка носит достаточно простой характер и для ее выполнения не обязательно привлечение профессиональных лингвистов, можно ограничиться просто носителями языка.

Остановимся на базовых методах КЛ, отсылая за детальным изложением вопроса к вышеупомянутым монографиям (Clark et al. 2013, Indurkhya & Damerau 2010).

Автоматический анализ текста обычно начинается с его предобработки, включающей сегментацию текста, т.е. его разбиение на слова и предложения.

Может показаться, что это несложные задачи, поскольку слова отделяются друг от друга пробелами, а предложения начинаются с заглавной буквы и заканчиваются точкой (редко – восклицательным или вопросительным знаками, многоточием) с последующим пробелом. Это простейший пример правила или шаблона: «точка – пробел – заглавная буква». Однако точка может стоять в середине предложения после первого инициала, за ней будет пробел и затем второй инициал с заглавной буквой. Здесь точка явно не указывает на разделение текста на предложения. В качестве примера можно привести такое предложение: «Лукашевич Н.В., Левчик А.В. Создание лексикона оценочных слов русского языка РуСентилекс // Труды конференции OSTIS-2016. С. 377–382». Тем не менее, несмотря на указанные сложности, проблема сегментации считается практически решенной. Еще в 1989 г. в (Riley 1989) была достигнута точность 99,8% в решении задачи разбиения текста на предложения. Для достижения такого результата потребовалась сложная система правил. В ней учитывались такие признаки, как длина слова перед точкой, длина слова после точки, наличие слова перед точкой в словаре аббревиатур и ряд других.

Следующий шаг в ходе анализа текста – морфологический. Рассмотрим в качестве примера язык со сложной морфологией – русский. Для русского языка морфологический анализ выполняется многими анализаторами: MyStem, Natasha, pymorphy2, SpaCy и др. В КЛ морфологический анализ, цель которого состоит в определении морфологических характеристик слова, основан на детальном описании парадигм словоизменения. Для русского языка справочник создан такого рода создан (Зализняк 1977), в котором представлены индексы парадигм почти 100 тыс. слов (лемм) русского языка. Наличие такого справочника позволило сгенерировать около 3 миллионов словоформ для зафиксированных лемм русского языка. Автоматический анализ текста находит соответствующую любой словоформе лемму и полный перечень морфологических характеристик. Главной сложностью, с которой существующие анализаторы пока не справляются полностью, является омонимия форм. Базовое решение состоит в том, что анализатор выдает все варианты разборов. Однако во многих задачах требуется указать единственное решение. В этом случае анализаторы выдают наиболее частотный вариант морфологического разбора, не учитывая значение слова в контексте. Еще одна проблема – это проблема разбора «несловарных» слов, т.е. слов, отсутствующих в словаре. Для их морфологического анализа, учитывая, что количество таких слов в среднем составляет около 3%, приходится разрабатывать специальные алгоритмы. В простейшем случае анализируется окончание несловарной единицы и ей приписывается типичная для этого окончания парадигма словоизменения.

Синтаксический анализ, или парсинг, намного более сложен. Результатом синтаксического парсинга предложения является дерево зависимостей,

отражающее структуру предложения либо в формализме генеративной грамматики, либо в формализме грамматики зависимостей (dependency grammar (Tesnière 2015)). Для успешного синтаксического разбора необходимо детальное описание синтаксиса языка. Для русского языка наиболее успешным признан анализатор проекта ЭТАП, разрабатываемый более 40 лет в Лаборатории компьютерной лингвистики Института проблем передачи информации РАН. Его последняя версия – ЭТАП-4 доступна по адресу (ENA, June 6, 2022)[18]. В основу синтаксического анализатора проекта ЭТАП положена хорошо известная модель «Смысл ⇔ Текст» (Мельчук 1974), ее формализованный вариант изложен в монографии (Апресян 1989). В последнее десятилетие конкурирующим стал подход на основе нейронных сетей (Chen & Manning 2014). Для обучения нейронных сетей используются базы данных предложений с их синтаксическим разбором. Для английского языка это, например, English Penn Treebank (ENA, June 6, 2022)[19]. Для русского языка можно использовать SynTagRus (ENA, June 6, 2022)[20], созданный в Лаборатории компьютерной лингвистики ИППИ РАН.

Еще более сложной следует признать задачу семантического анализа. Однако, если мы хотим, чтобы компьютер хотя бы в какой-то степени «понимал» смысл, необходимо, некоторым образом, формализовать семантику слов и предложений. Классическими в решении данной проблемы являются два направления. Первое направление инициировано Ч. Филлмором (Fillmore 1968), который ввел понятия семантических падежей или ролей именных групп в предложении. Правильное установление семантических ролей – важный шаг к пониманию предложения. Исходные идеи Ч. Филлмора были воплощены в компьютерной лексической базе данных FrameNet (ENA, June 6, 2022)[21].

Второе направление – это создание электронного тезауруса (лексической онтологии) WordNet (Fellbaum 1998) для английского языка и его аналогов – для многих других языков. Для русского языка было предпринято несколько попыток создания аналога WordNet, наиболее удачным из которых и широко используемым в настоящее время признан тезаурус RuWordNet (ENA, June 6, 2022 [22](Loukachevitch & Lashevich 2016)), содержащий более 130 тыс. слов. В WordNet-подобных тезаурусах эксплицированы семантические отношения между словами (понятиями), в том числе синонимия, гипонимия, гиперонимия и ряд других. Данные системные параметры в определенной степени уже определяют часть семантики слов. WordNet успешно использовался в большом числе как лингвистических, так и компьютерных исследований.

[18] http://proling.iitp.ru/ru/etap4
[19] https://catalog.ldc.upenn.edu/LDC99T42
[20] https://universaldependencies.org/treebanks/ru_syntagrus/index.html
[21] https://framenet.icsi.berkeley.edu/fndrupal/
[22] https://ruwordnet.ru/ru,

В последние годы была предложена идея векторного представления семантики слов (word embeddings), в основу которой положена дистрибутивная гипотеза: лингвистические единицы, встречающиеся в аналогичных контекстах, имеют близкие значения (Sahlgren 2008). Данная гипотеза подтверждена в ряде работ, в рамках которых созданы и исследованы векторы частот слов, зафиксированных в большом корпусе текстов в контексте изучаемых слов. Существует целый ряд уточнений и компьютерных реализаций этой идеи, однако используется преимущественно word2vec (Mikolov et al. 2013), доступная в библиотеке Gensim (ENA, June 6, 2022) [23] и пользующаяся наибольшей популярностью. Для русского языка существует система RusVectores (Kutuzov & Kuzmenko 2017), доступная по адресу: (ENA, June 6, 2022)[24] и выполняющая ряд операций со словами на основе их векторной семантики. RusVectores, например, может рассчитывать семантическую близость слов.

Разумеется, важнейшим инструментом исследований в КЛ, да и всей лингвистики в целом, являются корпусы текстов. Первым корпусом был созданный в 1960-е гг. *Brown Corpus*, содержащий на момент создания один миллион слов. С тех пор требования по объему корпусов стали неизмеримо выше. Для русского языка наиболее известен Национальный корпус русского языка (НКРЯ, ENA, June 6, 2022[25]). Созданный в 2004 г., он постоянно пополняется и в настоящий момент включает более 600 млн слов. В 2009 г. Google создал очень интересный многоязычный ресурс – Google Books Ngram (ENA, June 6, 2022)[26], содержащий 500 млрд слов, в том числе 67 млрд слов для русского языка (подробнее о данном ресурсе см. Michel 2011). Важной проблемой остается разметка корпусов, которая в сложных случаях осуществляется вручную. При этом важным является привлечение нескольких аннотаторов и контроль согласованности их разметок (Pons & Aliaga 2021). Несмотря на то, что корпусы стали неотъемлемым элементом лингвистических исследований, споры о репрезентативности, сбалансированности, дифференциальной полноте, предметной и жанровой отнесенности, корректности данных продолжаются. Обсуждение этих вопросов для корпуса Google Books Ngram можно найти в (Solovyev et al. 2020).

Подводя итог этому разделу статьи, отметим, что благодаря КЛ мы имеем такие уже ставшие привычными сервисы, как информационный поиск, автоматическая коррекция ошибок и многие другие. Это стало возможным благодаря принципиально важным достижениям не только в компьютерных науках, но и в лингвистике. В КЛ используются обширные словари и тезаурусы, детально проработанные модели синтаксиса, гигантские корпусы текстов. Автоматический морфологический анализ в современном виде

[23] https://github.com/rare-technologies/gensim
[24] https://rusvectores.org/ru/
[25] https://ruscorpora.ru/new/
[26] https://books.google.com/ngrams

просто не существовал бы без «Грамматического словаря русского языка» А.А. Зализняка (1977). Многие исследования в КЛ основаны на созданных вручную тезаурусах WordNet и RuWordNet. Компьютерные технологии, в свою очередь, вносят вклад в развитие лингвистики. Использование корпусов текстов, статистических методов стало уже общим местом, без этого проведение серьезных лингвистических исследований становится невозможным. Все ключевые технологии КЛ являются общедоступными. Программы для решения основных задач для ряда языков, но не для русского, доступны здесь (ENA, June 6, 2022)[27].

В одной статье, разумеется, невозможно дать исчерпывающее представление о столь обширной и быстро развивающейся области науки о языке, как компьютерная лингвистика. Многие важные вопросы остались незатронутыми. К ним можно отнести следующие: разрешение кореференции, снятие омонимии, тематическое моделирование и др., для знакомства с которыми следует обратиться к специальной литературе или указанным выше монографиям.

3. Сложность языка и текста как научная проблема

Ядром спецвыпуска является группа статей, посвященных оценке сложности текстов.

Оценка сложности языка в зависимости от количества имеющихся в его системе категорий представляется, на первый взгляд, весьма логичной, а сама задача – выполнимой. Иллюстрацией в данном случае могут служить, например, фонологический инвентарь языка, количество морфофонологических правил или форм глагола. Очевидной в данном случае становится возможность сравнительной оценки сложности разных языков и присвоения им некоторой объективной, абсолютной сложности (Miestamo et al. 2008). Добавим, что именно «объективная» сложность значима при освоении неродного языка. С другой стороны, если язык изучается как родной, он не представляет для детей сложности, и с этой точки зрения сложность всех языков абсолютно одинакова. Исследователи признаются, что сложность языка и текста «сопротивляется измерению», а ученые, работающие в этой области, сталкиваются с концептуальными и методологическими трудностями.

Значимым в свете изучаемой проблематики представляется описание взаимосвязи и взаимозависимости двух направлений изучения сложности: сложности языка (lingue), или языковой (лингвистической) комплексологии, с одной стороны, и сложности текста (parole) или дискурса (discourse complexity), или дискурсивной комплексологии, – с другой.

Трактовка самого понятия «сложность языка (lingue)» кардинально менялась в течение XIX–XX вв. В XIX в. общепринятым было выдвинутое В. Гумбольдтом положение о том, что различия в структуре языка и, следовательно, сложности определяют развитие говорящих на этом языке людей (Humboldt

[27] https://stanfordnlp.github.io/CoreNLP/

1999: 37). Признавая данное положение, ученые фактически соглашались с концепцией неравного статуса языков и народов. В XX в. на смену гумбольдианским взглядам, утверждающим неравные позиции языков и их носителей, пришла концепция единой, неизменной для всех языков мира сложности, получившая два названия: ALEC («All Languages are Equally Complex», букв. «Все языки одинаково сложны») (Deutscher 2009: 243) и linguistic equi-complexity dogma – букв. лингвистическая догма равной сложности (Kusters 2003: 5). В работах ученых, поддерживающих данную концепцию, доказательству подлежали две гипотезы: (1) сложность языка складывается из под-сложностей (sub-complexities) его элементов; (2) все под-сложности в лингвистических подсистемах компенсированы: простота в области А компенсируется сложностью в области В, и наоборот («компенсаторная гипотеза»). Аргументируя концепцию «Все языки одинаково сложны», Ч. Хоккет весьма смело заявил: «Объективное измерение сложности затруднено, но субъективно понятно, что общая грамматическая сложность любого языка, включая его морфологию и синтаксис, примерно одинакова. Это неудивительно, поскольку все языки выполняют одни и те же функции: что не может быть сделано «морфологически», должно быть сделано «синтаксически» (Hockett 1958: 180–181). К сожалению, в работах данного направления и периода традиционно не обсуждались критерии оценки сложности, а эмпирические доказательства попросту отсутствуют. Подробный обзор точек зрения о «догме равной сложности» представлен в основополагающей работе Дж. Сэмпсона, Д. Гила и П. Традгилла «Сложность языка как эволюционирующая переменная» (Sampson et al. 2009).

Начало XXI в. ознаменовалось появлением ряда критических обзоров теории равной сложности всех языков, с одной стороны (см. Miestamo, Sinnemäki & Karlsson 2008), и провокационным заявлением Дж. Маквортера о том, что «креольские грамматики – самые простые грамматики в мире» (McWhorter 2001). Сама же идея о том, что все языки одинаково сложны, была доказательно отвергнута социолингвистами, которые продемонстрировали, что языковой контакт может привести к упрощению языка. Это показано на примере африкаанс, пиджинов и койне. Если признать возможность упрощения языка, то отсюда неизбежно следует, что до упрощения язык был сложнее, чем после. И если язык может быть более или менее сложным на разных этапах своей истории, то очевидно, что одни языки могут быть более сложными, чем другие (Trudgill 2012).

В начале 2000-х гг. идея о лингвистической сложности и «догме равной сложности» начала активно обсуждаться на конференциях и семинарах (см. семинар «Сложность языка как развивающаяся переменная», организованный Институтом эволюционной антропологии им. Макса Планка в 2007 г. в Лейпциге ENA, June 6, 2022[28]), в ряде журнальных статей (Shosted 2006,

[28] https://www.eva.mpg.de/fileadmin/content_files/linguistics/pdf/ComplexityWS_Webpage_2007.pdf

Trudgill 2004) и монографий (Даль 2009, Kusters 2003, Miestamo et al. 2008, Sampson et al. 2009).

В России публикации по сложности языка до сих пор малочисленны и преимущественно представлены обзорами, выполненными зарубежными учеными, однако в последнее время некоторый интерес к данной проблеме начал возрастать. Из наиболее значимых следует указать на статью А. Бердичевского (2012) и рецензию на книгу Питера Трандгилла «Sociolinguistic Typology», опубликованную в 2011 г. (Вахтин 2014). Проблемы сложности языка обсуждались в Институте лингвистических исследований Российской академии наук (ИЛИ РАН) в 2018 г. на конференции «Балканские языки и диалекты: корпусные и квантитативные исследования».

Локальная и глобальная сложность

Развитие лингвистической комплексологии привело к выделению двух типов сложности: глобальной, т.е. сложности языка (или диалекта) в целом, и локальной сложности, т.е. сложности отдельного уровня языка или домена (Miestamo 2008). И если оценка глобальной сложности языка, по мнению ученых, является весьма амбициозной и, вероятно, безнадежной задачей, сравниваемой Г. Дойчером с «погоней за дикими гусями» (Deutscher 2009), то измерение локальной сложности рассматривается учеными как вполне выполнимая задача, состоящая в составлении перечня и оценке предикторов сложности, объективируемых на различных уровнях языка. Список предикторов *фонологической сложности* традиционно включает объем инвентаря фонем, частоту встречаемости маркированных[29] фонем, тональные различия, супрасегментные модели, фонотактические ограничения и максимальные кластеры согласных (Nichols 2009, Shosted 2006). При оценке *морфологической сложности* классическими «факторами неудобств» (термин Браунмюллера 1990: 627) признаны объем флективной морфологии языка (или языковой разновидности), специфика алломорфии и морфофонемных процессов и др. (Dammel & Kürschner 2008, Kusters 2003). Расчет *синтаксической сложности* осуществляется на основе данных о количестве предписываемых синтаксисом языка правил по принципу «чем больше, тем сложнее», а также способности языка порождать рекурсии и клаузы внутри синтаксического целого (Ortega 2003, Givón 2009, Karlsson 2009). *Семантическая и лексическая сложность* трактуется на основе следующих параметров: количества неоднозначных единиц языка, различия инклюзивных и эксклюзивных местоимений, лексического многообразия и др. (Fenk-Oczlon & Fenk 2008, Nichols 2009). *Прагматическая*, или «скрытая», сложность, имеющая в своей основе закон экономии, есть сложность умозаключений, необходимых для восприятия текстов на данном языке. Языки со скрытой сложностью допускают минималистские, весьма простые поверхностные структуры, интерпретация грамматических категорий в которых требует нетривиальных умозаключений. В качестве примера исследователи приводят языки Юго-Восточной

[29] Маркированными считаются фонемы, редко встречающиеся в языках мира (Бердичевский 2012).

Азии, достигшие особенно высокой степени скрытой сложности, в частности за счет опущения местоимений, множественной кореференции в относительных предложениях, отсутствия маркеров отношений и «голых», без модификаторов, существительных с широким диапазоном интерпретаций (Bisang 2009).

Исследования показали, что высокие уровни локальной сложности одного уровня в языке необязательно влекут за собой низкую локальную сложность другого уровня, как это прогнозируется «догмой равной сложности». Например, анализ метрик морфологической и фонологической сложности в 34 языках, осуществленных Р. Шостедом, не выявил ожидаемой статистически значимой корреляции (Shosted 2006). А наблюдаемые Г. Фенк-Озлог и А. Фенком отдельные «балансирующие эффекты» (trade-offs) между локальными сложностями, к сожалению, также недостаточны, чтобы валидировать «догму равной сложности» языков. Г. Фенк-Озлог и А. Фенк, в частности, выявили, что в английском языке тенденция к фонологической сложности и односложности связана с тенденцией к омонимии и многозначности, к твердому порядку слов и идиоматичности речи (Fenk-Oczlon & Fenk 2008: 63). Д. Гил убедительно доказал, что изолирующие языки не обязательно компенсируют простую морфологию более сложным синтаксисом (Gil 2008).

Факторы (или предикторы) сложности языка принято делить на внутренние и внешние. *Внутренними факторами* сложности признаются количество элементов и категорий в языке, избыточность и нерегулярность языковых категорий. При оценке внутренней сложности в современных исследованиях весьма распространенным является так называемый «списочный подход», при котором ученые составляют список языковых явлений, присутствие которых в языке увеличивает степень его сложности, т.е. фактически списки предикторов внутренней сложности суть списки локальной сложности, описанной выше. Например, список предикторов сложности, составленный Дж. Николз, содержит более 18 параметров и включает фонологические, морфологические, синтаксические и лексические параметры (Nichols 2009). Язык считается более сложным, если в нем больше маркированных фонем, тонов, синтаксических правил, грамматически выраженных семантических и/или прагматических различий, морфофонемных правил, больше случаев дополнения, алломорфии, согласования и др. Ученых, работающих в рамках данного направления, интересует, например, количество грамматических категорий в языке (Shosted 2006), число фонематических оппозиций (McWhorter 2008), длина «минимального описания» системы языка (Даль 2009). Для иллюстрации упрощения языка при утрате предиктора Макуортер (2001) сравнивает порядок слов, т.е. позицию глагола в германских языках, доказывая, что синтаксис английского языка имеет более низкую степень сложности, чем шведский и немецкий. Причина положения состоит в утрате английским языком правила V2 (verb-second), в соответствии с которым личный глагол в шведском и немецком занимает второе место в предложении.

В качестве «избыточных» внутренних предикторов сложности признаются элементы и функции в системе языка, которые несут «дублирующую» информацию или «излишнюю спецификацию», букв. overspecification, и поэтому являются коммуникативно необязательными элементами (McWhorter 2008). П. Традгилл именует такого рода элементы «историческим багажом», букв. historical baggage (Trudgill 1999: 149), В.М. Жирмунский – «гиперхарактеризацией» (Жирмунский 1976), Макуортер – «декоративным украшением», букв. ornamental elaboration, или «барочными образованиями», букв. baroque accretion[s] (McWhorter 2001). В качестве иллюстрации синтагматической избыточности традиционно называют косвенную (непрямую) номинацию и «семантическое согласование». Иллюстрацией парадигматической избыточности в языке выступает синтетическое выражение грамматических категорий, например маркирование при согласовании (Избыточность в грамматическом строе языка) и маркирование обвиатива (см. McWhorter 2001).

Нерегулярность или «непрозрачность» формо- и словообразовательных процессов как внутренний фактор сложности языка (см. Mühlhäusler 1974) реализуется в нерегулярных аффиксах, встречающихся в отдельных словах (приставки па- (пасынок), су- (сумрак), низ- (низводить), суффиксы -таш (патронташ), -ичок (новичок), -арник (кустарник) (см. Казак 2012).

Внешними факторами, детерминирующими сложность языка, признаются культура, возраст языка и языковые контакты. Считается, что старые языки, обслуживающие хорошо развитые многоуровневые культуры, являются более сложными, поскольку аккумулировали «зрелые языковые черты», букв. mature language features (термин О. Даля (2009) (Deutscher 2010, Parkvall 2008). Вместе с тем существенное влияние на сложность языков оказывают интенсивные контакты между языковыми сообществами. В начале нашего столетия П. Традгилл заявил, что «небольшие, изолированные сообщества с низким уровнем контактов, имеющие тесные социальные сети», развивают более сложные языки, чем сообщества с высоким уровнем контактов (Trudgill 2004: 306). Однако в своей более поздней работе исследователь уточняет, что динамика развития сложности языков при их взаимодействии определяется длительностью контактов и возрастом носителей, осваивающих суперстрат: упрощение языка имеет место при кратковременных контактах сообществ, когда иностранный (второй) язык усваивают взрослые. Усложнение языка может иметь место в тех случаях, когда контакт долговременный, а второй язык осваивается не взрослыми, а детьми (Trudgill 2011). Для доказательства влияния языковых контактов на сложность языка Б. Кортман и Б. Смречаньи (2004) сравнивают способы реализации 76 морфосинтаксических параметров, включая количество местоимений, модели именных групп, время и вид, модальные глаголы, морфологию глагола, наречия, способы выражения отрицаний, согласование, порядок слов и др., в 46 вариантах английского языка. Ученые делят все варианты английского языка на три большие группы:

(1) родные для их носителей и выполняющие все функции в языковом сообществе; (2) языки, функционирующие как второй официальный язык государства, и (3) креольские языки, имеющие в основе английский. Исследование подтвердило, что третья группа языков, т.е. креольские языки, имеющие в основе английский язык, наименее сложны, разновидности английского как родного (первого) языка являются наиболее сложными, а разновидности английского языка, используемого носителями в качестве второго языка, демонстрируют промежуточную сложность (Kortmann & Szmrecsanyi 2004).

В самых общих чертах *аналитические методы* оценки сложности делятся на *абсолютные* (теоретико-ориентированные и трактуемые как «объективные») и *относительные* (ориентированные на пользователя и, таким образом, «субъективные[30]») (Crossley et al. 2008). Абсолютный подход популярен в лингвистической типологии и используется для оценки сложности языка, в то время как в социолингвистике и психолингвистике используется относительный подход. П. Традгилл определяет относительную сложность как трудность изучения иностранного языка взрослыми (Trudgill 2011: 371). Сложность текста как конструкт также моделируется в дискурсологии, лингвистической персонологии, в психолингвистике и нейролингвистике. При этом изучается относительная сложность (трудность) текста для разных категорий реципиентов в различных условиях коммуникации, а также абсолютная и относительная (сравнительная) сложность текстов, генерируемых различными авторами (см. McNamara et al. 1996, Солнышкина 2015).

4. Краткий обзор статей выпуска

Современный подход к оценке сложности текстов характеризуется использованием как комплекса лингвистических методов исследования, так и достаточно сложного аппаратного и программного инструментария. Основные идеи весьма полно представлены в настоящем выпуске. Важным способом объективной оценки сложности текста для читающего является методика отслеживания движения глаз, осуществляемого с помощью специального оборудования – систем айтрекинга. Для русского языка исследования в этом направлении только начинаются. В качестве базовой ученые вудвигают задачу выбора параметров текста и глазодвигательной активности, а также меры сложности восприятия текста. Обычно в качестве параметров текста выбираются средняя длина слов и средняя частотность, а в качестве параметров глазодвигательной активности: относительная скорость чтения слова, длительность фиксаций и количество фиксаций. Мерой читабельности текста является скорость чтения вслух в словах в минуту. Айтрекингу посвящены статьи А.Н. Лапошиной с соавторами и А.А. Бонч-Осмоловской с соавторами.

[30] Характеристика этого типа сложности как субъективной может быть принята условно, поскольку она является вполне объективной для всех участников коммуникации. Более подходящим являлось бы определение этого типа сложности как «индивидуальной».

В первой из вышеуказанных работ показано, что число фиксаций на слове коррелирует с его длиной, а длительность фиксаций – с частотностью. Вторая статья посвящена более сложным элементам текста – элементарным дискурсивным единицам (ЭДЕ), трактуемой как «квант устного дискурса, минимальный шаг, при помощи которого говорящий продвигает дискурс вперед» (Подлесская, Кибрик 2009: 309). Структура ЭДЕ также влияет на читабельность текста и это фиксируется с помощью айтрекинга.

Оценке сложности текстов с помощью наиболее современных методов глубокого обучения нейронных сетей посвящены работы Д. Корталеску с соавторами, С.А. Шарова, Д.А. Морозова с соавторами и В.В. Иванова с А.В. Абрамовым. Объект исследования – тексты, предназначенные для изучающих русский язык как иностранный. Точная оценка их сложности позволит правильно выбирать тексты в той или иной образовательной ситуации. Как отмечалось в первом разделе статьи, в качестве инструмента исследований используется, в первую очередь, модель BERT. Ее применение позволяет достичь высокой точности в определении сложности этого типа текстов – 91–92%.

Применение нейронных сетей предполагает успешное решение важной исследовательской лингвистической проблемы, а именно, определение признаков текстов, влияющих на решение нейронной сети. Один из возможных подходом состоит в том, чтобы вычислить коэффициенты корреляции ряда лингвистических признаков текста с оценками сложности текста нейронной сетью. Исследование на обширном материале коллекций текстов разных жанров на английском и русском языках с учетом десятков языковых признаков позволило обнаружить ряд неочевидных эффектов. Например, оказалось, что большее число предлогов ассоциируется с более сложными текстами в русском и с более простыми текстами в английском. Очевидно, это связано с различием в типологической структуре языков. Впрочем, на взаимосвязь языковых признаков текста с его сложность даже в большой мере влияет жанр текста.

Широкий обзор применения иных средств компьютерной лингвистики в проблематике сложности текстов дан в работе М.И. Солнышкиной с соавторами. В этой работе описана динамика развития и предложена периодизация в виде 6 парадигм дискурсивной комплексологии: формирующей, классической, периода закрытых текстов, структурно-когнитивного периода, периода обработки естественного языка, периода искусственного интеллекта.

Важной отличительной особенностью статей данного спецвыпуска и его вклада в дискурсивную комплексологию является учет огромного числа разнообразных данных: несколько сот языковых признаков, разные языки, разные корпусы текстов, разные жанры. Сложность текста рассматривается на нескольких уровнях: лексическом, морфологическом, синтаксическом, дискурсивном. Столь многоплановые исследования позволяют глубже понять природу самого понятия сложность текста. В статьях выпуска используются

не только уже существующие готовые корпусы текстов и словари, но описывается создание новых.

Степень абстрактности также рассматривается в качестве важнейшего параметра сложности текста. Чем больше абстрактных слов текст содержит, тем он сложнее. Это означает необходимость создания словарей абстрактной/конкретной лексики и средств расчета степени абстрактности текста. Ранее словари абстрактных/конкретных слов были созданы для английского и некоторых других языков, но не для русского. В статье В.Д. Соловьева с соавторами подробно описывается методология создания такого словаря для русского языка. Показано, как этот словарь может быть использован и в других исследованиях, кроме проблематики сложности.

Лингвистическая сложность представляет собой междисциплинарную проблему, которая изучается не только компьютерной лингвистикой, но также в рамках нескольких научных направлений: философии, прикладной лингвистики, психологии, нейролингвистики. В XXI в. проблематика сложности обрела собственный терминологический аппарат, разработала и верифицировала широкий спектр лингвистических параметров сложности, а основным достижение новой парадигмы стала валидация когнитивных предикторов сложности, поднявшая проблематику текста на новый уровень – уровень дискурса. Этот успех, а также междисциплинарный подход к проблеме позволили интегрировать исследования сложности дискурса в отдельную область – дискурсивную комплексологию. Проблематики сложности – не «вещь в себе», поскольку результаты исследований релевантны как для лингвистического анализа текста, так и для прогнозирования успешности восприятия информации в широком спектре прагмалингвистических ситуаций.

Одной из таких ситуаций является когнитивный анализ ошибок, допускаемых при изучении иностранного языка. Этой проблематике посвящены работы О.Н. Ляшевской с соавторами и Л. Янды с соавторами. В них исследования выходят на уровень взаимосвязей между сложностью текстов и когнитивными ресурсами, необходимыми для их понимания. В первой работе получен следующий интересный результат: чем сложнее используемые обучающимся аффиксы, тем меньше он допускает ошибок в текстах. Во второй работе описана компьютерная система, предназначенная для анализа и адекватного объяснения ошибок изучающего русский язык как иностранный.

5. Заключение

Успехи компьютерной лингвистики последних лет во многом обеспечили достижения дискурсивной комплексологии и позволили ученым не только автоматизировать ряд операций лингвистического анализа, но и создать удобные для пользователей профайлеры текстов. Такие инструменты, как ReaderBench, Coh-Metrix и RuMOR (подробно описанные в статьях данного выпуска) способны решать как исследовательские, так и практиче-

ские задачи: осуществлять подбор текстов для целевой аудитории, редактировать и сокращать тексты, производить анализ когнитивных причин возникновения ошибок и даже предлагать стратегии вербального поведения. Алгоритмы, используемые разработчиками при создании инструментов автоматического анализа текстов, имеют в своей основе классические методы и методы машинного обучения, включая нейронные сети глубокого обучения и одну из новейших систем – систему BERT. В настоящее время, и это хорошо показано в ряде статей спецвыпуска, ученые успешно совмещают методы машинного обучения и «параметрического подхода».

Однако важнейшей особенностью современных исследований является значительное расширение научной проблематики и повышение точности расчетов за счет способности искусственных нейронных сети к обучению и модификации. Прорыв в области искусственного интеллекта был обусловлен тремя основными факторами: появлением новых, более совершенных алгоритмов самообучения, повышением скорости работы компьютеров, многократным увеличением объема данных для обучения. Современные базы данных, а также разработанные в последние годы словари и инструменты для русского языка позволили авторам спецвыпуска обратиться и успешно решить целый ряд проблем в области сложности текста.

Еще одним фундаментом успеха в области сложности текста послужили открытия ученых когнитологов, сделанные в начале нашего века и навсегда поменявшие научную парадигму комплексологии. Если основным достижением комплексологии текста XX в. являлся вывод о том, что «разные типы текстов сложны по-разному», то дискурсивная комплексология XXI в. не только сумела предложить и верифицировать предикторы сложности для различных типов текстов, но разработала инструментарий для оценки относительной сложности текста в различных коммуникативных ситуациях. С обращением к когнитивным наукам комплексология обрела две дополнительные переменные: языковую личность читателя и коммуникативную ситуацию процесса чтения.

Новая исследовательская парадигма лингвистической комплексологии также отражена в тех работах спецвыпуска, которые посвящены поиску новых критериев сложности текста: на смену экспертной оценке, тестам на понимание и скорости чтения пришли новые методы, позволяющие выявлять дискурсивные единицы, влияющие на сложность восприятия текста.

Исследования, публикуемые в специальном выпуске высветили и основные проблемы, стоящие перед отечественной лингвистической комплексологией: создание матрицы сложности текстов различных типов и жанров, расширение списка предикторов сложности, валидация новых критериев сложности, расширение баз данных для русского языка.

Благодарность

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ-2030).

Acknowledgments

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л., Лазурский А.В., Перцов Н.В., Санников В.З., Цинман Л.Л. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989. [Apresyan, Yurii D., Igor M. Boguslavskii, Leonid L. Iomdin, Aleksandr V. Lazurskii, Nikolai V. Pertsov, Vladimir Z. Sannikov, Leonid L. Tsinman. 1989. *Lingvisticheskoe obespechenie sistems ETAP-2* (*Linguistic support of the system STAGE-2*). Moscow: Nauka. (In Russ.)].

Бердичевский А. Языковая сложность // Вопросы языкознания. 2012. № 5. С. 101–124. [Berdichevskii, Aleksandr. 2012. Yazykovaya slozhnost' (Language complexity). *Voprosy yazykoznaniya* 5. 101–124.] (In Russ.)

Вахтин, Н. Рец. на кн.: Peter Trudgil. Sociolinguistic Typology: Social Determinants of Linguistic Complexity // *Антропологический форум.* 2014. № 2. С. 301–309. [Vakhtin, Nikolai. 2014. Review of Peter Trudgil. Sociolinguistic Typology: Social Determinants of Linguistic Complexity. *Antropologicheskii Forum* 2. 301–309. (In Russ.)].

Даль Э. Возникновение и сохранение языковой сложности. М.: ЛКИ, 2009. [Dahl, Osten. 1976. Vozniknovenie i sokhranenie yazykovoi slozhnosti (The emergence and persistence of language complexity). Moscow: LKI. (In Russ.)].

Жирмунский В.М. Общее и германское языкознание: Избранные труды. Л.: Наука, 1976. [Zhirmunskii, Viktor M. 1976. Obshchee i germanskoe yazykoznanie: Izbrannye trudy (General and Germanic Linguistics: Selected works). Leningrad: Nauka. (In Russ.)].

Зализняк А.А. Грамматический словарь русского языка. М.: Русский язык, 1977. [Zaliznyak, Andrei A. 1977. Grammaticheskii slovar' russkogo yazyka (Grammatical dictionary of the Russian language). Moscow. (In Russ.)].

Избыточность в грамматическом строе языка / под ред. М.Д. Воейковой. СПб.: Наука, 2010. [Voeikova, Mariya D. (ed.). 2010. *Izbytochnost' v grammaticheskom stroe yazyka (Redundancy in the Grammatical Structure of the Language).* Saint Petersburg: Nauka. (In Russ.)].

Казак М.Ю. Морфемика и словообразования современного русского языка. Теория. Белгород: ИД «Белгород», 2012. [Kazak, Mariya Yu. 2012. Morfemika i slovoobrazovaniya sovremennogo russkogo yazyka. Teoriya (Morphemics and word formation of the modern Russian language. Theory). Belgorod: ID «Belgorod». (In Russ.)].

Кибрик А.А., Подлесская В.И. (ред.). Рассказы о сновидениях. Корпусное исследование устного русского дискурса. М.: Языки славянских культур, 2009. [Kibrik, A. A. & V. I. Podlesskaya (eds.). 2009. *Night Dream Stories: A Corpus Study of Russian Spoken Discourse.* Moscow: Yazyki slavyanskikh kul'tur. (In Russ.)].

Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011. [Manning, Kristofer D., Prabkhakar Ragkhavan & Khinrich Shyuttse. 2011.

Vvedenie v informatsionnyi poisk (Introduction to Information Search). Moscow: Vil'yams. (In Russ.)].

Мельчук И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М., 1974. [Mel'chuk, Igor' A. 1974. Opyt teorii lingvisticheskikh modelei «Smysl ⇔ Tekst» (The experience of the theory of linguistic models «Meaning ⇔Text»). Moscow. (In Russ.)].

Подлесская В.И., Кибрик А.А. Дискурсивные маркеры в структуре устного рассказа: Опыт корпусного исследования // *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегод. Междунар. конф. «Диалог».* 2009. Вып. 8 (15). С. 390–396. [Podlesskaya, V.I. & Kibrik A.A. 2009. Diskursivnye markery v strukture ustnogo rasskaza: Opyt korpusnogo issledovaniya (Discursive mrkers in the structure of oral narrative: The Experience of Corpus Research). In *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Proceedings of the Annual international conference Dialogue* 8(15). 390–396].

Солнышкина М.И., Кисельников А.С. Сложность текста: Этапы изучения в отечественном прикладном языкознании // *Вестник Томского государственного университета. Филология.* 2015. № 6. С. 86–99. [Solnyshkina, M.I., Kise'nikov, A.S. 2015. Slozhnost' teksta: Ehtapy izucheniya v otechestvennom prikladnom yazykoznanii (Text complexity: Stages of study in domestic applied linguistics). *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya* 6. 86–99].

Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez & Krys Kochut. 2017. Text summarization techniques: A brief survey. *arXiv* 1707.02268, URL: https://arxiv.org/pdf/1707.02268.pdf. (accessed 20.01.2022).

Batrinca, Bogdan & Philip Treleaven. 2015. Social media analytics: A survey of techniques, tools and platforms. *AI & Soc* 30 (1). 89–116. https://doi.org/10.1007/s00146-014-0549-4

Bisang, Walter. 2009. On the evolution of complexity: Sometimes less is more in East and mainland Southeast Asia. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 34–49. Oxford, New York: Oxford University Press.

Braunmüller, Kurt. 1990. Komplexe flexionssysteme – (k)ein problem für die natürlichkeitstheorie? *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 43. 625–635.

Cambria, Erik, Dipankar Das, Sivaji Bandyopadhyay & Antonio Feraco (eds.). 2017. *A Practical Guide to Sentiment Analysis*. Cham, Switzerland: Springer International Publishing.

Chen, Danqi & Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 740–750. https://doi.org/10.3115/v1/D14-1082

Church, Kenneth & Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence* 4. 625341. https://doi.org/10.3389/frai.2021.625341

Cinelli, Matteo, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo & Antonio Scala. 2020. The COVID-19 social media infodemic. *Sci Rep* 10. 16598. https://doi.org/10.1038/s41598-020-73510-5

Clark, Alexander, Chris Fox & Shalom Lappin (eds.). 2013. *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons.

Crossley, S.A., Greenfield, J. & McNamara, D. S. 2008. Assessing Text Readability Using Cognitively Based Indices. *TESOL Quarterly*, 42 (3), 475–493.

Dammel, Antje & Sebastian Kürschner. 2008. Complexity in nominal plural allomorphy. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 243–262. Amsterdam, Philadelphia: Benjamins.

Deutscher, Guy. 2009. «Overall complexity»: A wild goose chase? In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 243–251. Oxford: Oxford University Press.

Deutscher, Guy. 2010. *Through the Language Glass: Why the World Looks Different in Other Languages*. New York: Metropolitan Books.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv 1810.04805v2*. URL: https://arxiv.org/pdf/1810.04805.pdf. (accessed 20.01.2022).

Domingue, John, Dieter Fensel & James A. Hendler (eds.). 2011. *Handbook of Semantic Web Technologies*. Springer Science & Business Media.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fenk-Oczlon, Gertraud & August Fenk. 2008. Complexity trade-offs between the subsystems of language. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 43–65. Amsterdam, Philadelphia: Benjamins.

Fillmore, Charles J. 1968. The case for case. In Emmon W. Bach & Robert T. Harms (eds.), *Universals in Linguistic Theory*, 1–88. New York, NY: Holt, Rinehart & Winston.

Ghani, Norjihan A., Suraya Hamida, Ibrahim AbakerTargio Hashemb & Ejaz Ahmedc. 2019. Social media big data analytics: A survey. *Computers in Human Behavior* 101. 417–428. https://doi.org/10.1016/j.chb.2018.08.039

Gil, David. 2008. How complex are isolating languages? In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 109–131. Amsterdam, Philadelphia: Benjamins.

Givón, Thomas. 2009. *The Genesis of Syntactic Complexity: Diachrony, Ontogeny, Neuro-Cognition, Evolution*. Amsterdam, Philadelphia: Benjamins.

Hoang, Mickel, Oskar Alija Bihorac & Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. In Mareike Hartmann & Barbara Plank (eds.), *Proceedings of the 22nd Nordic conference on computational linguistics*, 187–196. Turku, Finland: Linköping University Electronic Press Publ.

Hockett, Charles F. 1958. *A Course in Modern Linguistics.* New York: Macmillan.

Humboldt, Wilhelm von. 1999. *On Language: On the Diversity of Human Language Construction and its Influence on the Mental Development of the Human Species*. Cambridge, U.K. New York: Cambridge University Press.

Hutchins, John. 1999. Retrospect and prospect in computer-based translation. In *Proceedings of MT Summit VII «MT in the Great Translation Era»*. 30–44. Tokyo: AAMT.

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. *Handbook of Natural Language Processing*. CRC Press.

Jiang, Ridong, Rafael E. Banchs & Haizhou Li. 2016. Evaluating and combining name entity recognition systems. In Nancy Chen, Rafael E. Banchs, Xiangyu Duan, Min Zhang & Haizhou Li (eds.), *Proceedings of NEWS 2016. The Sixth named entities workshop*, 21–27. Berlin, Germany.

Karlsson, Fred. 2009. Origin and maintenance of clausal embedding complexity. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 192–202. Oxford: Oxford University Press.

Kortmann, Bernd & Benedikt Szmrecsanyi. 2004. Global synopsis: Morphological and syntactic variation in English. In Bernd Kortmann, Edgar Schneider Werner, Clive Upton,

Kate Burridge & Rajend Mesthrie(eds.), *A Handbook of varieties of English*, 1142–1202. Berlin, New York: Mouton de Gruyter.

Kusters, Wouter. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection.* Utrecht: LOT.

Kutuzov, Andrey & Elizaveta Kuzmenko. 2017. WebVectors: A toolkit for building web interfaces for vector semantic models. In Wil M. P. van der Aalst, Dmitry I. Ignatov, Michael Khachay, Sergei O. Kuznetsov, Victor Lempitsky, Irina A. Lomazova, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Andrey V. Savchenko &Stanley Wasserman (eds.), *Analysis of Images, Social Networks and Texts,* 155–161. Moscow: AIST.

Lauriola, Ivano, Alberto Lavelli & Fabio Aiolli. 2022. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing* 470. 443–456. https://doi.org/10.1016/j.neucom.2021.05.103

Loukachevitch, Natalia V. & Anatolii Levchik. 2016. Creating a general Russian sentiment lexicon. In *Proceedings of Language Resources and Evaluation Conference LREC-2016*.

Loukachevitch, Natalia V. & G. Lashevich. 2016. Multiword expressions in Russian Thesauri RuThes and RuWordNet. In *Proceedings of the AINL FRUCT*. 66–71. Saint-Petersburg.

McNamara, Danielle S., Elieen Kintsch, Nancy Butler Songer & Walter Kintsch. 1996. Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text. *Cognition and Instruction*, 14 (1), 1–43

McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 6. 125–166. https://doi.org/10.1515/LITY.2001.001

McWhorter, John. 2008. Why does a language undress? Strange cases in Indonesia. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 167–190. Amsterdam, Philadelphia: Benjamins.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian veres, Matthew K. Gray, The Google books team, Joseph P. Pickett & Dale Hoiberg. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014). 176–182. https://doi.org/10.1126/science.1199644

Miestamo, Matti, Kaius Sinnemäki & Fred Karlsson (eds.). 2008. *Language Complexity: Typology, Contact, Change.* Amsterdam, Philadelphia: John Benjamins.

Miestamo, Matti. 2008. Grammatical complexity in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 23–42. Amsterdam, Philadelphia: Benjamins.

Mikolov, Thomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv 1301.3781*. URL: https://arxiv.org/abs/1301.3781 (accessed 20.01.2022).

Miranda-Jiménez, Sabino, Alexander Gelbukh & Grigori Sidorov. 2013. Summarizing conceptual graphs for automatic summarization task. In *Conceptual Structures for STEM Research and* Education. 245–253. *Lecture Notes in Computer Science* 7735.

Moon, Chang Bae, Jong Yeol Lee, Dong-Seong Kim & Byeong Man Kim. 2020. Multimedia content recommendation in social networks using mood tags and synonyms. *Multimedia Systems* 26 (6). 1–18. https://doi.org/10.1007/s00530-019-00632-w

Mühlhäusler, Peter. 1974. *Pidginization and Simplification of Language*. Canberra: Dept. of Linguistics, Research School of Pacific Studies, Australian National University.

Nasirian, Farzaneh, Mohsen Ahmadian & One-Ki D. Lee. 2017. *AI-based Voice Assistant Systems: Evaluating from the Interaction and Trust Perspectives*. Twenty-third Americas Conference on Information Systems, Boston.

Nassif, Ali Bou, Ismail Shahin, Imtinan Attili, Mohammad Azzeh & Khaled Shaalan. 2019. Speech recognition using deep neural networks: A systematic review. *IEEE access* 7. 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

Nichols, Johanna. 2009. Linguistic complexity: A comprehensive definition and survey. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 64–79. Oxford: Oxford University Press.

Ojokoh, Bolanle & Emmanuel Adebisi. 2018. A review of question answering systems. *Journal of Web Engineering* 17 (8). 717–758. https://doi.org/10.13052/jwe1540-9589.1785

Ortega, Lourdes. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24. 492–518.

Parkvall, Mikael. 2008. The simplicity of creoles in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 265–285. Amsterdam, Philadelphia: Benjamins.

Patel, Krupa & Hiren B. Patel. 2020. A state-of-the-art survey on recommendation system and prospective extensions. *Computers and Electronics in Agriculture* 178. 105779. https://doi.org/10.1016/j.compag.2020.105779

Pons Bordería, Salvador & Pascual Aliaga E. 2021. Inter-annotator agreement in spoken language annotation: Applying uα-family coefficients to discourse segmentation. *Russian Journal of Linguistics 25*(2). 478–506. https://doi.org/10.22363/2687-0088-2021-25-2-478-506

Riley, Michael D. 1989. Some applications of tree-based modelling to speech and language indexing. In *Proceedings of the DARPA Speech and Natural Language Workshop*. 339–352. San Mateo, CA.

Sahlgren, Magnus. 2008. The Distributional Hypothesis. From context to meaning. In distributional models of the lexicon in linguistics and cognitive science (special issue of the Italian Journal of Linguistics). *Rivista di Linguistica* 20 (1). 33–53.

Sampson, Geoffrey, David Gil & Peter Trudgill. 2009. *Language Complexity as an Evolving Variable.* Oxford linguistics. Oxford, New York: Oxford University Press.

Schmidhuber, Jürgen. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61. 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Sharnagat, Rahul. 2014. *Named Entity Recognition: A Literature Survey*. Center for Indian Language Technology.

Shosted, Ryan K. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10 (1). 1–40.

Sigdel, Bijay, Gongqi Lin, Yuan Miao & Khandakar Ahmed. 2020. Testing QA systems' ability in processing synonym commonsense knowledge. *IEEE* [Special issue]. *24th International Conference Information Visualisation (IV).* 317–321. https://doi.org/10.1109/IV51561.2020.00059

Solovyev, Valery & Vladimir Ivanov. 2014. Dictionary-based problem phrase extraction from user reviews. In Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (eds.), *Text, speech and dialogue*, 225–232. Springer.

Solovyev, Valery D., Vladimir V. Bochkarev & Svetlana S. Akhtyamova. 2020. Google Books Ngram: Problems of representativeness and data reliability. *Communications in Computer and Information Science* 1223. 147–162. https://doi.org/10.1007/978-3-030-51913-1_10

Su, Xiaoyuan & Taghi M. Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*. 1–19. https://doi.org/10.1155/2009/421425

Tan, Xu, Tao Qin, Frank Soong & Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv 2106.15561*. URL: https://arxiv.org/pdf/2106.15561.pdf (accessed 20.01.2022).

Tesnière, Lucien. 2015. *Elements of Structural Syntax*. Amsterdam: John Benjamins Publishing Company.
Trudgill, Peter. 1999. Language contact and the function of linguistic gender. *Poznan Studies in Contemporary Linguistics* 35. 133–152.
Trudgill, Peter. 2004. Linguistic and Social Typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8(3). 305–320.
Trudgill, Peter. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity.* Oxford: Oxford University Press (reprinted 2012).
Trudgill, Peter. 2012. On the sociolinguistic typology of linguistic complexity loss. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts & Paul Trilsbeek (eds.), *Language documentation & conservation special publication No. 3 (August 2012): Potentials of language documentation: Methods, analyses, and utilization*, 90–95.
Valdez, Cruz & Monika Louize. 2021. *Voice Authentication Using Python's Machine Learning and IBM Watson Speech to Text*. Universitat Politècnica de Catalunya.
Wang, Yu, Yining Sun, Zuchang Ma, Lisheng Gao, Yang Xu & Ting Sun. 2020. Application of pre-training models in named entity recognition. In *2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. 23–26. Hangzhou, China.

Article history:
Received: 20 October 2021
Accepted: 06 February 2022

Bionotes:

Valery D. SOLOVYEV is Doctor Habil. of Physical and Mathematical Sciences, Professor, Chief Researcher of "Text Analytics" Research Lab, Institute of Philology and Intercultural Communication of Kazan Federal University, Kazan, Russia. He is a member of the Presidium of the Interregional Association for Cognitive Research, author of four monographs and more than 60 publications on the text complexity.

Contact information:
Kazan Federal University
18 Kremlevskaya str., Kazan, 420008, Russia
e-mail: maki.solovyev@mail.ru
ORCID: 0000-0003-4692-2564
Scopus ID: http://www.scopus.com/authid/detail.url?authorId=26665013000

Marina I. SOLNYSHKINA is Doctor Habil. of Philology, Professor of the Department of Theory and Practice of Teaching Foreign Languages, Head and Chief Researcher of "Text Analytics" Research Lab, Institute of Philology and Intercultural Communication of Kazan Federal University, Kazan, Russia. She is the author of over 60 publications on text complexity.

Contact information:
Kazan Federal University
18 Kremlevskaya str., Kazan, 420008, Russia
e-mail: mesoln@yandex.ru
ORCID: 0000-0003-1885-3039

Danielle S. MCNAMARA, Ph.D., is Professor of Psychology in the Psychology Department and Senior Scientist at Arizona State University. She is an international expert in the fields of cognitive science, comprehension, natural language processing, and intelligent systems.

Contact information:
Arizona State University Payne Hall, TEMPE Campus, Suite 108, Mailcode 1104, the USA
e-mail: Danielle.McNamara@asu.edu
ORCID: 0000-0001-5869-1420

Сведения об авторах:
Валерий Дмитриевич СОЛОВЬЕВ – доктор физико-математических наук, профессор, главный научный сотрудник НИЛ «Текстовая аналитика» Института филологии и межкультурной коммуникации Казанского федерального университета, Казань, Россия. Член президиума Межрегиональной ассоциации когнитивных исследований. Автор четырех монографий и более 60 публикаций по сложности текста.

Контактная информация:
Казанский федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: maki.solovyev@mail.ru
ORCID: 0000-0003-4692-2564

Марина Ивановна СОЛНЫШКИНА – доктор филологических наук, профессор, профессор кафедры теории и практики преподавания иностранных языков, заведующий и главный научный сотрудник НИЛ «Текстовая аналитика» Института филологии и межкультурной коммуникации Казанского федерального университета, Казань, Россия. Автор более 60 публикаций по сложности текста.

Контактная информация:
Казанский федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: mesoln@yandex.ru
ORCID: 0000-0003-1885-3039

Даниэль С. МАКНАМАРА – доктор наук, профессор кафедры психологии Университета штата Аризона, психолингвист, международный эксперт в области когнитивистики, понимания, обработки естественного языка и интеллектуальных систем.

Контактная информация:
Arizona State University Payne Hall, TEMPE Campus, Suite 108, Mailcode 1104, the USA
e-mail: Danielle.McNamara@asu.edu
ORCID: 0000-0001-5869-1420

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 317–341
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30171

Research article

Natural language processing and discourse complexity studies

Marina SOLNYSHKINA[1] ✉, Danielle MCNAMARA[2] and Radif ZAMALETDINOV[1]

[1]*Kazan Federal University, Kazan, Russia*
[2]*Arizona State University, Tempe, USA*
✉mesoln@yandex.ru

Abstract
The study presents an overview of discursive complexology, an integral paradigm of linguistics, cognitive studies and computer linguistics aimed at defining discourse complexity. The article comprises three main parts, which successively outline views on the category of linguistic complexity, history of discursive complexology and modern methods of text complexity assessment. Distinguishing the concepts of linguistic complexity, text and discourse complexity, we recognize an absolute nature of text complexity assessment and relative nature of discourse complexity, determined by linguistic and cognitive abilities of a recipient. Founded in the 19th century, text complexity theory is still focused on defining and validating complexity predictors and criteria for text perception difficulty. We briefly characterize the five previous stages of discursive complexology: formative, classical, period of closed tests, constructive-cognitive and period of natural language processing. We also present the theoretical foundations of Coh-Metrix, an automatic analyzer, based on a five-level cognitive model of perception. Computing not only lexical and syntactic parameters, but also text level parameters, situational models and rhetorical structures, Coh-Metrix provides a high level of accuracy of discourse complexity assessment. We also show the benefits of natural language processing models and a wide range of application areas of text profilers and digital platforms such as LEXILE and ReaderBench. We view parametrization and development of complexity matrix of texts of various genres as the nearest prospect for the development of discursive complexology which may enable a higher accuracy of inter- and intra-linguistic contrastive studies, as well as automating selection and modification of texts for various pragmatic purposes.

Keywords: *text complexity, discourse, cognitive model, automatic analyzer, natural language processing*

For citation:
Solnyshkina, Marina, Danielle McNamara & Radif Zamaletdinov. 2022. Natural language processing and discourse complexity studies. *Russian Journal of Linguistics* 26 (2). 317–341. https://doi.org/10.22363/2687-0088-30171

Научная статья

Обработка естественного языка и изучение сложности дискурса

М.И. СОЛНЫШКИНА[1] ✉, Д. МАКНАМАРА[2], Р.Р. ЗАМАЛЕТДИНОВ[1]

[1]Казанский (Приволжский) федеральный университет, Казань, Россия
[2]Университет штата Аризона, Темпе, США
✉mesoln@yandex.ru

Аннотация
В исследовании представлен обзор формирования и развития дискурсивной комплексологии – интегрального научного направления, объединившего лингвистов, когнитологов и программистов, занимающихся проблемами сложности дискурса. Статья включает три основных части, в которых последовательно изложены взгляды на категорию сложности, история дискурсивной комплексологии и современные методы оценки сложности текста. Разграничивая понятия сложности языка, текста и дискурса, мы признаем абсолютный характер оценки сложности текста и относительный, зависимый от языковой личности реципиента характер сложности дискурса. Проблематика теории сложности текста, основы которой были заложены в XIX в., сфокусирована на поиске и валидации предикторов сложности и критериев трудности восприятия текста. Мы кратко характеризуем пять предыдущих этапов развития дискурсивной комплексологии: формирующего, классического, периода закрытых тестов, конструктивно-когнитивного и периода обработки естественно языка, а также подробно описываем современное состояние науки в данной области. Мы представляем теоретическую базу автоматического анализатора Coh-Metrix – пятиуровневую когнитивную модель восприятия, позволившую обеспечить высокий уровень точности оценки сложности и включить в список предикторов сложности текста не только лексические и синтаксические параметры, но и параметры текстового уровня, ситуационной модели и риторических структур. На примере нескольких инструментов (LEXILE, ReaderBench и др.) мы показываем области применения данных инструментов, включающие образование, социальную сферу, бизнес и др. Ближайшая перспектива развития дискурсивной комплексологии состоит в параметризации и создании типологии сложности текстов различных жанров для обеспечения более высокой точности меж- и внутриязыкового сопоставления, а также для автоматизации подбора текстов в различных лингвопрагматических условиях.

Ключевые слова: *сложность текста, дискурсивная комплексология, когнитивная модель, автоматический анализатор, обработка естественного языка*

Для цитирования:
Солнышкина М.И., Макнамара Д., Замалетдинов Р.Р. Обработка естественного языка и изучение сложности дискурса. *Russian Journal of Linguistics.* 2022. Т. 26. № 2. С. 317–341. https://doi.org/10.22363/2687-0088-30171

1. Введение

Более семи десятилетий исследователи в области компьютерной лингвистики искали решения, которые позволили бы компьютерным системам осуществлять обработку естественного языка. Эта работа велась как с целью решения сугубо лингвистических задач, включая, например, разработку теории общения на естественном языке (Hendrix 1980), так и для создания компьютерных систем, осуществляющих общение с пользователями на естественном языке. И хотя задача свободного понимания машиной естественного языка по-прежнему не решена, значительный прогресс достигнут в создании автоматизированных систем обработки естественного языка (см. Coh-Metrix, Lextutor, ReaderBench, RuLinva, TextInspector, Текстометр и др.), широко востребованных в системах образования и здравоохранения, социальных службах, бизнесе и праве. Потребность в такого рода инструментах заставляет исследователей искать новые алгоритмы и инструменты для решения задач, стоящих перед обществом. А сложность исследовательской задачи – автоматизации лингвистического анализа – обусловлена сложностью самой языковой системы, процесса восприятия вербальной информации в целом и в текстовом формате в частности, а также многообразием текстовых типов и жанров. В известной работе «Науки об искусственном» (1996) Г. Саймон, определяя основную задачу науки в целом, фактически обозначил направление развития проблематики сложности текста: «сделать удивительное тривиальным, показать, что правильно рассматриваемая сложность является лишь маской простоты (англ. complexity, correctly viewed, is only a mask for simplicity), найти закономерность, спрятанную в кажущемся хаосе»[1] (Simon 1996: 1).

В представленной работе мы предлагаем свой взгляд на категорию сложности текста, а также кратко описываем историю и достижения лингвистической комплексологии, реализованные в автоматических анализаторах различного уровня и класса.

2. Проблема сложности: сложность языка, текста и дискурса

Сложность признана ключевой характеристикой бытия, и стремление понять сложные системы объединяет ученых различных направлений. Новым в современной исследовательской парадигме является не изучение отдельных сложных систем, но описание самого феномена сложности, а также способов и инструментов оценки сложности. Термин «сложность» определяется в науке при помощи широкого спектра понятий: «многообразность»,

[1] Здесь и далее перевод с английского выполнен авторами статьи.

«множественность», «иерархия», «неоднородность», «неаддитивность» или «эмерджентность», «холизм», «гештальт», «информация», «общие системы», «генетические алгоритмы» и др. Например, Н. Решер определяет сущность сложности через «(1) количество и многообразие дискретных компонентов, из которых состоит объект, а также (2) количество связей между составляющими компонентами» (Rescher 1998: 1). Г. Саймон пишет о «неаддитивности» или «эмерджентности» сложных систем, их несводимости к сложности элементов: «Под сложной системой мы понимаем систему, состоящую из большого числа частей и взаимодействующую между собой непростым образом. В таких системах целое больше, чем сумма частей, в том смысле, что по заданным свойствам частей и их взаимодействиям нельзя правильным образом выявить свойства всей системы» (Саймон 2004: 104–105). Ф. Андерсон указывает на иерархическую структуру сложных объектов (Anderson 1972), а Г. Саймон подчеркивает, что «…сложность часто проявляется в форме иерархии и что иерархические системы имеют некоторые общие свойства, не зависящие от их конкретного содержания. <…> иерархия – это одна из центральных структурных схем, которую использует архитектор сложности» (Simon 1996: 184).

Что касается лингвистической сложности, то в современной научной парадигме сформировано три взаимосвязанных понятия: сложность языка (англ. language complexity), сложность текста (англ. text complexity) и сложность дискурса (англ. discourse complexity), а термин «лингвистическая сложность» (англ. linguistic complexity) используется для обозначения каждого из вышеуказанных понятий (см. Kortmann & Szmrecsanyi 2012). В некоторых случаях и термин «сложность языка» используется как синоним термина «сложность текста» (Sun 2020). Рассмотрим каждое из этих понятий. Наибольшую известность получили относительный и абсолютный подходы к понятию сложности, в рамках которых относительная и абсолютная сложность трактуются как свойство функций языка (т.е. элементов, паттернов, конструкций, правил), их (под-)систем или способов использования этих функций. Однако исследователи расходятся во мнениях относительно того, можно ли говорить об абсолютной сложности языка, т.е. существует ли некая независимая мера, с помощью которой оценивается сложность любого языка, либо сложность языка следует оценивать только в сравнении с другим языком. Можно ли оценивать сложность/простоту языка в целом (холистическая сложность) или, поскольку различные параметры сложности объективируются в разных структурах, оценка возможна только для отдельных уровней (уровневая сложность)? Весьма болезненным является вопрос о «количественной», т.е. структурной сложности языка, когда элементам или категориям в одном языке присваивается более высокая степень сложности, чем их аналогам в других языках. Поскольку при оценке абсолютной количественной сложности качественные аспекты сложности, такие как внутренняя сложность отдельных категорий, функций, правил, не учитываются, развитие

мысли о более высокой относительной сложности одного языка по сравнению с другим без учета степени «когнитивной сложности» языковых фактов может привести к выводу о том, что структурная простота, например изолирующих языков, означает примитивность развития их носителей (Steger & Schneider 2012: 159).

При оценке *сложности текста* разрабатываются шкалы уровней сложности текстов, предназначенных для различных категорий реципиентов, а присвоению тексту определенного уровня сложности предшествует создание «инвентарного» списка элементов, составляющих сложность каждого уровня. В этом случае текст оценивается как более или менее сложный на основании присутствия/отсутствия в тексте элементов, объективирующих «сложности» соответствующей шкалы. Несмотря на относительный характер *сложности текста* и самой процедуры, речь в данном случае идет об объективации «списочного подхода», т.е. абсолютной сложности. А относительной сложностью принято называть индивидуальную трудность текста для отдельного читателя. К сожалению, в ряде случаев индивидуальную трудность ошибочно называют субъективной сложностью (Bulté & Housen 2012: 31–33), хотя для читателя она весьма объективна.

И хотя достижения когнитологов и специалистов в области искусственного интеллекта позволили реализовать сложные модели верификации предикторов сложности и ввести в список обязательных семантические и дискурсивные параметры, термины «сложность текста» и «сложность дискурса» в современных исследованиях до сих пор используются как синонимы, обозначающие перечень параметров текста, детерминирующих трудность его восприятия читателями. Очевидно, что ускользающая от исследователя сущность лингвистической сложности текста и дискурса детерминирована в первую очередь неоднозначностью самих объектов, а также многообразием подходов к данным феноменам. Обратимся к истории разработки проблематики лингвистической сложности.

3. История автоматизации оценки сложности текста и дискурса

Приступая к краткому изложению истории автоматизации оценки сложности, следует отметить, что почти за 30 лет до появления первых формул читабельности русский ученый Н.А. Рубакин опубликовал свой знаменитый труд «Заметки по литературе для народа» (Рубакин 1890), в котором убедительно доказал, что сложность текста зависит от «знакомости» слов и длины предложения. А в 1893 г. вышла в свет работа профессора Л.А. Шермана «Analytics of Literature» («Аналитика литературы») (Sherman 1893), в которой ученый сформулировал два и по сей день никем не опровергнутых предиктора сложности текста: длина предложения и степень абстрактности. Еще одним важным событием, предопределившим появление первой формулы читабельности, следует признать идею использования частот слов в качестве параметра сложности. Именно реализация этой идеи в частотном списке слов Е. Торндайка (1921) обеспечила основу первой формулы читабельности.

Принцип, лежащий в основе частотных списков, заключается в том, что именно частота слова влияет на его узнавание в тексте, а значит, и на легкость/трудность восприятия информации в тексте.

Что касается автоматизации оценки сложности текста, то начало данного процесса авторы обзоров различных периодов развития комплексологии (см. Collins-Thompson 2015) традиционно связывают с именами Б. Лайвли и С. Прессли, разработавшими в 1923 г. первую формулу сложности текста на английском языке (Lively & Pressey 1923). Важным вкладом в теорию текстовой комплексологии того времени стали разработки М. Фогеля и К. Вошборна (Vogel & Washburne 1928), активно использовавшиеся практически без изменений до начала XXI в. Отправной точкой их исследования стало установление корреляции предикторов сложности текста с выбранными ими критериями: тестом на понимание, продолжительностью чтения текста, экспертной оценкой. В список предикторов сложности вошли лексические и синтаксические параметры, доли различных частей речи в тексте, а также информация о структуре абзаца и книги. Формула сложности М. Фогеля и К. Вошборна для английского языка имела следующий вид: $X_1 = 17{,}43 + 0{,}085X_2 + 0{,}101X_3 + 0{,}604X_4 – 0{,}411X_5$, где X_1 – сложность текста; X_2 – количество разных слов в выборке из 1000 слов; X_3 – количество предлогов в тексте; X_4 – количество слов, отсутствующих в списке Торндайка (1921); X_5 – количество простых предложений в выборке из 75 предложений. Первый этап развития дискурсивной комплексологии – формирующий – продолжался вплоть до 1935 г. и ознаменовался четырьмя важными достижениями: адаптацией способов оценки сложности, разработанных на детской учебной литературе, для других категорий читателей; включением в список предикторов «плотности идей» в тексте (Ojemann 1934); расширением списка критериев сложности текста и введением в него скорости чтения (McClusky 1934), а также включением 44 переменных в расчеты сложности текста (Gray & Leary 1935). Показательно, что уже в первых работах по сложности исследователи обращаются к широкому спектру параметров текста. Более того, именно в этот период, в 1925 г., вышел в свет первый сборник стандартизированных тестов по чтению У. МакКолла и Л. Краббс (McCall & Crabbs 1925), включавший 376 текстов и созданных на их основе тестов с множественным выбором. В течение длительного времени их валидность, установленная в ходе экспериментальных исследований с 2000 школьников Нью-Йорка, не подвергалась сомнению, а ранжированные тексты и тесты использовались в качестве критериев степени сложности. Надежность этого сборника, а вместе с тем и валидность основанных на нем формул впервые были подвергнуты жесткой критике в статье К.С. Стивенс, вышедшей в 1980 г. (Stevens 1980). Автор заявила, что инструмент плохо стандартизирован и никогда не предназначался для тех исследовательских целей, в которых он использовался.

Второй период развития комплексологии – *классический* (Klare 1963). Характерной чертой этого периода стало стремление к простоте и эффективности, нашедшее выражение в попытке разработать универсальные формулы

читабельности текстов, основанные на двух-трех предикторах. Попытки сократить количество переменных в формулах были обусловлены еще и отсутствием автоматизированных процедур расчетов переменных, что делало процедуру весьма утомительной. Именно в этот период появились простые в использовании и ставшие впоследствии классическими формулы: формулы Р. Флеша (Flesch 1948), предназначенной для взрослых, и формулы Е. Дейла и Дж. Чолл (Dale & Chall 1948), предназначенной для школьников. Обе формулы имели в своем составе два предиктора: синтаксический (среднее количество слов в предложении) и лексический (среднее количество слогов на 100 слов в формуле Р. Флеша и количество слов, не входящих в список Е. Торндайка, для формулы Е. Дейла и Дж. Чолл). Эти два исследования вдохновили множество последователей, сохранивших оригинальную методологию: использование в качестве предикторов не более двух переменных (лексических или синтаксических), а в качестве критериев – стандартизированные тексты и тесты У. МакКолла и Л. Краббс (см. McCall & Crabbs 1925).

Точкой отсчета следующего периода – периода так называемых *закрытых тестов*[2], пришедшего на смену классическому, стал выход в 1953 г. статьи У. Тейлора «Закрытый тест: новый инструмент для измерения читабельности» (Taylor 1953). В 1965 г. была опубликована книга Е. Коулмэна, в которой автор предлагал использовать закрытые тесты в качестве нового критерия сложности, утверждая, что количество правильно заполненных читателем пробелов является более достоверным критерием восприятия текста, а следовательно, и его сложности, чем ответы на вопросы по тексту (Coleman 1965). Формулы, рассчитанные на основе этого критерия, показали более эффективные результаты. Например, если коэффициенты множественной корреляции (R), полученные по формулам Р. Флеша или Е. Дейла и Дж. Чолл, не превышали 0,7, то результаты достоверности исследований Е. Коулмэна (1965) находились в пределах от 0,86 до 0,9, а Дж. Бормута (Bormuth 1969) – составляли более 0,92.

Еще одной особенностью этого периода стало то, что формулы на данном этапе «революции закрытого теста» (англ. the revolution of the cloze) включали больше предикторов, чем классические формулы. Этому способствовало появление и активное использование компьютеров, которые позволили исследователям автоматизировать подсчет некоторых переменных и обучение статистических моделей. В 1961 г. Э. Смит опубликовал первую формулу, расчеты которой были полностью автоматизированы (Smith & Quackenbush 1960), – формулу Деверо (Devereux). В качестве параметров Э. Смит использовал количество символов (знаков) в слове, аргументируя свой выбор тем, что это проще и быстрее, чем считать слоги или выбирать слова из списка (Smith & Quackenbush 1960: 333). Уже в 1963 г. У. Дэнилсон и С. Брайан адаптировали формулу Э. Смита для расчета на компьютере UNIVAC 1105 (Danielson & Bryan 1963). Одним из выдающихся исследователей того

[2] Cloze test – тест на заполнение пропусков в связном тексте.

периода был Дж. Бормут (Bormuth 1969). Именно Дж. Бормут впервые обратился к ряду методологических вопросов текстовой комплексологии: он первым показал, что связь между предикторами и критериями сложности не всегда является линейной. Он первым использовал дерево синтаксического анализа в качестве предиктора сложности и доказал, что оно менее эффективно, чем количество слов в предложении. Ему принадлежит и первенство расчета сложности на трех уровнях: уровне текста, предложения и слова. Он подвергает критике исследования, в которых рассчитывается коэффициент корреляции не из тестового, а обучающего набора данных.

В конце XX в. интерес к проблематике сложности текста наблюдается в отечественной науке изучается сложность текстов различных стилей и жанров, выходят ряд публикаций, появляются формулы читабельности для русского языка. В качестве переменных используются исключительно лексические и синтаксические предикторы: доля многозначных слов, абстрактных слов, трехсложных слов, средняя длина предложений и некоторые др. (см. обзор Солнышкина, Кисельников 2015). К сожалению, возможность автоматизации большого количества параметров привела к тому, что исследователи в ряде случаев стали выбирать параметры текста, расчет которых не представлял особых сложностей. В конечном итоге это способствовало формированию более поверхностного взгляда на сложность текста. Именно поэтому, а также благодаря достижениям когнитологов в области изучения процессов восприятия информации новые формулы читабельности подвергались все более резкой критике (Kintsch & Vipond 1979). Основные замечания когнитологов были нацелены на неспособность классических формул учитывать связность, когерентность, композицию и другие характеристики текста, а также интерактивность самого процесса чтения. Кроме того, дискредитации формул способствовала и повсеместная практика их применения на текстах таких типов, критерии сложности которых не были изучены. Практика такого рода приводила к недостоверности расчетов. В это же время появляются работы, авторы которых заявляют, что «смыслы несут люди, но не тексты» (Spivey 1987: 171), а процесс чтения и оценка сложности текста требуют учета индивидуальных способностей читателя – памяти, мотивации и др. На смену периоду «закрытых тестов» пришел новый – *структурно-когнитивный* или *когнитивный*. Появившись как следствие достижений в области лингвистики и когнитивных наук, он стал прорывом в лингвистической комплексологии: критика формул читабельности сопровождалась появлением исследований с новыми предикторами, измеряющими сложные психолингвистические характеристики, такие как информационная нагрузка текста (англ. inferential load) (Kemper 1983), концептуальная плотность текста (англ. conceptual density) (Kintsch & Vipond 1979) или композиция и тип текста (Meyer 1982). Следует, однако, признать, что, несмотря на гораздо более сложные предикторы, сочетание структурно-когнитивистских характеристик с классическими переменными привело к минимальному улучшению достоверности

(R = 0,78 vs 0,76) (см. Kemper 1983). Привлечение психолингвистических и даже нейролингвистических данных в комплексологию текста позволило поднять исследования на новый уровень, а «погружение» текста в коммуникативную ситуацию повлекло за собой включение в осуществляемые исследования сложности двух новых переменных: ситуативного контекста и языковой личности читателя. Текстовая комплексология стала дискурсивной комплексологией.

В 1990–2000-е гг. продолжают разрабатываться формулы, имеющие в своей основе традиционный подход: например, формула Дж. Чолл и Е. Дейла 1995 г. по сути является всего лишь обновленным вариантом старой формулы (Chall & Dale 1995). В это же время появляется серьезный интерес к данной проблематике в России и формулы сложности текстов на русском языке: формула И.В. Оборневой для художественных текстов и формула SIS (Solovyev-Ivanov-Solnyshkina) для учебных текстов (см. Solovyev et al. 2018, Solnyshkina et al. 2020). Именно в эти годы были заложены успехи современного этапа развития дискурсивной комплексологии – этапа *искусственного интеллекта*. Фокус парадигмы того времени был нацелен не только на поиск новых методов оценки сложности объекта, но и на создание аналитических инструментов, способных оптимально быстро дать характеристику текста и осуществить анализ его сложности. Формулируя требования к инструментам оценки сложности в 1993 г., М. Рабин подчеркивает значимость их способности рассчитывать три параметра: (1) длину оцениваемых последовательностей внутри системы, определяющую время их оценки; (2) «глубину вычисления», т.е. число уровней параллельных шагов, на которые последовательность может быть разложена; (3) объем памяти, требуемый для расчетов (Рабин 1993: 382). Именно в эти годы была создана система SATOdication, позволявшая осуществить автоматическую оценку 120 лингвистических переменных с весьма высоким коэффициентом корреляции расчетов, достигавшим отметки R = 0,86 (Daoust et al. 1996). Вскоре после этого была предложена новая метрика лексической связности текста, основанная на латентном семантическом анализе (LSA). Латентный семантический анализ может проводиться на семантически размеченных больших корпусах данных. П. Фольц и др. (Foltz et al. 1998) доказали, что предложения, расположенные в одном семантическом пространстве корпуса, имеют высокую вероятность внутренних связей. Это свойство стало использоваться для оценки глобальной когерентности текста и рассчитываться как среднее косинусное сходство всех пар смежных предложений. Для дискурсивной комплексологии это открытие оказалось весьма значимым: было доказано, что это – новый предиктор, поскольку лексическая связность коррелирует с критерием сложности текста.

А уже в 2001 г. Л. Си и Дж. Каллан определяют сложности текста как проблему классификации, расширяя таким образом спектр используемых для оценки методов, и применяют методы машинного обучения. Ученым удалось

выявить корреляции языковых моделей униграмм с типичными для разного уровня сложности контентами. Классификаторы уровней сложности при этом формируются как линейные комбинации языковой модели и поверхностных лингвистических признаков (Si & Callan 2001).

Существенный вклад в развитие теории сложности текста внесли исследования Ф. Даоста и др. (Daoust et al. 1996), П. Фольца и др. (Foltz et al. 1998), Л. Си и Дж. Каллан (Si & Callan 2001), которые и фактически поменяли научную парадигму. Исследования в новой парадигме по-прежнему предполагают наличие размеченного корпуса и выбор предикторов, анализ и комбинации которых верифицируются в статистических моделях, однако использование методов обработки естественного языка позволяет автоматизировать расчеты большого количества предложенных в рамках структурно-когнитивного подхода параметров. В первое десятилетие нашего века появляются исследования, нацеленные на валидацию не только лексических и синтаксических предикторов, но в первую очередь семантических, дискурсивных и когнитивных. Например, на основе расчетов, сделанных при помощи профайлера Coh-Metrix, исследовательская группа в составе С. Кроссли, Д. Дафти, П. МакКарти и Д. МакНамара разработала первую формулу сложности текста, сочетающую лексические и синтаксические параметры с параметрами связности (Crossley et al. 2007). Приблизительно в эти же годы исследователи делают еще одно важное открытие: синтаксические модели являются предиктором сложности текста для неносителей языка, в то время как на восприятие носителей языка сложность синтаксических моделей существенного влияния не оказывает (Schwarm & Ostendorf 2005).

Современные исследования в области дискурсивной комплексологии показывают, что эффективными моделями сложности текста являются только модели, ориентированные на конкретную категорию реципиентов, а выявление корреляций предикторов и критериев сложности предполагает учет целевой аудитории для конкретного текста. Фактически открытие этого важного закона предопределило качественно новый этап в развитии теории сложности и появление *дискурсивной комплексологии*, в которой, в отличие от комплексологии текста, критерии сложности текста выстраиваются в зависимости от категорий реципиентов. Два основных вызова времени – недостаточно качественная аннотация корпусов, искажающая данные, и необходимость уточнения и корректировки критериев сложности на различных категориях читателей, включая читателей с нарушениями речи, а также носителей и неносителей языка – решаются в новейшей истории комплексологии по-разному. Сочетание машинного обучения и методов обработки естественного языка должны обеспечивать лучшую точность предикторов, однако для современных более сложных статистических моделей нужно большее (по сравнению с предыдущими периодами) количество параметров для их обучения, следовательно, значительно возрастает потребность в больших, хорошо размеченных обучающих корпусах. Именно поэтому ученые все чаще обращаются к уже

размеченным корпусам, т.е. к корпусам с уже присвоенными уровнями сложности. Это может быть, например, корпус линейки учебников, сложность которых возрастает, например, в диапазоне от 2-го до 11-го класса (Solovyev et al. 2019, Gatiyatullina et al. 2020); корпус текстов для изучающих язык как неродной, в котором каждому тексту присвоен уровень сложности по Общеевропейской шкале (Лапошина, Лебедева 2021). Современный период комплексологии – период междисциплинарных исследований, в которых для изучения сложности привлекаются лингвисты, когнитологи, программисты и специалисты в области статистики. В настоящее время совершенно очевидно, что модели оценки сложности текста должны строиться на типологии и параметризации текстов различных регистров и жанров, типологии когнитивных моделей восприятия текста, а также на типологии реципиентов.

4. Инструменты оценки сложности текста и дискурса

В начале нового века появилась потребность в обработке больших массивов данных: компании, организации и государственные органы все чаще стали испытывать необходимость в анализе информации, полученной из нескольких текстовых источников, таких как онлайн-обзоры, электронные письма, транскрипты встреч и конференций или, например, приложений для обмена сообщениями. В настоящее время столь сложные задачи вполне осуществимы с помощью методов текстовой аналитики, а известные примеры практического применения методов обработки естественного языка включают автоматизированный машинный перевод, системы ответов на вопросы, извлечения и интерпретации информации, а также анализа настроений. Как ответ на практические запросы общества с конца прошлого века начинают появляться инструменты (системы, анализаторы, профайлеры) автоматического анализа текстов.

Появление автоматических анализаторов текстов не решило всех стоящих перед учеными проблем, но в определенной степени обострило их. Известно, что разработка и функционирование инструментов автоматического анализа естественного языка во многом зависят от имеющихся в распоряжении конструкторов баз данных, способов доступа к ним, а также средств и способов представления данных. Именно базы данных являются основой, на которой может строиться работа любой автоматизированной системы обработки естественного языка. При этом очевидно, что информация, доступная даже в нескольких базах данных, как правило, довольно ограничена (см. об этом Соловьев и др. 2022). Более того, пользователи, которым предоставляется доступ к базе данных, во многих случаях ожидают не только извлечения нужной им информации, хранящейся в базе, но и способности системы осуществлять на этих данных расчеты производной информации. При разработке эффективных систем автоматизированной обработки языка возникает также проблема создания средств представления данных. По мере того как компьютерные системы используют более сложные базы знаний, им

требуются более эффективные средства представления данных. Извлечение информации из текстов предполагает последующую ее визуализацию в виде таблиц, карт памяти, диаграмм, которые позднее интегрируются в базы данных и используются для описательной, предписывающей или прогнозной аналитики. Однако для того, чтобы система могла вести диалог с пользователем, ожидается, что она не только умеет интерпретировать вводимые данные, но и соответствующим образом реагирует на запросы пользователя, генерируя ответы, уже адаптированные к предполагаемым потребностям пользователя. Иллюстрацией решения такого рода проблемы в области обработки естественного языка является создание разработчиками Coh-Metrix «сокращенного» и более удобного в пользовании приложения TERA (Text Ease and Readability Assessor, букв. Анализатор легкости и удобочитаемости текста), предназначенного преимущественно для педагогов (ENA, April 18, 2022)[3]. TERA рассчитывает и визуализирует интегральные индексы пяти кластеров параметров: повествовательность, синтаксическая простота, конкретность слова, референциальная связность и глубокая связность, востребованных методистами и тестологами при отборе текстов для соответствующей аудитории читателей (см. Solnyshkina, Harkova & Kiselnikov 2014). Coh-Metrix же анализирует тексты по более чем 200 параметрам и предназначена преимущественно для исследовательских целей (McNamara & Graesser 2012).

Среди наиболее известных и хорошо зарекомендовавших себя инструментов для анализа текстов выделим следующие: ATOS (Advantage/TASA Open Standard), Lexile Framework (Lexile), REAP, проект университета Карнеги-Меллон, TextIspector, проект Открытого университета, Coh-Metrix, проект университетов Мемфиса и Штата Аризоны, США и Readerbench (ENA, April 18, 2022)[4], многоязычный профайлер Политехнического университета Бухареста. Для текстов на русском языке созданы две успешно функционирующие системы: Текстметр[5] (ENA, April 18, 2022), онлайн-инструмент определения уровня сложности текста РКИ, и RuLingva (ENA, April 18, 2022)[6], функционал которой поддерживает учебные тексты для носителей и неносителей русского языка.

В 2000-е гг. в Институте школьного возрождения (School Renaissance Institute) и компании Touchstone Applied Science Associates[7] была создана платформа ATOS (Advantage-TASA Open Standard), осуществляющая оценку текста по шести параметрам: количество слов в предложении, символов в слове, слогов в слове, частотность слова, процент знакомых слов и год овладения словом (ATOS Readability Formula). При разработке платформы

[3] http://129.219.222.70:8084/Coh-Metrix.aspx

[4] http://www.readerbench.com/

[5] https://textometr.ru/

[6] https://rulingva.kpfu.ru/

[7] В марте 2007 г. компания была переименована в Questar Assessment, Inc. (https://www.questarai.com/).

был создан корпус объемом 474 млн слов, 28 тыс. книг, 650 стандартизированных текстов, предназначенных для читателей нескольких уровней образования. В проекте были использованы записи чтения более 30 тыс. участников исследования. Параметр «год овладения словом» оценивался на основе ранжированного по годам обучения списка слов (Graded Vocabulary List). Список создан на основе «Словаря живых слов» (Living Word Vocabulary) (Dale & O'Rourke 1981), «Руководства по частотности слов» (Educator's Word Frequency Guide) (Zeno et al. 1995), «Списка слов для обучения (Words Worth Teaching) (Biemiller 2009), а также слов, используемых в стандартных тестах. Для удобства пользователей сложность текста шкалируется по году обучения в диапазоне от 0 до 15+. На платформе ATOS также были размещены списки книг, ранжированных по уровням сложности (Renaissance, Development of the ATOS, Special Collections. Accelerated Reader).

В современной версии анализатора Lexile, запущенного также в 2000-е гг. на платформе Lexile Platform (Lexile Framework for Reading), оценка сложности текста осуществляется при помощи двух индексов: семантического и синтаксического. Первый имеет в своей основе меру «незнакомости» слова, рассчитываемую при помощи частотности слов в корпусе, а второй – среднюю длину предложения (Lennon & Burdick 2004). Платформа активно используется в англоязычных странах для подбора учебников, тестовых материалов, а также развивающих книг для различных категорий читателей.

При создании анализатора DRP (Degrees of Reading Power, степени читательской способности), разработанного для платформы Questar (ENA, April 18, 2022) [8], была использована формула сложности Дж. Бормута с четырьмя предикторами: длина слова и предложения, количество знаков препинания и «узнаваемость» слова (Nelson et al. 2012). Сложность текста DRP ранжируется на непрерывной шкале от 0 до 100: более высокие значения маркируют более сложные тексты.

Отдельную нишу среди анализаторов текстов занимают так называемые «поисковые роботы» (англ. web crawlers), предназначенные для поиска вебтекстов определенной тематики и уровня сложности. Из наиболее известных укажем на REAP (Heilman et al. 2008) и TExtEvaluator (Sheehan et al. 2014). Профайлер REAP (REAder-specific Practice, букв. читательские практики) (ENA, April 18,2022)[9] основан на оценке сложности отдельных слов и не оценивает параметры высоких уровней, такие как семантика текста и связность. Ключевым компонентом REAP является расширенная модель поиска, способная находить документы, удовлетворяющие набору разнообразных лингвопрагматических запросов, включая тему текста, уровень образования (год обучения), тип синтаксических структур и словарный запас, соответствующий определенному уровню образования. REAP использует тематический классификатор SVM (Support Vector Machine, букв. метод опорных векторов).

[8] https://www.questarai.com

[9] https://www.lti.cs.cmu.edu/projects/language-technologies-education/reap-reader-specific-lexical-practice-improved-reading

Поиск выполняется на коллекции автоматически собранных из интернета документов, каждый из которых аннотирован и прошел процедуру метаразметки (Heilman et al. 2008).

TextEvaluator, заменивший SourceRater (ENA, April 18, 2022)[10], автоматизированный профайлер сложности текста, разработанный компанией ETS, имеет аналогичный функционал и адресован методистам, издателям учебников и разработчикам тестов для отбора текстов, соответствующих рекомендациям по сложности текста.

Преимущественное большинство инструментов автоматического анализа создано для английского языка, однако в последнее время стало появляться все больше инструментов для других языков: французского (François & Naets 2011), румынского (Dascalu 2014), итальянского (READ-IT), португальского (Antunes 2019, Marujo et al. 2009), русского (Gatiyatullina et al. 2020), голландского (readability 0.3.1, ENA, April 18, 2022)[11] языков.

Разработчики всех без исключения автоматических анализаторов признают, что обработка текстов пьес, интервью, стихов, рецептов или даже списков с нестандартной пунктуацией дает весьма противоречивые результаты. Наиболее существенным ограничением для современных систем обработки естественного языка считается буквальность интерпретации смыслов, т.е. неспособность систем осуществлять обработку переносных смыслов. Однако область применения профайлеров весьма широка и включает социальную сферу (Vergara & Lintao 2020), систему образования, медицину (Antunes 2019), юриспруденцию (Hall et al. 2006) и ряд других. Во всех этих сферах автоматические анализаторы применяются прежде всего для обеспечения так называемых «дискурсивных ограничений» (language-discourse constraints) (ENA, April 18, 2022)[12], т.е. ограничений на использование определенных языковых единиц в выбранных типах текстов.

5. Coh-Metrix: реализация подходов к оценке сложности дискурса

Достижения современной парадигмы дискурсивной комплексологии наилучшим образом реализованы в автоматическом анализаторе Coh-Metrix, предназначенном для оценки языковой сложности и тестирования когнитивных моделей восприятия. Инструмент прошел проверку временем и валидирован в ряде исследований (Hall et al. 2006, Graesser et al. 2014, Solnyshkina et al. 2014, Солнышкина, Кисельников 2015). Coh-Metrix создавалась на корпусе TASA (Touchstone Applied Science Associates), репрезентативность которого обеспечена не только его объемом (37 000 текстов), но и тем, что он «содержит примерно такое же количество и качество текстов, которые среднестатистический студент колледжа прочел за всю свою жизнь» (Jones et al.).

[10] https://texteval-pilot.ets.org/TextEvaluator/Default2.aspx; https://www.ets.org/research/policy_research_reports/publications/report/2015/junz

[11] https://pypi.org/project/readability/

[12] https://benjamins.com/catalog/pbns.172

Coh-Metrix осуществляет глубокую обработку текста на лексическом и синтаксическом уровнях, определяет степень связности текста и дискурсивные параметры текста, эксплицитность ситуационной модели. В основу разработки Coh-Metrix положена пятиуровневая когнитивная модель восприятия текста (Graesser & McNamara 2011) (см. табл. 1). Характеристика данной модели выходит за пределы данной статьи (см. Солнышкина и др. 2022), однако следует указать, что для восприятия представленной в тексте информации реципиент должен иметь способность понимать отдельные слова, извлекать информацию из долговременной памяти, использовать синтаксические знания для синтеза пропозиций и макропропозиций, применять эксплицитные и имплицитные коннекторы для установления связей между предложениями и частями текста, а также использовать фоновые знания и опыт для формирования когерентной модели текста (т.е. ситуационной модели (по Кинчу)).

Таблица 1. **Уровни восприятия дискурса /**
Table 1. **Levels of discourse (Graesser & McNamara 2011)**

Levels of discourse	Уровни восприятия дискурса
(1) Surface code Word composition (graphemes, phonemes, syllables, morphemes, lemmas, tense, aspect) Words (lexical items) Part of speech categories (noun, verb, adjective, adverb, determiner, connective) Syntactic composition (noun-phrase, verb-phrase, prepositional phrases, clause) Linguistic style and dialect	**(1) Параметры поверхностного кода** Состав слова (графемы, фонемы, слоги, морфемы, леммы, время, вид) Слова (лексические единицы) Частеречные категории (существительное, глагол, прилагательное, наречие, модификатор, связка) Синтаксис (именная группа, глагольная конструкция, предложная группа, предложение) Языковой стиль и диалект
(2) Text base Explicit propositions Referents linked to referring expressions Connectives that explicitly link clauses Constituents in the discourse focus versus linguistic presuppositions	**(2) Уровень текста** Эксплицитно выраженные пропозиции Референты и их связи с наименованиями референтов Эксплицитно выраженные коннекторы частей предложения Фокус дискурса и языковые факты
(3) Situation model Agents, objects, and abstract entities Dimensions of temporality, spatiality, causality, intentionality Inferences that bridge and elaborate ideas Given versus new information Images and mental simulations of events Mental models of the situation	**(3) Ситуационная модель** Агенты, объекты и абстрактные сущности Измерения темпоральности, пространственности, причинности, интенциональности Умозаключения, обеспечивающие связь и уточнение основных идей Данная и новая информация Образы и ментальные симуляции событий Ментальные модели ситуации
(4) Genre and rhetorical structure Discourse category (narrative, persuasive, expository, descriptive) Rhetorical composition (plot structure, claim + evidence, problem + solution, etc.)	**(4) Жанр и дискурсивные параметры текста** Категории дискурса (типы текстов) (повествовательный, аргументативный, информативный, описательный)

Levels of discourse	Уровни восприятия дискурса
Epistemological status of propositions and clauses (claim, evidence, warrant, hypothesis) Speech act categories (assertion, question, command, promise, indirect request, greeting, expressive evaluation) Theme, moral, or point of discourse	Структура текста (структура сюжета, заявление + доказательство, проблема + решение и т.д.) Эпистемологический статус суждений и предложений (утверждение, доказательство, предписание, гипотеза) Категории речевого акта (утверждение, вопрос, команда, обещание, косвенная просьба, приветствие, экспрессивная оценка) Тема, мораль или идея
(5) Pragmatic communication Goals of speaker/ writer and listener/reader Attitudes (humor, sarcasm, eulogy, deprecation) Requests for clarification and backchannel feedback (spoken only)	**(5) Прагматический уровень коммуникации** Цели говорящего/писателя и слушателя/читателя Отношение (юмор, сарказм, восхваление, осуждение) Запросы разъяснений и обратной связи по обратному каналу (только в устной форме)

Рассмотрим, каким образом и при помощи каких параметров Coh-Metrix осуществляет анализ текста на английском языке.

Параметры поверхностного кода. Лексические и семантические метрики оцениваются в Coh-Metrix на основе баз данных и корпусов, включая одну из наиболее полных – MRC (ENA, April18, 2022)[13]. Coh-Metrix рассчитывает многозначность слов, возраст освоения, образность, абстрактность, долю знаменательных слов в тексте и др. (см. McNamara et al. 2014: 247–251). Частота слов рассчитывается на основе данных CELEX (ENA, April 18, 2022)[14], неоднозначность и «гиперонимический уровень»[15] – на основе WordNet (ENA, April 18, 2022)[16].

Синтаксическая сложность текста рассчитывается при помощи синтаксического анализатора Е. Чарняк (2000) (Charniak 2000: 132–139) на основе количества именных и глагольных групп, предложных словосочетаний, встроенных конструкций, модификаторов именных групп, слов перед сказуемым в главном предложении и др. Полный перечень индексов синтаксической сложности представлен в (McNamara et al. 2014: 247–251).

Уровень текста содержит эксплицитно выраженные пропозиции, референциальные связи и подтексты (англ. inferences), необходимые для понимания (van Dijk & Kintsch 1983). В качестве иллюстрации А. Грейсер

[13] https://websites.psychology.uwa.edu.au/school/mrcdatabase/mrc2.html

[14] https://www.ldc.upenn.edu/language-resources/data/obtaining

[15] «Гиперонимический уровень» есть способность слова вступать в гиперо-гипонимические отношения, коррелирующая со степенью абстрактности слова. Например, существительное *стол* имеет семь гиперонимических уровней: место – мебель – обстановка – инструментальность – артефакт – объект – сущность и обладает высокой степенью конкретности. Абстрактные слова, например, humanity, имеют более низкий «гиперонимический уровень»: humaneness (человечность) – quality (качество) – attribute (признак) – abstraction, abstract entity (абстракция, абстрактная сущность) – entity (сущность) (http://wordnetweb.princeton.edu/perl/webwn).

[16] https://wordnet.princeton.edu/

и Д. Макнамара (Graesser & McNamara 2011) предлагают следующее предложение и его пропозициональный анализ:

> When the board met on Friday, they discovered they were bankrupt. They needed to take some action, so they fired the president. *Букв. Когда совет собрался в пятницу, они поняли, что обанкротились. Им нужно было принимать меры, поэтому они уволили президента.*

Первое предложение содержит следующие пропозиции:

PROP 1: meet (board, TIME = Friday)	ПРОП-Я 1: встречаться (совет, ВРЕМЯ = пятница)
PROP 2: discover (board, PROP 3)	ПРОП-Я 2: понимать (совет, ПРОП-Я 3)
PROP 3: bankrupt (corporation)	ПРОП-Я 3: банкротство (корпорация)
PROP 4: when (PROP 1, PROP 2)	ПРОП-Я 4: когда (ПРОП-Я 1, ПРОП-Я 2)

Понимание второго предложения обеспечивается благодаря анафорической референциальной связи (referential cohesion) местоимения *they* и антецедента *board*. Очевидно, что ситуационная модель данного текста может содержать указания на «глубинные связи» (англ.deep cohesion), такие как увольнение президента советом, некомпетентность президента как причина банкротства, новый президент, платежеспособность корпорации. Coh-Metrix рассчитывает следующие типы *лексической кореферентности:* повторы (overlaps) нарицательных существительных, местоимений, основ и знаменательных частей речи как в смежных предложениях, так и во всем тексте. *Глубинная связность* на уровне предложений и текста рассчитывается на основе частоты встречаемости следующих типов дискурсивных маркеров: аддитивные (*также, кроме того*), темпоральные (*а затем, после, во время*), каузальные (*потому что, так*) и логические (*следовательно, если, и, или*). Частота вхождений каждого класса дискурсивных маркеров нормализуется на 1000 словоформ. *Лексическое разнообразие* рассчитывается как отношение неповторяющихся в тексте слов (англ. types) и словоформ (англ. tokens).

Важным условием понимания текста является способность моделирования *ситуационной (или ментальной) модели*, «референциального содержания или микромира того, о чем текст» (Graesser et al. 1994: 375). Ситуационная модель текста эксплицируется в причинно-следственных связях, интенциональности, темпоральности, пространственности и количестве субъектов коммуникативной ситуации (Zwaan & Radvansky 1998), рассчитываемых при помощи следующих метрик: количество каузальных глаголов, каузальных структур (*потому что, как следствие, как результат*), интенциональных глаголов, интенциональных структур (*чтобы, с помощью, посредством*), морфологических повторов (*времени и вида глагола*), отношение каузальных структур к каузальным глаголам, отношение интенциональных структур к интенциональным глаголам и др. Очевидно, например, что маркеры несовпадающих временных форм затрудняют построение ситуативной модели, а присутствие темпоральных маркеров (*позднее, до того, как, накануне*),

наоборот, снижают сложность текста. *Пространственность* как «пространственная плотность» (spatial density) и «пространственная связность» (spatial cohesion) оценивается путем подсчета доли существительных с семантикой местоположения и глаголов движения.

В дополнение к обсуждавшимся ранее предикторам кореференции, Coh-Metrix также оценивает *концептуальное сходство* между предложениями и абзацами при помощи латентного семантического анализа (LCA), статистического метода представления знаний о мире, основанного на идее семантического сходства слов, имеющих аналогичное окружение. На методе LCA основан еще один предиктор Coh-Metrix – *отношение объема заданной и новой* информации в тексте – рассчитываемый в двух вариантах: как среднее значение и стандартное отклонение (LSA given/new, sentences, mean, LSA given/new, sentences, standard deviation).

Разработчики Coh-Metrix нашли весьма изящное решение проблемы оценки сложности текстов различных жанров. Поскольку «разные типы текстов сложны по-разному», а формулы сложности жанрозависимы (см. Solnyshkina et al. 2020), определение жанровой принадлежности весьма затруднительно. В качестве причин укажем на отсутствие общепринятого набора текстовых категорий отдельных жанров и вероятностный характер присутствия в текстах одного жанра текстовых категорий нескольких жанров. Учитывая, что объекты «могут иметь категориальное отношение друг к другу только посредством обладания общими категориальными признаками» (см. Rosch & Mervis 1975: 603), можно было бы рассчитывать количество жанровых признаков, присутствующих в каждом конкретном тексте. При этом значимыми остаются два вопроса: нужно ли при оценке сложности оценивать присутствие категориальных признаков всех жанров в каждом конкретном тексте или достаточно выбрать, например, два–три жанра, имеющих стандартизированные критерии сложности и категориальные признаки которых можно аннотировать и рассчитать в большом корпусе. В качестве таких жанров разработчиками Coh-Metrix были выбраны нарративы (художественные тесты), учебные тексты по истории и естественнонаучные тексты, поскольку (1) обязательная программа школьного образования строится преимущественно на данных типах текстов и (2) именно для этих жанров были рассчитаны предикторы сложности (количество глаголов прошедшего времени, длина и частотность слов (McCarthy et al. 2009). Гипотеза, положенная в основу исследования, была следующая: категории текстов различных жанров можно ранжировать по сложности, а затем, определяя долю этой категории в тексте, определять их сложность. Например, если текст X содержит текстовую категорию C, и если категория C имеет уровень сложности D, то текст X унаследует уровень сложности D (McNamara et al. 2014: 13). Именно поэтому разработчики Coh-Metrix предлагают осуществлять классификацию текстов для оценки их сложности не по жанровой принадлежности, а по присутствию в них категорий того или иного жанра (McNamara et al.

2014: 5–6). Применение таких методик позволяет, например, выявить в тексте 70 % повествовательности и 30 % информативности.

Многочисленные исследования валидировали алгоритмы Coh-Metrix, доказав ее значимость как для исследовательских, так и практических целей: Coh-Metrix обеспечивает надежный анализ пяти уровней дискурса.

6. Заключение

Современная парадигма дискурсивной комплексологии как интегрального научного направления, предметом которого является оценка абсолютной и относительной сложности дискурса, сформирована на фундаменте лингвистических и психолингвистических достижений, а также успехов в области компьютерной лингвистики. Пройдя пять этапов своего развития – формирующего, классического, закрытых тестов, структурно-когнитивного и обработки естественного языка – дискурсивная комплексология разработала и валидировала более двухсот предикторов абсолютной сложности и десятки критериев относительной сложности. Дискурсивная комплексология подняла проблематику сложности текста на новый уровень, доказав, что оценка сложности текста должна быть дополнена оценкой когнитивных и лингвистических способностей языковой личности реципиента текста, а также анализом коммуникативной ситуации.

Специфика современного этапа – этапа искусственного интеллекта – состоит в использовании как традиционного «параметрического подхода», так и новых методов – методов машинного обучения для создания текстовых профайлеров, осуществляющих оценку сложности, подбор и модификацию текстов. Одним из наиболее успешных проектов в данной области следует признать разработку автоматического анализатора текстов на английском языке Coh-Metrix, в которой реализована пятиуровневая когнитивная модель восприятия текста. В дополнении к дескриптивным и «классическим» параметрам, таким как длина текста и индекс читабельности, Coh-Metrix успешно осуществляет оценку параметров текстового уровня, ситуационной модели и жанровых характеристик.

Перспектива научных исследований лингвистической комплексологии состоит в параметризации текстов – разработке перечня типологических параметров сложности текстов различных типов и жанров. Параметризация текстов разных языков позволит автоматизировать процесс подбора текстовых материалов для решения образовательных, социальных и профессиональных задач, а также добиться большей точности внутри- и межъязыковых исследований.

Благодарность

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ-2030).

Acknowledgments

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Anderson, Philip. 1972. More is different: Broken symmetry and the hierarchical nature of science. *Science* 177 (4047). 393–396.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/S0022226700014201

Biemiller, Andrew. 2009. *Words Worth Teaching*. Columbus, OH: SRA/McGraw-Hill.

Bormuth, John R. 1969. *Development of Readability Analysis*. Technical report, Projet number 7-0052, U.S. Office of Education, Bureau of Research, Department of Health, Education and Welfare, Washington, DC.

Bulté, Bram & Alex Housen. 2012. Defining and operationalising L2 complexity. In Housen Alex, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 21–46. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Chall, Jeanne S. & Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge: Brookline Books.

Charniak, Eugene. 2000. A maximum-entropyinspired parser. *In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*. 132–139.

Coleman, Edmund B. 1965. *On Understanding Prose: Some Determiners of Its Complexity*. NSF Final Report GB2604, Washington, D.C, National Science Foundation.

Collins-Thompson, Kevyn. 2015. Computational assessment of text readability: A survey of current and future research. *ITL – International Journal of Applied Linguistics* 165 (2). 97–135.

Crossley, Scott A., Philip M. Mccarthy, David F Duffy & Danielle McNamara. 2007. Toward a new readability: A mixed model approach. *In Proceedings of the 29th Annual Conference of the Cognitive Science Society*. 197–202.

Dale, Edgar & Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin* 27. 11–20, 37–54.

Dale, Edgar & Joseph O'Rourke. 1981. *Living Word Vocabulary*. Chicago: World Book – Childcraft International.

Danielson, Wayne A. & Sam D. Bryan. 1963. Computer automation of two readability formulas. *Journalism Quarterly* 40 (2). 201–205. https://doi.org/10.1177%2F107769906304000207

Daoust, François, Léo Laroche & Lise Ouellet. 1996. SATO-CALIBRAGE: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue Québécoise de Linguistique* 25 (1). 205–234.

Dascalu, Mihai. 2014. Analyzing discourse and text complexity for learning and collaborating. *In Analyzing Discourse and Text Complexity for Learning and Collaborating*, 1–3. Springer, Cham. https://doi.org/10.1007/978-3-319-03419-5

Flesch, Rudolf. 1948. A new readability yardstick. *Journal of Applied Psychology* 32 (3). 221–233. https://doi.org/10.1037/h0057532

Foltz, Peter W., Walter Kintsch & Thomas Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes* 25 (2). 285–307. https://doi.org/10.1080/01638539809545029

Gatiyatullina, Galya, Marina Solnyshkina, Valery Solovyev, Andrey Danilov, Ekaterina Martynova & Iskander Yarmakeev. 2020. Computing Russian morphological distribution patterns using RusAC Online Server. *In 13th International Conference on Developments in eSystems Engineering (DeSE)*. 393–398. https://doi.org/10.1109/DeSE51703.2020.9450753

Graesser, Arthur C. & Danielle S. McNamara. 2011. Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science* 3. 371–398.

Graesser, Arthur C., Matthew Singer & Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological Review* 101. 371–395.

Gray, William & William Leary. 1935. *What Makes a Book Readable*. University of Chicago Press, Chicago: Illinois.

Hall, Charles, Debra S. Lee, Gwenyth Lewis, Phillip M. McCarthy & Danielle S. McNamara. 2006. Language in law: Using Coh-Metrix to assess differences between American and English/Welsh language varieties. *In Proceedings of the Annual Meeting of the Cognitive Science Society* 28.

Heilman, Michael, Le Zhao, Juan Pino & Maxine Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. *In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. 80–88. https://doi.org/10.3115/1631836.1631846

Hendrix, Gary G. 1980. Future prospects for computational linguistics. *In ACL '80: Proceedings of the 18th Annual Meeting on Association for Computational Linguistics*. 131–135. Association for Computational Linguistics, United States. https://doi.org/10.3115/981436.981476

Jones, Michael N., Walter Kintsch & Douglas J. Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of Memory and Language* 55(4). 534–552.

Kemper, Susan. 1983. Measuring the inference load of a text. *Journal of Educational Psychology* 75 (3). 391–401.

Kintsch, Walter & Vipond Douglas. 1979. Reading comprehension and readability in educational practice and psychological theory. In Lars-Göran Nilsson (ed.), *Perspectives on memory research*, 329–365. Hillsdale, NJ, Lawrence Erlbaum.

Klare, George R. 1963. *The Measurement of Readability*. Iowa State University Press.

Kortmann, Bernd & Benedikt Szmrecsanyi (eds.). 2012. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: De Gruyter.

Laposhina, Antonina N. & Maria Yu. Lebedeva. 2021. Tekstometr: Online-instrument opredeleniya urovnya slozhnosti teksta po russkomu yazyku kak inostrannomu. *Rusistika* 19(3). 331–345. (In Russ.) http://dx.doi.org/10.22363/2618-8163-2021-19-3-331-345

Lively, Bertha & Sidney Pressey. 1923. A method for measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision* 9. 389–398.

Marujo, Luis, Jorge Baptista, José Lopes, Maxine Eskenazi, Ceu Viana, Juan Pino & Isabel Trancoso. 2009. Porting reap to European Portuguese. *In SLaTE*. 69–72. Citeseer.

McCall, William & Lelah Crabbs. 1925. *Standard Test Lessons in Reading*. New York: Teacher's College Press.

McCarthy, Philip M., John C. Myers, Stephen Briner & Arthur C. Graesser. 2009. A psychological and computational study of sub-sentential genre recognition. *JLCL* 24 (1). 23–55.

McClusky, Howard. 1934. A quantitative analysis of the difficulty of reading materials. *The Journal of Educational Research* 28. 276–282. https://doi.org/10.1080/00220671.1934.10880487

McLaughlin, G. Harry. 1969. Smog-grading – a new readability formula. *Journal of Reading* 13. 639–646.

McNamara, Danielle & Arthur C. Graesser. 2012. *Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing*. IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch011

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy & Zhiqiang Cai. 2014. *Coh-Metrix: Theoretical, Technological, and Empirical Foundations. In Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664.006

Meyer, Bonnie J. F. 1982. Reading research and the composition teacher: The importance of plans. *College Composition and Communication* 33 (1). 37–49. https://doi.org/10.2307/357843

Nelson, Jessica, David Liben, Meredith Liben & Charles Perfetti. 2012. *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance*. New York, NY: Student Achievement Partners.

Ojemann, Ralph. 1934. The reading ability of parents and factors associated with the reading difficulty of parent education materials. *University of Iowa Studies in Child Welfare* 8. 11–32.

Rabin, Mikhael'. 1993. Slozhnost' vychislenii. *In ACM Turing Award Lectures*. 371–391. Moscow: Mir. (In Russ.)

Rescher, Nicholas. 1998. *Complexity: A Philosophical Overview*. London: Transaction Publishers.

Rosch, Eleanor & Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7. 573–605.

Rubakin, Nikolai A. 1890. Notes on literature for the people. *Russkoe Bogatstvo* 10. 221–231. (In Russ.)

Saimon, Gerbert. 2004. *The Sciences of the Artificial*. Moscow: Editorial URSS. (In Russ.)

Schwarm, Sarah E. & Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. *In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. 523–530. https://doi.org/10.3115/1219840.1219905

Sheehan, Kathleen M., Irene Kostin, Diane Napolitano & Michael Flor. 2014. The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal* 115 (2). 184–209. https://doi.org/10.1086/678294

Sherman, Lucius A. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston: Ginn.

Si, Luo & Jamie Callan. 2001. A statistical model for scientific readability. *In Proceedings of the Tenth International Conference on Information and Knowledge Management*. 574–576. ACM New York, NY, USA. https://doi.org/10.1145/502585.502695

Simon, Herbert A. 1996. *The Sciences of the Artificial*. Cambridge: The MIT Press.

Smith, Edgar A. & John Quackenbush. 1960. Devereux teaching aids employed in presenting elementary mathematics in a special education setting. *Psychological Reports* 7. 333–336. https://doi.org/10.2466/PR0.7.6.333-336

Solnyshkina, Marina I., Elena V. Harkova & Aleksander S. Kiselnikov. 2014. Comparative Coh-metrix analysis of reading comprehension texts: Unified (Russian) state exam in English vs Cambridge first certificate in English. *English Language Teaching* 7 (12). 65–76. https://doi.org/10.5539/elt.v7n12p65

Solnyshkina, Marina I. & Kisel'nikov Aleksandr. S. 2015. Slozhnost' teksta: Etapy izucheniya v otechestvennom prikladnom yazykoznanii. *Vestnik Tomskogo Gosudarstvennogo Universiteta. Filologiya* 6(38). (In Russ.)

Solnyshkina, Marina I., Elena V. Harkova & Maria B. Kazachkova. 2020. The structure of Cross-Linguistic differences: Meaning and context of 'Readability' and its Russian

equivalent 'Chitabelnost'. *Journal of Language & Education* 6 (1). 103–119. https://jle.hse.ru/article/view/7176/12052. https://doi.org/10.17323/jle.2020.v6.i1

Solnyshkina, Marina I., Ehl'zara Gizzatullina-Gafiyatova, Ekaterina V. Martynova & Valery Solovyev. 2022. Text complexity as an interdisciplinary problem. *Voprosy Kognitivnoi Lingvistiki 1*. (In Russ.)

Solovyev, Valery D., Vladimir V. Ivanov & Marina I. Solnyshkina. 2018. Assessment of reading difficulty levels in Russian academic texts: Approaches and Metrics. *Journal of Intelligent & Fuzzy Systems* 34 (5). 3049–3058. https://doi.org/10.3233/JIFS-169489

Solovyev, Valery, Marina Solnyshkina, Vladimir Ivanov & Ildar Batyrshin. 2019. Prediction of reading difficulty in Russian academic texts. *Journal of Intelligent & Fuzzy Systems* 36 (5). 4553–4563. https://doi.org/10.3233/JIFS-179007

Solovyev, Valerii, Yulia Volskaya, Maria Andreeva & Artem Zaikin. 2022. Russian dictionary with concreteness/abstractness indexes. *Russian Journal of Linguistics* 2. 514–548. (In Russ.)

Spivey, Nancy N. 1987. Construing constructivism: Reading research in the United States. *Poetics* 16 (2). 169–192. https://doi.org/10.1016/0304-422X%2887%2990024-6

Steger, Maria & Edgar W. Schneider. 2012. Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In Kortmann Bernd & Benedikt Szmrecsanyi (eds.), *Linguistic complexity: Second language acquisition, indigenization, contact*, 156–191. Berlin: De Gruyter.

Stevens, Kathleen C. 1980. Readability Formulae and McCall-Crabbs Standard Test Lessons in Reading. *The Reading Teacher* 33 (4). 413–415.

Sun, Haimei. 2020. Unpacking reading text complexity: A dynamic language and content approach. *Studies in Applied Linguistics & TESOL at Teachers College* 20 (2). 1–20. https://doi.org/10.7916/salt.v20i2.7098

Taylor, Wilson L. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly* 30 (4). 415–433. https://doi.org/10.1177%2F107769905303000401

Thorndike, Edward. 1921. Word knowledge in the elementary school. *The Teachers College Record* 22 (5). 334–370.

van Dijk, Teun A. & Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. New York: Academic.

Vergara, Fermina & Rachelle Lintao. 2020. War on drugs: The readability and comprehensibility of illegal drug awareness campaign brochures. *International Journal of Language and Literary Studies* 2 (4). 98–121. https://doi.org/10.36892/ijlls.v2i4.412

Vogel, Mabel & Carleton Washburne. 1928. An objective method of determining grade placement of children's reading material. *The Elementary School Journal* 28 (5). 373–381. https://doi.org/10.1086/456072

Zwaan, Rolf A. & Gabriel A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin* 123. 162–185. https://doi.org/10.1037/0033-2909.123.2.162

Zeno, Susan, Robert T. Millard & Raj Duvvuri. 1995. *The Educator's Word Frequency Guide*. Brewster: Touchstone Applied Science Associates, Inc.

Internet Resources / Электронные ресуры

Antonini, Alessio, Francesca Benatti, Edmund King, François Vignale & Guillaume Gravier. 2019. *Modelling Changes in Diaries, Correspondence and Authors' Libraries to Support Research on Reading: The READ-IT Approach*. URL: https://hal.archives-ouvertes.fr/hal-02130008/document (accessed 25 January 2022).

Antunes, Hélder M. M. 2019. *Automatic Assessment of Health Information Readability.* URL: https://repositorio-aberto.up.pt/bitstream/10216/121810/4/345408.pdf (accessed 25 January 2022).

Development of the ATOS Readability Formula. 2014. URL: https://webcache.googleusercontent.com/search?q=cache:lWV4zvGcnhMJ:https://doc.renlearn.com/KMNet/R004250827GJ11C4.pdf+&cd=14&hl=ru&ct=clnk&gl=ru (accessed 25 January 2022).

François, Thomas & Hubert Naets. 2011. Dmesure: A readability platform for French as a foreign language. URL: https://cental.uclouvain.be/team/tfrancois/articles/CLIN21.pdf (accessed 25 January 2022).

Lennon, Colleen & Hal Burdick. 2004. *The Lexile Framework as an Approach for Reading Measurement and Success*. URL: http://www.lexile.com/m/resources/materials/Lennon__Burdick_2004.pdf (accessed 25 January 2022).

Renaissance. 2022. URL: https://ukhosted43.renlearn.co.uk/2171850/ (accessed 25 January 2022).

Special Collections. Accelerated Reader (ATOS Level: 5.0-5.9). Bookshare a Benetech Initiative. 2002–2022. URL: https://www.bookshare.org/browse/collection/371895 (accessed 25 January 2022).

T.E.R.A.: The Coh-Metrix Common Core Text Ease and Readability Assessor. 2012–2022. URL: http://129.219.222.70:8084/Coh-Metrix.aspx (accessed 25 January 2022).

The ATOS Readability Formula for Books and How it Compares to Other Formulas. 2000. URL: https://files.eric.ed.gov/fulltext/ED449468.pdf (accessed 25 January 2022).

The Lexile Framework for Reading. 2022. URL: https://lexile.com (accessed 25 January 2022).

Article history:
Received: 20 October 2021
Accepted: 06 February 2022

Bionotes:

Marina I. SOLNYSHKINA is Doctor Habil. (Philology), Professor of the Department of Theory and Practice of Foreign Language Teaching, Head of "Text Analytics" Research Lab at the Institute of Philology and Intercultural Communication of Kazan Federal University (Russia). Her research interests include linguistic complexology, corpus linguistics, and lexicography.

Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya str., Kazan, 420008, Russia
e-mail: mesoln@yandex.ru
ORCID: 0000-0003-1885-3039

Danielle S. MCNAMARA, Ph.D., is Professor of Psychology in the Psychology Department and Senior Scientist at Arizona State University. She is an international expert in the fields of cognitive science, comprehension, natural language processing, and intelligent systems.

Contact information:
Arizona State University
Payne Hall, TEMPE Campus, Suite 108, Mailcode 1104, the USA
e-mail: Danielle.McNamara@asu.edu
ORCID: 0000-0001-5869-1420

Radif R. ZAMALETDINOV is Doctor Habil. (Philology), Professor, Director of the Institute of Philology and Intercultural Communication of Kazan Federal University. His research interests embrace cognitive linguistics, linguoculturology, comparative linguistics, history and patterns of functioning of the Tatar and Russian languages, and bilingualism.

Contact information:
Kazan Federal University
18 Kremlevskaya str., Kazan, 420008, Russia
e-mail: director.ifmk@gmail.com
ORCID: 0000-0002-2692-1698

Сведения об авторах:
Марина Ивановна СОЛНЫШКИНА – доктор филологических наук, профессор кафедры теории и практики преподавания иностранных языков, руководитель НИЛ «Текстовая аналитика» Института филологии и межкультурной коммуникации Казанского федерального университета (Россия). Сфера ее научных интересов включает лингвистическую комплексологию, корпусную лингвистику и лексикографию.
Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: mesoln@yandex.ru
ORCID: 0000-0003-1885-3039

Даниэль С. МАКНАМАРА – доктор наук, профессор кафедры психологии Университета штата Аризона, психолингвист, международный эксперт в области когнитивистики, понимания, обработки естественного языка и интеллектуальных систем.
Контактная информация:
Даниэль С. МакНамара
Университет штата Аризона
Пэйн Холл, Кампус TEMPE, ком. 108, 1104, США
e-mail: Danielle.McNamara@asu.edu
ORCID: 0000-0001-5869-1420

Радиф Рифкатович ЗАМАЛЕТДИНОВ – доктор филологических наук, профессор, директор Института филологии и межкультурной коммуникации Казанского федерального университета. В сферу его научных интересов входят когнитивная лингвистика, лингвокультурология, сопоставительное языкознание, история и закономерности функционирования татарского и русского языков, билингвизм.
Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: director.ifmk@gmail.com
ORCID: 0000-0002-2692-1698

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 342–370
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30145

Research article

ReaderBench: Multilevel analysis of Russian text characteristics

Dragos CORLATESCU[1] ✉, Ștefan RUSETI[1] and Mihai DASCALU[1,2]

[1] *University Politehnica of Bucharest, Bucharest, Romania*
[2] *Academy of Romanian Scientists, Bucharest, Romania*
✉dragos.corlatescu@upb.ro

Abstract
This paper introduces an adaptation of the open source ReaderBench framework that now supports Russian multilevel analyses of text characteristics, while integrating both textual complexity indices and state-of-the-art language models, namely Bidirectional Encoder Representations from Transformers (BERT). The evaluation of the proposed processing pipeline was conducted on a dataset containing Russian texts from two language levels for foreign learners (A – Basic user and B – Independent user). Our experiments showed that the ReaderBench complexity indices are statistically significant in differentiating between the two classes of language level, both from: a) a statistical perspective, where a Kruskal-Wallis analysis was performed and features such as the "nmod" dependency tag or the number of nouns at the sentence level proved the be the most predictive; and b) a neural network perspective, where our model combining textual complexity indices and contextualized embeddings obtained an accuracy of 92.36% in a leave one text out cross-validation, outperforming the BERT baseline. ReaderBench can be employed by designers and developers of educational materials to evaluate and rank materials based on their difficulty, as well as by a larger audience for assessing text complexity in different domains, including law, science, or politics.
Keywords: *ReaderBench framework, text complexity indices, language model, neural architecture, multilevel text analysis, assessing text difficulty*

For citation:
Corlatescu, Dragos, Ștefan Ruseti & Mihai Dascalu. 2022. ReaderBench: Multilevel analysis of Russian text characteristics. *Russian Journal of Linguistics* 26 (2). 342–370. https://doi.org/10.22363/2687-0088-30145

Научная статья

ReaderBench: многоуровневый анализ характеристик текста на русском языке

Драгош КОРЛАТЕСКУ[1] ✉, Штефан РУСЕТИ[1], Михай ДАСКАЛУ[1,2]

[1]*Политехнический университет Бухареста, Бухарест, Румыния*
[2]*Академия румынских ученых, Бухарест, Румыния*
✉dragos.corlatescu@upb.ro

Аннотация
В статье представлена новая версия платформы ReaderBench с открытым исходным кодом. В настоящее время Readerbench поддерживает многоуровневый анализ параметров текстов на русском языке, интегрируя при этом как индексы текстовой сложности, так и современные языковые модели, в частности, BERT. Оценка предлагаемого алгоритма обработки проводилась на корпусе русских текстов двух языковых уровней, используемых при обучении русскому языку как иностранному (A – базовый пользователь и B – независимый пользователь). Наши эксперименты показали, что (а) индексы сложности текстов различных уровней по Общеевропейской шкале, рассчитываемые при помощи ReaderBench, статистически значимы (по критерию Краскела-Уоллиса), при этом количество существительных на уровне предложения оказалось наилучшим предиктором сложности; б) а наша нейронная модель, сочетающая индексы сложности текста и контекстуализированные вложения, при перекрестной валидации достигла точности 92,36 % и превзошла базовый уровень BERT. ReaderBench может использоваться разработчиками учебных материалов для оценки и ранжирования текстов в зависимости от их сложности, а также более широкой аудиторией для оценки сложности восприятия текста в различных областях, включая юриспруденцию, естествознание или политику.

Ключевые слова: *фреймворк ReaderBench, индексы сложности текста, языковая модель, нейронная архитектура, многоуровневый анализ текста, оценка сложности текста*

Для цитирования:
Corlatescu D., Ruseti Ș, Dascalu M. ReaderBench: Multilevel analysis of Russian text characteristics. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 342–370. https://doi.org/10.22363/2687-0088-30145

1. Introduction

The Natural Language Processing (NLP) field focuses on empowering computers to process and then understand written or spoken language texts in order to perform various tasks. The performance of Artificial Intelligence or Machine Learning approaches on common NLP tasks has increased over the years, but there are still many tasks where computers are far from human performance. Nonetheless, the processing speed of computer programs is not to be neglected, and the current tradeoff between the response time of an algorithm and its errors is shifting the balance towards automated analyses – for example, a human invests tens of hours to correctly extract all the parts of speech from a novel, while the computer can perform the same task in a couple of minutes, with an error of only 1–5% mislabeled

words. As such, NLP tools are becoming more widely used to provide valuable inputs to further develop and test various hypotheses.

Tailoring reading materials for learners is a practical and essential field where NLP tools can have a high impact. Designing such materials can prove to be a difficult task since texts below readers' level of understanding will make them lose interest, while texts too difficult to comprehend will demotivate learners. Automated NLP frameworks provide valuable insights in those situations, especially the ones that focus on identifying the complexity of a text. One such tool is the ReaderBench (Dascalu et al. 2013) framework, which previously supported other languages besides English, namely French (Dascalu et al. 2014), Dutch (Dascalu et al. 2017), and Romanian (Gifu et al. 2016), and has now been adapted to also support Russian.

The new version of ReaderBench[1] is a Python library that extracts multilevel textual characteristics from texts in multiple languages. These characteristics, named also textual complexity indices, provide valuable insights of text difficulty on multiple levels, namely surface, word, morphology, syntax, and semantics (i.e., cohesion), all described in the following sections. The purpose of this study is to present the adaptation process of ReaderBench to support the Russian language, starting from the computation of Russian complexity indices, and followed by the integration of new methods for building the Cohesion Network Analysis (CNA, Dascalu et al. 2018) graph using state-of-the-art language models, namely Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019).

Various neural network architectures and statistical analyses were employed to assess the performance of our processing pipeline. Our experiment uses Russian texts from two language level groups that reflect an individual's language proficiency: A (Basic User) and B (Independent User). The corpus is a part of the Russian as a Foreign Language Corpus (RuFoLC) compiled by language experts from the "Text Analysis" laboratory, Kazan Federal University. Our goal is to build an automated model and to perform statistical analyses of the texts in order to differentiate between the two classes, while assessing the importance of the textual complexity indices in making this decision.

2. Assessing Russian text complexity

The manner in which people understand and study languages has changed during the last two centuries; Russian is no different. A brief history of the approaches used to analyze textual complexity in Russian texts is presented by Guryanov et al. (2017). The authors documented that such analyses were conducted by linguists mostly by hand in the beginning of the 20th century. Even though the key terms such as readability or text complexity were not completely defined, the general understanding of the concepts existed and simple indices, such as word

[1] https://github.com/readerbench/ReaderBench

length or number of words, were considered. Moving on to the end of the 20th century – beginning of the 21st century, researchers started to include semantic features, such as, for example: the polysemy of the words. In recent times, additional features were introduced, detailed in the next subsection dedicated to automated measures of text complexity.

One important part of research on text complexity revolves around its educational theme: are texts appropriate in terms of complexity for the students reading or studying them? Linguists can provide an expert opinion to this question; however, this requires a lot of resources, including a considerable amount of time. Thus, a system that can provide meaningful insights into the difficulties encountered when reading a text is desirable.

McCarthy et al. (2019), who are language experts, developed a Russian language test to assess text comprehension. The test was conducted on approximately 200 students (~100 fifth graders, ~100 ninth graders) and the results showed that they struggle to understand the ideas of the texts. Additionally, the paper provided an overview of the entire evaluation process in the Russian educational system, and it offered a viable evaluation alternative designed by linguists in the form of a test.

One of the initial papers on the same matter, but written from a more statistical perspective, was the work by Gabitov et al. (2017). In their study, the problem of text complexity in Russian manuals was addressed. Specifically, the investigation focused on the 8th grade manual on social studies made by Bogolyubov. All analyses were performed mostly manually, starting from selecting 16 texts from the book and then computing readability indices formulas, such as Flesch-Kincaid, Coleman-Liau, Dale-Chale readability formula, Automated Readability Index, and Simple Measure of Gobbledygook (SMOG). The unevenness of those indices across the texts raised questions whether the texts were suitable for students and represented the underlying reason for further research in this domain.

The syntactic complexity of social studies texts was explored by Solovyev et al. (2018). The authors used ETAP-3 (Boguslavsky et al. 2004), a syntactic analyzer for Russian grammar, to compute the dependency parse tree for each sentence. Fourteen indices were extracted based on the dependency tree that looked at key components of the Russian sentence structure in order to deduct its complexity, namely the length of the path between two nodes and various counts of nodes, leaves, verbal participles, verbal adverb phrases, modifiers in a nominal group, syndetic elements, participial constructs, compound sentences, coordinating chains, subtrees, and finite dependent verbs. Their statistical analyses showed high correlation between the extracted features and grade level; however, syntactic features were less correlated than the lexical ones.

Solovyev et al. (2020) also explored how predictive specific quantitative indices were in ranking academic Russian texts and in determining their complexity. Their corpus was composed of texts from the field of Social Studies grouped by grade level, i.e. 5th–11th grades. The texts were extracted from manuals

written by two authors (Bogolyubov and Nikitin) used at that time for teaching social studies. The corpus required a preprocessing step, where the parts of speech were extracted using TreeTagger (Schmid et al. 2007) for Russian, the texts were split into sentences, and outliers (i.e., sentences that were even too short or too long) were eliminated. The following indices were used in their analysis: Flesch-Kincaid Grade, Flesch Readability Ease, frequency of content words, average words per sentence, average syllables per word, and additional features based on the part of speech tags (such as the number of nouns or verbs). The authors performed a statistical analysis using both Pearson (1895) and Spearman (1987) coefficients to inspect the correlation between the indices and the complexity of the texts (i.e., their grades level). All features proved to be statistically significant, except for "average words per sentence" and "average syllables per word". Additionally, the authors proposed slightly modified formulas for the Flesch-Kincaid Grade and Flesch Readability Ease that better reflect the field of Social Studies.

Further studies of quantitative indices on the corpus containing texts from Social Sciences manuals, Churunina et al. (2020) introduced new indices such as type-token ratio (TTR), abstractness index, and words frequency based on Sharoff's dictionary (Sharoff et al. 2014) that proved to be statistically significant in differentiating the grades of the texts. Out of the specified indices, abstractness, was proven to be closely related to textual complexity. In fact, the study by Sadoski et al. (2000) claimed that the concreteness (which is the opposite of abstractness) is the most predictive feature for comprehensibility. As a follow-up, Solovyev et al. (2020) provided an in-depth analysis of the abstractness of words in the Russian Academic Corpus (RAC, Solnyshkina et al. 2018) and in a corpus containing students recalls of academic texts. The core of the experiments was the Russian dictionary of concrete/abstract words (RDCA, Akhtiamov 2019). A notable result was obtained in terms of students recall, where texts provided by students used more concrete words than the original ones, underlining the idea that abstract terms are harder to digest.

Quantitative indices provide significant insights into the textual complexity of writings, but they are not the only concept that can be applied to analyze text difficulty. One example can be topic modelling, as applied in an experiment performed by Sakhovskiy et al. (2020) on the Social Studies corpus. The authors implemented Latent Dirichlet Allocation with Additive regularization of topic models (ARTM, Vorontsov & Potapenko 2015). Topics were extracted at three granularity levels: paragraph, segment (i.e., sequences of 1000 words maximum), and full text level. The topics were manually verified by linguist experts, and they were further used in an experiment to determine the correlations between topics and grades of the texts, in four different ways: a) correlation between grade and topic weight, b) correlation between grade and the distance between topic words in a semantic space, c) correlation between grade and topic coherence, and d) correlation between topic properties and complexity-based topic proportion growth. The conclusion of their study highlighted that topic models can be successfully used to assess text complexity.

3. Textual complexity as a Natural Language Processing task

Readability reflects the level of easiness in understanding of a text. Extracting features using NLP techniques is a common approach, when exploring the readability of a given text. There are multiple tools readily available; however, most of them support only English. Nevertheless, the underlying ideas can be extrapolated to other languages, as well. We further describe recent tools that cover the most frequently integrated textual complexity indices and that are also present to some extent in the Russian version of ReaderBench.

One of the first freely available systems is Coh-Metrix (Graesser et al. 2004) which is at its 3rd version at present. Coh-Metrix provides 108 textual complexity indices from eleven categories: descriptive, text easability principal components scores, referential cohesion, LSA, lexical diversity, connectives, situation model, syntactic complexity, syntactic pattern density, word information, and readability. The framework can be freely accessed on a website, but the code is not open-sourced. Coh-Metrix offers support for other languages than English, namely Traditional Chinese, while adaptations for other languages exist – for example, Coh-Metrix-Esp (Quispesaravia et al. 2016) for Spanish.

The Automatic Readability Tool for English (ARTE, Choi 2020) is a Java library available on all platforms that processes plain text files and outputs a CSV file with all the computed indices. The list of indices includes the Flesch Reading Ease Formula (Flesch 1949), Flesch Kincaid Grade Level Formula (Kincaid et al. 1975), and Automated Readability Index (Senter & Smith 1967), which take into consideration the average number of words per sentence, the average number of syllables per word, and the difference between them consisting of the weights for each parameter. Other examples of indices are SMOG Grading (Mc Laughlin 1969) and the New Dale-Chall Readability Formula (Dale & Chall 1948). Lastly, there are multiple "crowdsourced" indices that are computed by aggregation of different counts and statistics from other libraries.

The following library is the Constructed Response Analysis Tool (CRAT, Crossley et al. 2016) which provides over 700 indices that also take into consideration text cohesion. The indices are grouped in specific categories, namely: a) indices that count or compute percentages for words, sentences, paragraphs, content words, function words, and parts of speech, or b) indices based on the MRC Psycholinguistic Database (Coltheart 1981), the Kuperman Age of Acquisition scores (Kuperman et al. 2012), the Brysbaert Concreteness scores (Brysbaert et al. 2014), the SUBTLEXus corpus (Brysbaert et al. 2012), the British National Corpus (BNC, BNC Consortium 2007), the COCA corpus (Davies 2010). Complementary, the Custom List Analyzer (CLA, Kyle et al. 2015) is a library written in Python that computes various occurrences of text sequences (i.e., a word, an n-gram, or a wildcard) in a corpus.

The Grammar and Mechanics Error Tool (GAMET, Crossley et al. 2019) is a Java library that identifies errors in a plain text file from the perspective of grammar, spelling, punctuation, white space, and repetitions. The core of the library

integrates two packages, one from Java, Java LanguageTool (LanguageTool 2021), and one from Python, language-check (Myint 2014). The GAMET project was also tested and evaluated on two datasets (Crossley et al. 2019): a) a TOEFL-iBT corpus containing 480 essays written by English as a Second Language Learners, and b) 100 essays written by high school students in the Writing Pal Intelligent Tutoring System project (Roscoe et al. 2014). The errors reported by GAMET were evaluated by two expert raters, and the results showed that GAMET offered relevant feedback throughout the experiments.

Next, we explore a collection of four tools (TAACO, TAALEED, TAALES and TAASC) that cover a wide spectrum of analysis levels. All the tools have a graphical interface that accepts plain text files as input to produce CSV files with all indices as outputs. First, the Tool for the Automatic Analysis of Cohesion is a framework that focuses on text cohesion. The indices are separated into multiple categories: a) TTR and Density, where TTR stands for type-token ratio computed as the number of unique words/lemmas in a category, divided by the total number of words/lemma in the same category; b) Sentence overlap, where statistics regarding the repetition of the same word with certain properties in the following sentences are computed; c) Paragraph overlap, which is similar to the sentence overlap, only that the metrics are computed at paragraph level; d) Semantic overlap, where the scores of similarity between adjacent blocks (sentences and paragraphs) are computed on three methods: Latent semantic analysis (Landauer et al. 1998), Latent Dirichlet allocation (LDA, Blei et al. 2003), and word2vec (Mikolov et al. 2013); e) Connectives, where statistics are computed based on the types of the English connectives (e.g. conjunctions, disjunctions); f) Givenness, which is a measure of new information in the context of previous information, based on pronouns counts and repeated content lemmas. Second, the Tool for the Automatic Analysis of Lexical Diversity (TAALED, Kyle et al. 2021) provides 9 indices for measuring the language diversity of a text.

Third, the Tool for the Automatic Analysis of Lexical Sophistication (TAALES, Kyle et al. 2018) offers 484 indices addressing lexical sophistication divided into 4 major categories: a) Academic Language containing wordlists and formulas based on counts and percentages of words, b) indices based on the COCA corpus, c) indices based on other corpora (BNC, MRC, SUBTLEXus), and d) other types of indices, such as Age of Exposure or Contextual Distinctiveness. Fourth, the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC, Kyle 2016) focuses on analyzing the sentence components and the relations between them. It provides statistics at the clause and noun phrase level for measuring complexity. The syntactic sophistication is computed based on indices that focus on verbs and lemmas.

Textstat (Bansal 2014) is a Python library available online on the pypi archives which provides textual complexity indices for multiple languages. Textstat includes 16 indices, out of which most are English readability formulas: Flesch Reading Ease, Flesch Kincaid grade, SMOG, Coleman Liau, Automated Readability, and Dale Chall.

ReaderBench (Dascalu et al. 2013) is an open-source framework that offers multiple natural language processing tools. ReaderBench was initially developed in Java, but the library migrated to Python given that all major NLP frameworks, including Tensorflow (Abadi et al. 2016), Scikit learn (Pedregosa et al. 2011), spaCy (Honnibal & Montani 2017), and Gensim (Rehurek & Sojka 2010) are written in Python to enable Graphics Processing Unit (GPU) optimizations. ReaderBench is grounded in Cohesion Network Analysis (CNA, Dascalu et al. 2018), a method similar to Social Network Analysis, but instead of representing relations between people or entities, the CNA graph contains links between text elements. The weights of the links are given by the semantic similarity between the components using different semantic models, such as LSA, LDA, or word2vec. Both local and global cohesion are computed based on the strength of intra and inter-paragraph edges extracted from the CNA graph. The library comes with a demo website (Gutu-Robu et al. 2018), making it available to multiple audiences. On one hand, the Python library can be installed and used by machine learning/NLP developers using the Pip library archives[2]). On the other hand, the website provides multiple interactive interfaces, where linguists or any other person interested in studying text can perform their own analysis using the capabilities of the library, without having any programming knowledge – demos include for example: Multi document CNA (i.e., a detailed analysis and visualization of multiple documents grounded in Cohesion Network Analysis), Keywords extraction (i.e., a list and a graph of the keywords from a text), AMoC (Automated Model of Comprehension, a model that simulates reading comprehension), Sentiment Analysis (i.e., extracting the polarity of a text in terms of expressed sentiments), and Textual Complexity (i.e., provide an export of the complexity indices applied on the input text). All publicly available analyses cover multiple languages, not just English, and all the additional information required for each experiment is also present on the website. In this study, we focus on the extension of the framework to also accommodate textual complexity indices and prediction models for Russian texts.

It is important to note that ReaderBench provides a viable alternative to all other previously mentioned software for text analysis. ReaderBench leverages state of the art NLP models to explore the semantics of texts and was effectively employed in various comprehension tasks, in multiple languages including English, French, Dutch, and Romanian. The project is open sourced under an Apache 2 license, the library can be easily integrated into multiple Python projects, whereas the presentation website can be used freely for the remote processing of texts.

4. Current study objectives

Our study focuses on an in-depth multilevel analysis of Russian texts by employing textual complexity indices and the CNA graph updated with language models, together with neural network models and statistical analyses, all integrated

[2] https://pypi.org/project/rbpy-rb/

into the ReaderBench framework. As such, we assess to what extent the Russian textual complexity indices integrintoed into ReaderBench are predictive of the differences between Russian texts from two language levels (i.e., A – Basic User and B – Independent User). We perform this analysis to explore the predictive power of our models and underline the most predictive features for this task.

5. Method

Corpus

This study considers Russian texts from two language levels for foreign learners (A-Basic User and B-Independent User) with the aim to predict a text's difficulty class. The selection of texts in terms of complexity assessment was performed by Russian linguists, members of the "Text Analytics" Laboratory from the Kazan Federal University. The corpus used in the follow-up experiments is a subpart of the Russian as a Foreign Language Corpus (RuFoLC). The initial corpus was in a raw format containing texts from 3 language levels A1 (Breakthrough or beginner), A2 (Waystage or elementary), and B1 (Threshold or intermediate). However, since only 3 texts were available for the A1 level, we decided to merge the A1 and A2 together (see Table 1 for corpus statistics). Since the overarching number of examples was too low for a neural network to learn meaningful representations, we decided to use paragraphs as input in order to ensure an increased number of samples.

Table 1. **Language levels corpus statistics.**

Class	# Documents	# Paragraphs	# Sentences	# Words
A	37	465	1663	18,307
B	48	333	1105	13,741

The ReaderBench Framework adapted for Russian

A specific set of resources is required for a new language to be integrated into ReaderBench. From this list, part are mandatory, while others are nice to have. One mandatory requirement is to have a language model available in spaCy (Honnibal & Montani 2017), an open-source library written in Python that offers support for NLP pre-processing tasks, such as part of speech tagging, dependency parsing, and named entity recognition. SpaCy offers a unified pipeline structure for any language and, at the moment of writing, spaCy reached version 3.1 with support for 18 languages, including Russian which has been integrated for reproducibility reasons. Additionally, spaCy includes a multi-language model that can be used for any language, but with lower performance. All languages have multiple models (i.e., small, medium, and large) available to address memory or time constraints. Smaller models are faster to run and require fewer resources, but yield lower performance.

Semantic models are a key component for the ReaderBench pipeline and for building the CNA graph. All indices that are calculated based on the meaning of the

words, the relations between words, sentences and paragraphs need a semantic language model. ReaderBench generally uses word2vec as a language model because it is available for most languages from multiple sources. During the development of this paper, we also considered it fit to align the semantic models across all the languages available in ReaderBench. Thus, we added support for the MUSE (Conneau et al. 2018) version of word2vec, where the semantic spaces are similar across the three languages.

Previous versions of ReaderBench used to compute similarity scores between textual elements from the CNA graph using LSA, LDA, and word2vec; however, these models have been outperformed by BERT-based (Devlin et al. 2019) derivatives. The Transformer architecture introduced by Vaswani et al. (2017) obtained state of the art results in most NLP tasks, especially with its encoder component, namely the Bidirectional Encoder Representations from Transformers (BERT). The original BERT was trained on two tasks: language modeling (where 15% of the tokens were masked and the model tried to predict the best word that fitted the mask, given the context) and next sentence prediction (given a pair of sentences, the model tried to predict if the second sentence made sense to follow the first sentence). The language modeling component is used to represent words in a latent vector space.

Nowadays, almost all languages have a custom BERT model available, and Russian is no exception. The ReaderBench library now integrates the DeepPavlov rubert-base-cased (Kuratov & Arkhipov 2019) BERT-base model to compute contextualized embeddings. It is important to note that this is the first study in which ReaderBench indices are computed using BERT-based embeddings.

Besides the above-mentioned libraries and models, ReaderBench can also benefit from specific word lists which were adapted for Russian, including: list of stop words (i.e., words with no semantic meaning ignored in preprocessing stages), list of connectives and discourse markers, and list of pronouns grouped by type and person; all previous word lists were provided by Russian linguists.

Additional improvements were made to the ReaderBench Python codebase, including performance optimizations and a refactoring to provide a more efficient and cleaner implementation of the textual complexity indices. New cohesion-centered textual complexity indices were added in ReaderBench, as well as a new aggregation function on top of them – the maximum value at a certain granularity level (more details are presented in the next section).

Textual Complexity Indices for Russian

The textual complexity indices provided by ReaderBench ensure a multilevel analysis of text characteristics and are grouped by their scope (see dedicated Wiki page[3]). Table 2–6 present the names of the indices, their description, what component or components from the above enumeration are used, as well as availability in terms of granularity. Note that, as previously mentioned, all indices

[3] https://github.com/readerbench/ReaderBench/wiki/Textual-Complexity-Indices

require the spaCy pre-processing pipeline to be executed; thus, SpaCy does not appear as a dependency. The "Granularity" column reflects four possible levels on which the index is calculated: Document (D), Paragraph (P), Sentence (S), or Word (W). In general, the value of one level of granularity is computed recursively as a function of values coming from one level below. For example, word counts are calculated at the sentence level by considering word occurrences from each sentence; follow-up at paragraph level, we report the count of the words from all sentences belonging to a targeted paragraph. The final values presented as indices are the results of three aggregation functions: mean (abbreviated "M"), standard deviation (abbreviated "SD"), and maximum (abbreviated "Max"). Thus, an index can look like "M(Wd / Sent)", which can be translated as the mean value of words per sentence in a text. In terms of consistency across languages, all ReaderBench indices, their acronyms and descriptions, are provided in English.

The surface indices available in ReaderBench are presented in Table 2. These indices are computed using simple algorithms that involve counting appearances of words, punctuations, and sentences. Starting from the Shannon's Information Theory (Shannon 1948), the idea of entropy at word level is also included as an index; the hypothesis is that a more varied vocabulary (i.e., higher entropy) may result in a more difficult text to understand.

Table 2. **ReaderBench Surface indices**

Abbreviation	Description	Dependencies	Granularity			
			D	P	S	W
Wd	Words	-	X	X	X	
UnqWd	Unique words	-	X	X	X	
Comma	Commas	-	X	X	X	
Punct	Punctuation marks (including commas)	-	X	X	X	
Sent	Sentences	-	X	X		
WdEnt	Word Entropy	-	X	X	X	

The morphology category (see Table 3) contains indices computed using the part of speech tagger from spaCy. Statistics are computed for each part of speech (e.g., nouns, verbs), while more detailed statistics are considered for sub-types of pronouns provided by linguists as predefined lists.

Table 3. **ReaderBench Morphology indices**

Abbreviation	Description	Dependencies	Granularity			
			D	P	S	W
PosMain	Words with a specific POS	-	X	X	X	
UnqPosMain	Unique words with a specific POS	-	X	X	X	
Pron	Specific pronoun types	Pronoun lists	X	X	X	

From the syntax point of view (see Table 4), ReaderBench provides indices derived from the dependency parsing tree. An index is computed for each dependency type available in the spaCy parser, such as "nsubj" or "cc". The depth

of the parsing tree is also an important feature in quantifying textual complexity: if the depth is high, then the text may become harder to understand.

Table 4. **ReaderBench Syntax indices**

Abbreviation	Description	Dependencies	Granularity			
			D	P	S	W
Dep	Dependencies of specific type	-	X	X	X	
ParseTreeDpth	Depth of the parsing tree	-			X	

Table 5 presents the indices that take into consideration text cohesion derived from the CNA graph. Cohesion is an important component when assigning text difficulty, as a lack of cohesion or cohesion gaps can make a text harder to follow (Dascalu 2014). As expected, a semantic model is required, either word2vec or the newly introduced BERT-base models. Note that the indices AdjSentCoh, AdjParCoh, IntraParCoh and InterParCoh were newly added to ReaderBench for this research.

Table 5. **ReaderBench Cohesion indices**

Abbreviation	Description	Dependencies	Granularity			
			D	P	S	W
AdjSentCoh	Cohesion between two adjacent sentences	Semantic Model	X	X		
AdjParCoh	Cohesion between two adjacent paragraphs	Semantic Model	X			
IntraParCoh	Cohesion between sentences contained within a given paragraph	Semantic Model	X	X		
InterParCoh	Cohesion between paragraphs	Semantic Model	X			
StartEndCoh	Cohesion between first and last text element	Semantic Model	X	X		
StartMiddleCoh	Cohesion between start and all middle text elements	Semantic Model	X	X		
MiddleEndCoh	Cohesion between all middle and last elements	Semantic Model	X	X		
TransCoh	Cohesion between the last sentence of the current paragraph and the first sentence from the upcoming paragraph	Semantic Model	X			

ReaderBench also provides statistics at individual words level (see Table 6). Name entity features are computed based on the Named Entity Recognizer from spaCy, while specific tags depend on the corpus on which the NER model was trained. For example, the Russian model is trained on a Wikipedia corpus and offers only 3 tags: location ("LOC"), organization ("ORG"), person ("PER"), while other models such as the English one offer 18 categories. This may affect the global statistics when comparing the complexity of texts from two languages, as observed in follow-up experiments. The syllables are computed using the "Pyphen" library for each language (Kozea 2016).

For other languages besides Russian, ReaderBench also includes additional textual complexity indices. For example, none of the Wordnet indices (e.g., sense counts, depths in hypernym trees) are currently available as the Russian WordNet (Loukachevitch et al. 2016) is in a different format when compared to the models integrated in Natural Language Toolkit (NLTK). Additionally, specific word lists

like Age of Acquisition, Age of Exposure, and discourse connectors are not yet available for Russian; as such, their corresponding indices are not computed.

Table 6. **ReaderBench Word indices**

Abbreviation	Description	Dependencies	Granularity			
			D	P	S	W
WdLen	Number of characters in a word	-				X
WdDiffLemma	Distance in characters between word (inflected form) and its corresponding lemma	-				X
Repetition	Number of occurrences of the same lemma	-	X	X	X	
NmdEnt	Number of specific types of named entity	Named Entity Recognizer	X	X	X	
Syllab	Number of syllables in a word	Rules or Dictionary				X

6. Neural Network Architectures combining Textual Complexity Indices and Language Models

Our first approach for predicting text difficulty involved using ReaderBench to extract the complexity indices available for the Russian language that were further introduced into a neural network depicted in Figure 1.a. The architecture started with an input layer which received the complexity indices for each text as a list. An optional layer with 128 units and Rectified Linear Unit ("RELU") activation function can be added to increase the complexity of the function computed by the neural network. Next, a dense layer with 32 units and with "RELU" as the activation function is used as a hidden layer. Finally, the output layer is a dense layer with only one output and the activation function 'sigmoid', which provides the class of the text.

Second, BERT and its derived models hold state-of-the-art results in multiple text classification tasks. Thus, we decided to test an architecture that uses only RuBERT, a BERT-base model trained for the Russian language. We obtained a semantic representation for each text by computing the mean of the last hidden state from the RuBERT output. Then, the embedding was feed into a neural network with an architecture similar to the previous one (see Figure 1.b).

Third, we tested a combination of the two inputs, as the RuBERT embeddings were concatenated with the ReaderBench indices and fed as input into the neural network. The architecture of the neural network can be observed in Figure 1.c

7. Statistical Analyses

A statistical approach was adopted to determine which features were significant in differentiating between textual complexity classes. The Shapiro normality test (Shapiro & Wilk 1965), as well as the skewness and kurtosis tests (Hopkins & Weeks 1990), were used to filter ReaderBench indices in terms of normality. Since most indices were not normally distributed, the Kruskal-Wallis analysis of variance (Kruskal & Wallis 1952) was employed to determine the statistical importance of the indices.

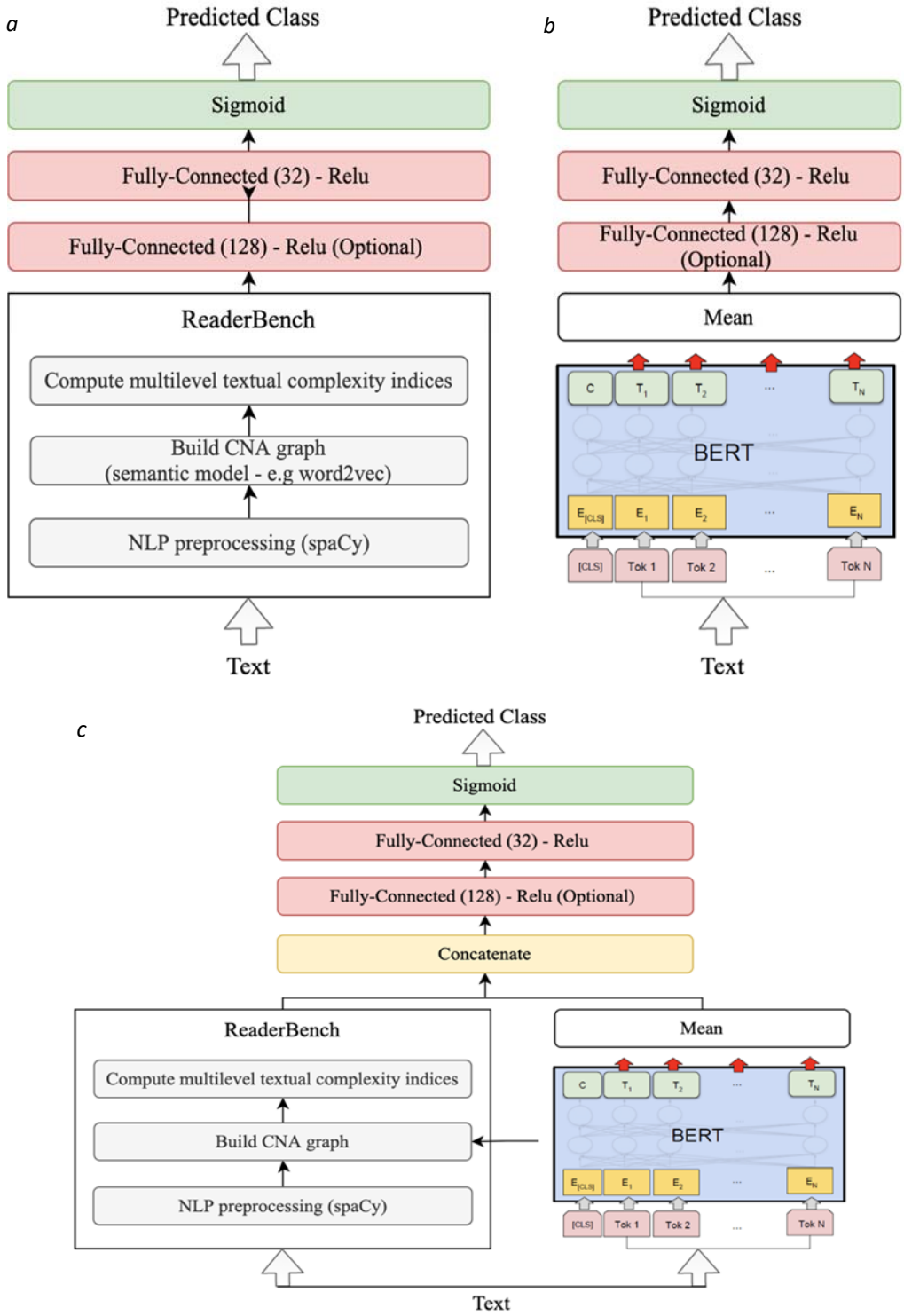

Figure 1. Neural network architectures:
a) Neural Network with ReaderBench indices as input; b) Neural Network with RuBERT embeddings as input; C) Neural network with both ReaderBench indices and RuBERT embeddings as input

8. Experimental Setup

The process of training neural networks requires the setup of hyperparameters. Thus, the Adam optimizer was considered with a learning rate of 1e-3. The loss function was binary cross-entropy, given that only two classes were predicted. Finally, each model was trained for 64 epochs with a batch size of 16.

The neural network architectures were used to classify the Russian texts on the two language levels: A (Basic user) and B (Independent user). The paragraphs were extracted from each text and labeled as the category the source text belonged to. We decided to perform cross-validation to evaluate the models due to the limited number of examples. There are multiple ways in which cross-validation can be performed, the most common ones being the 5-fold or 10-fold cross-validations. However, employing those methods involves limiting even more the input of the neural network, which in turn requires a substantial amount of data to be trained. Thus, given the limited number of entries, we elected to use a "leave-one-out" approach, where the entire corpus except a single entry is used for training a model at one iteration, followed by evaluation on the remaining entry. The process is repeated for each entry until the corpus is exhausted and performance is computed as the mean of all evaluation scores. Our corpus was composed of paragraphs and leaving one out would have meant that the other paragraphs from the same text would have been used in the training process which, again, could have generated biased. Thus, we decided to employ the "leave one text out" cross-validation. In this approach, an entire text (i.e., all the paragraphs belonging to the selected text) was left out, while the models were trained on all the other paragraphs. The final accuracy was reported as the mean of the results for each text.

9. Results

Table 7 depicts the results for the three neural architectures. The complexity indices from ReaderBench, as well as the RuBERT embeddings, were used as input to two different architectures: the first with only one hidden layer of 32 units, and the second with two hidden layers of 128 and 32 units. The scenario where the two input sources were combined is also presented.

Table 7. **Neural networks results**

Model Input Features	Hidden Layers	Leave one text out cross-validation (%)
Complexity Indices	1 hidden layer – 32 units	90.58
RuBERT	1 hidden layer – 32 units	87.49
Complexity Indices	2 hidden layers – 128, 32 units	87.05
RuBERT	2 hidden layers – 128, 32 units	88.69
Complexity Indices + RuBERT	1 hidden layer – 32 units	88.23
Complexity Indices + RuBERT	2 hidden layers – 128, 32 units	**92.36**

Table 8 presents a summary of the results obtained by applying Kruskal-Wallis test. The indices are divided by categories and subcategories, and each slot

introduces specific indices that are either statistically significant in differentiating the texts from the classes A and B or not. The notation is condensed and indices with the same characteristics are grouped using the "|" character. For example, the first entry considers the category "Surface" and subcategory word ("Wd"); the notation "M|Max(Wd / Doc|Par|Sent)" can be expanded to all the possibilities where "|" appears: M(Wd / Doc), M(Wd / Par), M(Wd / Sent), Max(Wd / Doc), Max(Wd / Par), Max(Wd / Sent). Additionally, in the "Dep" subcategory there is a list of dependency types that fitted the same pattern, and they are represented in a mathematical manner as a set. An important observation is that all features at document granularity were disregarded in this analysis, given the structure our data – i.e., all documents in the dataset contain only 1 paragraph; as such, the indices for the two granularities were the same. Similarly, the maximum values at paragraph level were ignored since the maximum and the mean values of only one entry are the same. An extended table with the descriptive statistics and corresponding $\chi 2$ and p values for all statistically significant textual complexity indices is provided in Appendix 1.

Table 8. **Summary of the predictive power of textual complexity indices.**

Indices Category	Indices Subcategory	Significant Indices ($p < .05$)	Not Significant Indices ($p > .05$)
Surface	Wd	M|Max(Wd / Sent), M(Wd / Par), SD(Wd / Sent)	SD(Wd / Par)
	UnqWd	M|Max(UnqWd / Sent), M(UnqWd / Par), SD(UnqWd / Sent)	SD(UnqWd / Par)
	Comma	M|Max(Commas / Sent), M(Commas / Par), SD(Commas / Sent)	SD(Commas / Par)
	Punct	M(Punct / Par), SD(Punct / Sent)	M|Max(Punct / Sent), SD(Punct / Par)
	Sent	M(Sent / Par)	SD(Sent / Par)
	WdEntr	M|Max|SD(WdEntr / Sent), M|SD(WdEntr / Par)	-
	NgramEntr	M|Max|SD(NgramEntr_2 / Word)	-
Morphology	POS	M|Max(POS_noun|_adj|_adv / Sent), M(POS_noun|_adj|_adv / Par), SD(POS_noun|_adj|_adv / Sent)	SD(POS_noun|_adj|adv / Par)
		SD(POS_pron / Sent)	M|Max(POS_noun / Sent), M(POS_noun / Par), SD(POS_noun / Par)
		M(POS_verb / Par), SD(POS_verb / Sent)	M|Max(POS_verb / Sent), SD(POS_verb / Par)
	UnqPOS	M|Max(UnqPOS_noun|_adj|_adv / Sent), M(UnqPOS_noun|_adj|_adv / Par), SD(UnqPOS_noun|_adj / Sent)	SD(UnqPOS_noun|_adj|_adv / Par)
		SD(UnqPOS_pron / Sent)	M|Max(UnqPOS_noun / Sent), M|SD(UnqPOS_noun / Par)
		M(UnqPOS_verb / Par), SD(UnqPOS_verb / Sent)	M|Max(UnqPOS_verb / Sent), SD(UnqPOS_verb / Par)

Indices Category	Indices Subcategory	Significant Indices ($p < .05$)	Not Significant Indices ($p > .05$)
	Pron	M|Max(Pron_indef / Sent), M(Pron_indef / Par), SD(Pron_indef / Sent)	SD(Pron_indef / Par)
		-	M|Max|SD(Pron_fst|Pron_thrd / Sent), M|SD(Pron_fst|Pron_thrd / Par)
		SD(Pron_snd / Sent)	M|Max(Pron_snd / Sent), M(Pron_snd / Par), SD(Pron_snd / Par)
Syntax	Dep	M|Max(Dep_**X** / Sent), M(Dep_**X** / Par), SD(Dep_**X** / Sent)	SD(Dep_**X** / Par)
		X ∈ {nmod, amod, case, acl, obl, det, xcomp, nummod, conj, appos, mark, cc, objt}	
		M(Dep_nsubj / Par), SD(Dep_nsubj / Sent)	M|Max(Dep_nsubj / Sent), SD(Dep_nsubj / Par)
		* All other types of dependencies were not significant	
	ParseDepth	M|Max(ParseDepth / Sent), M(ParseDepth / Par), SD(ParseDepth / Sent)	SD(ParseDepth / Par)
Cohesion	AdjCoh	M|Max(AdjCoh / Par)	SD(AdjCoh / Par)
	IntraCoh	M|Max(IntraCoh / Par)	SD(IntraCoh / Par)
	* StartEndCoh, StartMidCoh, MidEndCoh, TransCoh – Not Relevant for this analysis		
Word	Chars	M|Max|SD(Chars / Sent|Word), M|SD(Chars / Par)	-
	LemmaDiff	Max|SD(LemmaDiff / Word)	Max|M|SD(LemmaDiff / Sent), M(LemmaDif / Word), M|SD(LemmaDiff / Par)
	Repetitions	M|Max|SD(Repetitions / Sent), M|SD(Repetitions / Par)	-
	NmdEnt	M|Max(NmdEnt_loc|_org / Sent|Word), SD(NmdEnt_loc|_org / Sent|Word), M(NmdEnt_loc|_org / Par),	SD(NmdEnt_loc|_org / Par), *All for NmdEnt_per
	Syllab	M|Max(Syllab / Sent|Word), M(Syllab / Par), SD(Syllab / Sent|Word)	SD(Syllab / Par)

* mean (abbreviated "M"), standard deviation (abbreviated "SD"), and maximum (abbreviated "Max") are the aggregation functions applied at various granularities.

Two methods were employed to determine the efficiency of textual indices from ReaderBench in differentiating texts from two language levels (i.e., A versus B): neural networks and statistical analyses. In the first approach, the ReaderBench features performed better than RuBERT embeddings (see Table 7). Nonetheless, the neural networks that used only the RuBERT embeddings as input performed well (i.e., accuracy of 88.69%), even though the BERT embeddings are recognized for their capabilities to model the meaning of a text. Note that this result does not imply that ReaderBench indices are better than BERT on text classification tasks in

general, but rather argue that ReaderBench textual complexity indices can be successfully employed to assess text difficulty.

Both inputs, ReaderBench textual complexity indices and RuBERT embeddings, were used in different versions of initial neural network. The results from Table 7 indicate that adding an extra hidden layer for the neural network with only textual complexity indices decreased performance, thus arguing that the function that maps the inputs to the predicted class should be a simple one. In contrast, the BERT embeddings benefitted from the additional layer, therefore arguing that the mapping between the encodings and the complexity of a text is more complex than in the previous case. In the third configuration the two input sources were combined and tested on the same task; this architecture achieved the highest score (92.36%) with two hidden layers, benefiting from both handcrafted features and BERT contextualized embeddings. The intuition behind the performance increase is that the two approaches complement each other.

The statistical analysis using Kruskal-Wallis statistical test showed that the majority of indices were significant in differentiating between the two classes. In general, the indices aggregated with the standard deviation function were not so statistically significant, while the mean and the maximum related indices proved to be more predictive. While considering Appendix 1, the "nmod" dependency category was the most influential one, ranking first in the Kruskal-Wallis $\chi2(1)$ score with the index *Max(Dep_nmod / Sent)* ($\chi2 = 84.48$, $p < .001$), as well as having 6 appearances in top 10 most influential features. The nominal modifier appeared more frequently in more complex texts (B) than in the less complex texts (A). In the same syntactic category, the "amod" dependency also exhibited similar patterns.

In terms of morphology, the number of nouns was higher in B texts than in A text, both as rough count and unique count. The mean value of nouns at sentence level was ranked 2nd in terms of effect size (M(POS_noun / Sent); $\chi2 = 84.31$, $p < .001$), while other 3 related indices made it to top 10 most predictive features. The number of adjectives was also statistically significant, with the most predictive index in this subcategory (i.e., M(POS_adj / Par); $\chi2 = 69.28$, $p < .001$) ranking in top 5% of all the indices.

From the Word category, character indices performed best in terms of separating the two types of texts (e.g., M(Chars/Word); $\chi2 = 76.03$, $p < .001$), all the three variations being close to each other in the ranking. This finding supports the intuition that easier texts generally have shorter words in their composition. Strongly related to this subcategory is the syllables subcategory that also had an important impact (e.g., M(Syllab / Word); $\chi2 = 73.08$, $p < .001$).

From the remaining two categories, Surface and Cohesion, the highest impact was obtained by the features regarding the number of unique words (e.g., M(UnqWd / Par); $\chi2 = 32.74$, $p < .001$) and, respectively, the middle end cohesion feature (e.g., M(MidEndCoh / Par); $\chi2 = 25.89$, $p < .001$). As it can be seen from Table 8, these features were still statistically significant in differentiating the two categories of texts, but they are in the middle of overall rankings in terms of predictive power (i.e., ranks between places 70 and 100).

10. Discussion

Our findings indicate that the ReaderBench textual complexity indices, which span across multiple levels of analysis, provide valuable insights into the differences between two language levels for foreign Russian learners (A-Basic User and B-Independent User). From a machine learning perspective, the results are interesting, as a simple neural network using the features extracted with ReaderBench outperformed the Russian version of BERT, namely RuBERT, in the task of text classification. Nonetheless, this result likely occurred given that the complexity indices were specifically fitted for this task. In addition, we observed that the combination of features from both methods improved the overall classification scores. As such, the methods complement one another and the texts from the two categories differ from each other in terms of both textual complexity features and underlying themes (represented by meaning).

A follow-up analysis was centered on the textual complexity features; as such, the Kruskal-Wallis test was used to identify the most predictive indices, individually and per category. From the syntactic point of view, we can observe that the two most impactful features were "nmod" and "amod". The nominal modifier (i.e., "nmod") consists of a noun or a noun phrase that is expressed in Russian using genitive, while showing the possessiveness of another noun; "amod" is similar, with the difference that the syntactical formation is an adjectival phrase. Thus, both "amod" and "nmod" modify the meaning of a noun. In other words, adding more information to the nouns seems to make texts more difficult to comprehend.

From the surface category, the most significant feature is the number of unique words. Although this feature is not that impactful, it suggests that B texts tend to be longer than A texts. Nonetheless, it is more interesting to emphasize the underlying reason: from the morphological category, the number of nouns and of adjectives influence most the differences between the two types of texts; thus, additional concepts (i.e., nouns) are introduced, with corresponding descriptions (i.e., adjectives). In contrast, the number of verbs indicative of actions has a lower $\chi 2$ value in comparison to the previously mentioned parts of speech. Thus, texts that are ranked as being more difficult include more descriptive passages rather than action centered.

When considering semantics, text cohesion does not differ that much in comparison to the other categories, although is statistically significant at in-between sentences from the input paragraph. Yet again, this was an expected result, given that text cohesion is a measure of how well ideas relate to one another and flow throughout the text; nevertheless, texts are well written by experts and should be cohesive.

Our statistical analysis pinpointed that the difficulty of Russian texts comes from the usage of more descriptive passages that include phrases rich in nouns and adjectives. Other characteristics, such as the number of (unique) words, are logical implications of the previous idea. Given that the considered corpus was developed

by language experts and can be considered of reference for the Russian educational system, our findings can further support the design of new materials for L2 education. In addition, ReaderBench can be used in other experiments or domains where textual complexity is an important factor, as it can be used to quantify the differences between B and C language level texts, between manuals from two different grade levels, or to estimate the difficulty of science, politics, or law texts.

11. Conclusions and future work

This paper introduced the adaptation of the open-source ReaderBench framework to support multilevel analyses of Russian language in terms of identifying text characteristics reflective of its difficulty. Numerous improvements were made, starting from code refactoring, the addition of new indices (e.g., adjacent cohesion for sentences and for paragraphs, inter-paragraph cohesion) and of the maximum aggregation function, the integration of BERT language model as input for building the CNA graph, as well as the usage of the MUSE version of word2vec that provides multilingual word embeddings.

The ReaderBench textual complexity indices together with BERT contextualized embeddings were used as inputs to predict the language level of texts from two classes: A (Basic User) and B (Independent User). Both approaches, namely neural network architectures and the statistical analyses using the Kruskal-Wallis test, confirmed that the complexity indices from ReaderBench are reliable predictors for text difficulty. The best performance of the neural network using both handcrafted features and BERT embeddings achieved a 92.36% leave one text out cross-validation, thus arguing for the model's capability to distinguish between text of various difficulties.

ReaderBench can be used to assess the complexity of Russian texts in different domains, including law, science, or politics. In addition, our framework can be employed by designers and developers of educational materials to evaluate and rank learning materials.

In terms of future work, we want to further extend the list of Russian textual complexity indices available in ReaderBench, including discourse markers and the Russian WordNet which currently is not aligned with the Open Multilingual Wordnet format. In addition, we envision performing additional studies regarding the complexity of the Russian texts and focusing on textbooks used in the Russian educational system, as well as multilingual analyses highlighting language specificities.

Acknowledgements

This research was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number TE 70 *PN-III-P1-1.1-TE-2019-2209,* ATES – "Automated Text Evaluation and Simplification".

REFERENCES

Abadi, Martin. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* Savannah, GA, USA: {USENIX} Association. 265–283.

Akhtiamov, Raouf B. 2019. Dictionary of abstract and concrete words of the Russian language: A methodology for creation and application. *Journal of Research in Applied Linguistics*. Saint Petersburg, Russia: Springer. 218–230.

Bansal, S. 2014. Textstat. Retrieved September 1st, 2021. URL: https://github.com/shivam5992/textstat (accessed 26.05.2022).

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(4–5). 993–1022.

BNC Consortium. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Boguslavsky, Igor, Leonid Iomdin & Victor Sizov. 2004. Multilinguality in ETAP-3: Reuse of lexical resources. In *Proceedings of the Workshop on Multilingual Linguistic Resources.* Geneva, Switzerland: COLING. 1–8.

Brysbaert, Marc, Boris New & Emmanuel Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods* 44(4). 991–997.

Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3). 904–911.

Choi, Joon Suh & Scott A. Crossley. 2020. ARTE: Automatic Readability Tool for English. NLP Tools for the Social Sciences. linguisticanalysistools.org. Retrieved September 1st, 2021. URL: https://www.linguisticanalysistools.org/arte.html (accessed 26.05.2022).

Churunina, Anna A., Ehl'zara Gizzatullina-Gafiyatova, Artem Zaikin & Marina I. Solnyshkina. 2020. Lexical Features of Text Complexity: The case of Russian academic texts. In *SHS Web of Conferences*. Nizhny Novgorod, Russia: EDP Sciences.

Coltheart, Max. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33(4). 497–505.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer & Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations*. Vancouver, BC, Canada: OpenReview.net.

Crossley, Scott A., Franklin Bradfield & Analynn Bustamante. 2019. Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research* 11(2). 251–270.

Crossley, Scott A., Kristopher Kyle, Jodi Davenport & Danielle S. McNamara. 2016. Automatic assessment of constructed response data in a Chemistry Tutor. In *International Conference on Educational Data Ining.* Raleigh, North Carolina, USA: International Educational Data Mining Society. 336–340.

Dale, Edgar & Jeanne S. Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin* 27(1). 37–54.

Dascalu, Mihai. 2014. *Analyzing Discourse and Text Complexity for Learning and Collaborating, Studies in Computational Intelligence* (534). Switzerland: Springer.

Dascalu, Mihai, Philippe Dessus, Stefan Trausan-Matu & Maryse Bianco. 2013. ReaderBench, an environment for analyzing text complexity and reading strategies. In H. Chad Lane, Kalina Yacef, Jack Mostow & Philip Pavlik (eds.), *16th Int. Conf. on Artificial Intelligence in Education (AIED 2013),* 379–388. Memphis, TN, USA: Springer.

Dascalu, Mihai, Danielle S. McNamara, Stefan Trausan-Matu & Laura K. Allen. 2018. Cohesion Network Analysis of CSCL Participation. *Behavior Research Methods* 50(2). 604–619. https://doi.org/10.3758/s13428-017-0888-4

Dascalu, Mihai, Lucia Larise Stavarache, Stefan Trausan-Matu & Philippe Dessus. 2014. Reflecting comprehension through French textual complexity factors. In *26th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2014).* 615–619. Limassol, Cyprus: IEEE.

Dascalu, Mihai, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu & Hub J. Kurvers. 2017. ReaderBench learns Dutch: Building a comprehensive automated essay scoring system for Dutch. In Anne E. Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, Benedict du Boulay, Ryan Baker (eds.), *18th Int. Conf. on Artificial Intelligence in Education (AIED 2017),* 52–63. Wuhan, China: Springer.

Davies, Mark. 2010. The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4). 447–464.

Delvin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Minneapolis, MN, USA: Association for Computational Linguistics. 4171–4186.

Flesch, Rudolf F. 1949. *Art of Readable Writing.*

Gabitov, Azat, Marina Solnyshkina, Liliya Shayakhmetova, Liliya Ilyasova & Saida Adobarova. 2017. Text complexity in Russian textbooks on social studies. *Revista Publicando* 4(13 (2)). 597–606.

Gifu, Daniela, Mihai Dascalu, Stefan Trausan-Matu & Laura K. Allen. 2016. Time evolution of writing styles in Romanian language. In *28th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2016).* San Jose, CA: IEEE. 1048–1054.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse & Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2). 193–202.

Guryanov, Igor, Iskander Yarmakeev, Aleksandr Kiselnikov & Iena Harkova. 2017. Text complexity: Periods of study in Russian linguistics. *Revista Publicando* 4(13 (2)). 616–625.

Gutu-Robu, Gabriel, Maria-Dorinela Sirbu, Ionut S Cristian Paraschiv, Mihai Dascălu, Philippe Dessus & Stefan Trausan-Matu. 2018. Liftoff – ReaderBench introduces new online functionalities. *Romanian Journal of Human – Computer Interaction* 11(1). 76–91.

Honnibal, Montani & I. Montani. 2017. Spacy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing* 7(1).

Hopkins, Kenneth D. & Douglas L. Weeks. 1990. Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement* 50(4). 717–729.

Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers & Brad S. Chissom. 1975. *Derivation of New Readability Formulas: (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Naval Air Station Memphis: Chief of Naval Technical Training.

Kozea. 2016. Pyphen. Retrieved September 1st, 2021. URL: https://pyphen.org/ (accessed 20.05.2022).

Kruskal, William H. & Allen W. Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47(260). 583–621.

Kuperman, Victor, Hans Stadthagen-Gonzalez & Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4). 978–990.

Kuratov, Yuri & Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv preprint arXiv:1905.07213*.

Kyle, Kristopher. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-based Indices of Syntactic Sophistication*.

Kyle, Kristopher, Scott A. Crossley & Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50(3). 1030–1046.

Kyle, Kristopher, Scott A. Crossley & Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly* 18(2). 154–170.

Kyle, Kristopher, Scott A. Crossley & Youjin J. Kim. 2015. Native language identification and writing proficiency. *International Journal of Learner Corpus Research* 1(2). 187–209.

Landauer, Thomas K., Peter W. Foltz & Darrell Laham. 1998. An introduction to Latent Semantic Analysis. *Discourse Processes* 25(2/3). 259–284.

LanguageTool. 2021. Language Tool. Retrieved September 1st, 2021. URL: https://languagetool.org/ (accessed 20.05.2022).

Loukachevitch, Natalia V., G. Lashevich, Anastasia A. Gerasimova, Vyacheslav V. Ivanov. Boris V. Dobrov. 2016. Creating Russian wordnet by conversion. In *Computational Linguistics and Intellectual Technologies: Annual conference Dialogue 2016.* Moscow, Russia. 405–415.

Mc Laughlin, G. H. 1969. SMOG grading-a new readability formula. *Journal of Reading* 12(8). 639–646.

Mccarthy, Kathryn, Danielle Siobhan, Marina I. Solnyshkina, Fanuza Kh. Tarasove & Roman V. Kupriyanov. 2019. The Russian language test: Towards assessing text comprehension. *Vestnik Volgogradskogo Gosudarstvennogo Universiteta. Seriya 2: Yazykoznanie* 18(4). 231–247.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representation in Vector Space. In *Workshop at ICLR*. Scottsdale, AZ.

Myint. 2014. language-check. Retrieved September 1st, 2021. URL: https://github.com/myint/language-check (accessed 23.05.2022).

Pearson, Karl. 1895. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58. 240–242.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12. 2825–2830.

Quispesaravia, Andre, Walter Perez, Marco Sobrevilla Cabezudo & Fernando Alva-Manchego. 2016. Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 4694–4698.

Rehurek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* Valletta, Malta: ELRA. 45–50.

Roscoe, Rod, Laura K. Allen, Jennifer L. Weston & Scott A. Crossley. 2014. The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition* 34. 39–59.

Sadoski, Mark, Ernest T. Goetz & Maximo Rodriguez. 2000. Engaging texts: Effects of concreteness on comprehensibility, interest, and recall in four text types. *Journal of Educational Psychology* 92(1). 85.

Sakhovskiy, Andrey, Valery D. Solovyev & Marina Solnyshkina. 2020. Topic modeling for assessment of text complexity in Russian textbooks. In *2020 Ivannikov Ispras Open Conference (ISPRAS)*. Moscow, Russia: IEEE. 102–108.

Schmid, Helmut, Marco Baroni, Erika Zanchetta & Achim Stein. 2007. Il sistema 'tree-tagger arricchito'–The enriched TreeTagger system. *IA Contributi Scientifici* 4(2). 22–23.

Senter, R.J. & E.A. Smith. 1967. Automated readability index: CINCINNATI UNIV OH.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423.

Shapiro, S.S. & M.B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52(3/4). 591–611.

Sharoff, Serge, Elena Umanskaya & James Wilson. 2014. *A Frequency Dictionary of Russian: Core Vocabulary for Learners*. Routledge.

Solnyshkina, Marina I., Valery Solovyev, Vladimir Ivanov & Andrey Danilov. 2018. Studying text complexity in Russian academic corpus with Multi-Level Annotation. *CEUR WORKSHOP PROCEEDINGS. Proceedings of Computational Models in Language and Speech Workshop, co-located with the 15th TEL International Conference on Computational and Cognitive Linguistics, TEL 2018.*

Solovyev, Valery, Marina Solnyshkina, Mariia Andreeva, Andrey Danilov & Radif Zamaletdinov. 2020. Text complexity and abstractness: Tools for the Russian language. In *International Conference "Internet and Modern Society" (IMS-2020).* St. Petersburg, Russia: CEUR Proceedings. 75–87.

Solovyev, Valery, Marina I. Solnyshkina & Vladimir Ivanov. 2018. Complexity of Russian academic texts as the function of syntactic parameters. In *19th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing*. Hanoi, Vietnam: Springer Lecture Notes in Computer Science.

Spearman, Carl. 1987. The proof and measurement of association between two things. *The American Journal of Psychology* 100(3/4). 441–471.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems.* Long Beach, CA, USA: Curran Associates, Inc. 5998–6008.

Vorontsov, Konstantin & Anna Potapenko. 2015. Additive regularization of topic models. *Machine Learning* 101(1) 303–323.

Appendix 1. **Statistically significant ReaderBench indices**

Index	A M (SD)	B M (SD)	χ2(1)	*p*
Max(Dep_nmod / Sent)	0.49 (0.84)	1.24 (1.33)	84.48	<.001
M(POS_noun / Sent)	1.50 (1.25)	2.58 (1.84)	84.31	<.001
M(Dep_nmod / Sent)	0.22 (0.41)	0.64 (0.79)	83.55	<.001
Max(POS_noun / Sent)	2.27 (2.12)	3.82 (2.63)	82.50	<.001
M(Dep_nmod / Par)	1.10 (3.57)	2.14 (2.67)	81.24	<.001
M(UnqPOS_noun / Sent)	1.50 (1.24)	2.51 (1.76)	79.97	<.001
Max(UnqPOS_noun / Sent)	2.26 (2.09)	3.73 (2.54)	78.43	<.001
Max(NgramEntr_2 / Word)	2.05 (0.34)	2.20 (0.45)	76.74	<.001
M(Chars / Word)	3.97 (0.98)	4.43 (1.12)	76.03	<.001
Max(Chars / Word)	9.21 (2.46)	10.76 (3.39)	74.83	<.001
M(POS_noun / Par)	6.96 (18.39)	8.81 (8.13)	73.83	<.001
M(Dep_amod / Par)	1.91 (6.70)	2.41 (2.72)	73.77	<.001
M(Syllab / Word)	1.73 (0.32)	1.89 (0.46)	73.08	<.001
Max(Syllab / Word)	3.66 (1.04)	4.27 (1.39)	72.10	<.001
M(UnqPOS_noun / Par)	5.80 (13.29)	7.96 (7.10)	69.45	<.001
M(POS_adj / Par)	2.65 (8.85)	3.28 (3.39)	69.28	<.001
M(UnqPOS_adj / Par)	2.42 (7.52)	3.18 (3.25)	69.05	<.001
Max(ParseDepth / Sent)	4.06 (1.61)	5.09 (2.06)	66.62	<.001

Index	A M (SD)	B M (SD)	χ2(1)	*p*
Max(Dep_amod / Sent)	0.70 (1.12)	1.29 (1.23)	66.28	<.001
M(Dep_amod / Sent)	0.35 (0.57)	0.73 (0.82)	65.66	<.001
SD(Dep_nmod / Sent)	0.18 (0.36)	0.47 (0.61)	63.21	<.001
Max(POS_adj / Sent)	0.97 (1.30)	1.68 (1.46)	62.27	<.001
Max(UnqPOS_adj / Sent)	0.97 (1.29)	1.66 (1.43)	62.25	<.001
SD(Syllab / Word)	0.88 (0.28)	1.01 (0.41)	62.20	<.001
M(NgramEntr_2 / Word)	0.88 (0.27)	0.98 (0.27)	62.15	<.001
M(POS_adj / Sent)	0.51 (0.67)	0.97 (0.98)	60.85	<.001
M(UnqPOS_adj / Sent)	0.51 (0.66)	0.96 (0.96)	60.81	<.001
SD(Chars / Word)	2.71 (0.71)	3.03 (0.94)	58.24	<.001
M(ParseDepth / Sent)	3.43 (1.00)	4.07 (1.47)	53.99	<.001
SD(Dep_amod / Sent)	0.23 (0.41)	0.46 (0.53)	47.14	<.001
M(Dep_case / Par)	3.03 (7.92)	3.45 (3.74)	39.03	<.001
SD(POS_noun / Sent)	0.61 (0.83)	1.07 (1.15)	38.47	<.001
SD(POS_adj / Sent)	0.34 (0.52)	0.58 (0.62)	38.30	<.001
SD(UnqPOS_adj / Sent)	0.34 (0.52)	0.58 (0.61)	37.76	<.001
SD(UnqPOS_noun / Sent)	0.61 (0.82)	1.04 (1.11)	36.88	<.001
SD(ParseDepth / Sent)	0.54 (0.69)	0.85 (0.81)	34.05	<.001
M(UnqWd / Par)	27.37 (48.36)	31.79 (24.87)	32.74	<.001
SD(NgramEntr_2 / Word)	0.78 (0.18)	0.82 (0.21)	31.88	<.001
Max(UnqWd / Sent)	11.72 (7.00)	14.93 (8.52)	31.81	<.001
M(WdEntr / Par)	2.58 (0.93)	2.84 (1.06)	31.78	<.001
M(Wd / Par)	39.37 (93.19)	41.26 (36.14)	31.65	<.001
Max(WdEntr / Sent)	2.23 (0.67)	2.4 (0.82)	31.30	<.001
Max(Wd / Sent)	12.76 (8.31)	16.56 (10.35)	29.49	<.001
Max(Dep_case / Sent)	1.17 (1.36)	1.69 (1.50)	29.39	<.001
SD(Dep_acl / Sent)	0.03 (0.11)	0.09 (0.20)	29.16	<.001
M(Pron_indef / Par)	1.34 (3.69)	1.65 (2.08)	28.83	<.001
M(Dep_case / Sent)	0.66 (0.78)	0.94 (0.85)	28.77	<.001
SD(Pron_indef / Sent)	0.24 (0.38)	0.4 (0.48)	28.44	<.001
M(Dep_obl / Par)	2.66 (7.21)	2.72 (3.06)	27.91	<.001
M(Dep_det / Par)	0.88 (2.34)	1.22 (1.72)	27.26	<.001
Max(Dep_det / Sent)	0.45 (0.74)	0.77 (1.01)	27.20	<.001
M(Dep_xcomp / Sent)	0.11 (0.26)	0.21 (0.35)	26.88	<.001
SD(Dep_case / Sent)	0.39 (0.54)	0.64 (0.70)	26.82	<.001
M(Dep_xcomp / Par)	0.60 (1.89)	0.76 (1.21)	26.80	<.001
Max(Dep_xcomp / Sent)	0.29 (0.55)	0.51 (0.74)	26.43	<.001
SD(Wd / Sent)	2.47 (3.25)	3.86 (3.92)	26.32	<.001
Max(Pron_indef / Sent)	0.61 (0.85)	0.92 (0.99)	26.19	<.001
M(MidEndCoh / Par)	0.23 (0.32)	0.35 (0.35)	25.89	<.001
Max(LemmaDiff / Word)	1.31 (0.87)	1.63 (0.99)	25.80	<.001
M(Dep_acl / Sent)	0.02 (0.12)	0.06 (0.14)	24.68	<.001
SD(Dep_det / Sent)	0.19 (0.36)	0.33 (0.45)	24.57	<.001
SD(UnqWd / Sent)	2.17 (2.77)	3.31 (3.28)	24.31	<.001
M(Dep_acl / Par)	0.17 (1.02)	0.26 (0.61)	24.23	<.001
Max(Dep_acl / Sent)	0.08 (0.30)	0.20 (0.43)	24.18	<.001
M(Dep_nummod / Sent)	0.04 (0.22)	0.10 (0.30)	23.29	<.001

Index	A M (SD)	B M (SD)	χ2(1)	*p*
M(Dep_det / Sent)	0.19 (0.33)	0.33 (0.62)	22.94	<.001
Max(Dep_nummod / Sent)	0.12 (0.41)	0.29 (0.64)	22.90	<.001
M(Dep_nummod / Par)	0.16 (0.62)	0.35 (0.86)	22.10	<.001
Max(Dep_obl / Sent)	1.02 (1.25)	1.38 (1.27)	21.88	<.001
M(UnqWd / Sent)	9.02 (4.39)	10.94 (6.11)	21.57	<.001
M(Pron_indef / Sent)	0.29 (0.43)	0.43 (0.56)	21.20	<.001
SD(Dep_xcomp / Sent)	0.12 (0.24)	0.23 (0.35)	21.19	<.001
SD(Dep_obl / Sent)	0.36 (0.51)	0.54 (0.60)	20.91	<.001
SD(Dep_nummod / Sent)	0.05 (0.18)	0.13 (0.30)	20.37	<.001
M(Sent / Par)	3.58 (7.30)	3.32 (2.59)	20.15	<.001
M(Wd / Sent)	9.57 (4.95)	11.84 (7.39)	19.55	<.001
SD(Repetitions / Sent)	0.34 (0.68)	0.53 (0.82)	18.27	<.001
M(Dep_conj / Par)	2.18 (5.95)	2.08 (2.48)	18.14	<.001
M(Dep_obl / Sent)	0.54 (0.68)	0.73 (0.71)	17.79	<.001
M(Commas / Par)	3.01 (7.55)	3.12 (3.43)	16.97	<.001
M(Dep_appos / Sent)	0.09 (0.27)	0.18 (0.41)	16.95	<.001
M(WdEntr / Sent)	1.99 (0.57)	2.09 (0.72)	16.82	<.001
SD(POS_adv / Sent)	0.37 (0.52)	0.51 (0.57)	16.19	<.001
M(StartMidCoh / Par)	0.23 (0.32)	0.32 (0.34)	16.09	<.001
SD(UnqPOS_adv / Sent)	0.36 (0.52)	0.50 (0.55)	15.72	<.001
M(Dep_obj / Par)	1.74 (4.35)	1.89 (2.43)	15.49	<.001
SD(Dep_cc / Sent)	0.27 (0.42)	0.39 (0.46)	15.46	<.001
Max(Dep_appos / Sent)	0.21 (0.51)	0.36 (0.66)	15.25	<.001
Max(Dep_conj / Sent)	0.94 (1.33)	1.23 (1.34)	15.03	<.001
SD(Dep_advmod / Sent)	0.40 (0.56)	0.57 (0.66)	14.70	<.001
M(Dep_appos / Par)	0.33 (1.07)	0.42 (0.86)	14.34	<.001
M(UnqPOS_adv / Par)	1.94 (3.90)	2.21 (2.50)	13.75	<.001
Max(UnqPOS_adv / Par)	1.94 (3.90)	2.21 (2.50)	13.75	<.001
SD(Pron_int / Sent)	0.15 (0.28)	0.24 (0.34)	13.58	<.001
M(Pron_int / Par)	0.57 (1.42)	0.80 (1.22)	13.47	<.001
M(POS_adv / Par)	2.19 (4.85)	2.33 (2.72)	13.44	<.001
M(Punct / Par)	8.34 (18.53)	7.74 (6.6)	13.39	<.001
M(Dep_mark / Par)	0.61 (1.54)	0.84 (1.41)	13.22	<.001
Max(Dep_obj / Sent)	0.75 (0.91)	0.99 (0.99)	12.82	<.001
SD(Dep_conj / Sent)	0.35 (0.56)	0.48 (0.61)	12.76	<.001
M(Dep_nsubj / Par)	4.18 (10.36)	3.77 (3.7)	12.74	<.001
SD(Commas / Sent)	0.45 (0.61)	0.60 (0.66)	12.74	<.001
Max(POS_adv / Sent)	1.01 (1.20)	1.29 (1.27)	12.66	<.001
SD(Dep_mark / Sent)	0.15 (0.29)	0.23 (0.34)	12.62	<.001
Max(Dep_mark / Sent)	0.35 (0.59)	0.53 (0.73)	12.61	<.001
SD(WdEntr / Sent)	0.23 (0.29)	0.30 (0.29)	12.49	<.001
Max(Commas / Sent)	1.32 (1.41)	1.68 (1.53)	12.41	<.001
Max(UnqPOS_adv / Sent)	1.00 (1.18)	1.27 (1.22)	12.32	<.001
Max(Pron_int / Sent)	0.36 (0.57)	0.55 (0.73)	12.32	<.001
SD(Dep_obj / Sent)	0.29 (0.41)	0.40 (0.45)	12.25	<.001
M(Dep_cc / Par)	1.73 (4.53)	1.71 (2.16)	12.14	<.001
M(Repetitions / Par)	1.85 (5.64)	1.96 (3.06)	11.87	<.001

Index	A M (SD)	B M (SD)	χ2(1)	*p*
M(SentAdjCoh / Par)	0.35 (0.33)	0.43 (0.33)	11.81	<.001
M(Dep_mark / Sent)	0.15 (0.31)	0.22 (0.39)	11.66	<.001
M(Dep_advmod / Par)	2.65 (5.95)	2.72 (3.19)	11.36	<.001
M(StartEndCoh / Par)	0.35 (0.34)	0.42 (0.35)	11.32	<.001
M(POS_verb / Par)	5.65 (13.16)	5.16 (5.15)	11.19	<.001
M(UnqPOS_verb / Par)	5.19 (11.26)	4.92 (4.83)	10.90	<.001
SD(NmdEnt_loc / Sent)	0.09 (0.30)	0.15 (0.35)	10.82	.001
SD(POS_verb / Sent)	0.54 (0.67)	0.70 (0.72)	10.74	.001
SD(Punct / Sent)	0.66 (0.84)	0.87 (0.97)	10.32	.001
Max(Repetitions / Sent)	0.98 (1.78)	1.37 (2.13)	10.11	.001
SD(UnqPOS_verb / Sent)	0.54 (0.67)	0.68 (0.71)	9.91	.002
M(Pron_int / Sent)	0.15 (0.28)	0.22 (0.39)	9.69	.002
Max(Dep_advmod / Sent)	1.18 (1.33)	1.48 (1.44)	9.67	.002
Max(Dep_cc / Sent)	0.73 (0.94)	0.93 (0.99)	9.59	.002
M(Dep_fixed / Sent)	0.03 (0.13)	0.08 (0.24)	9.51	.002
SD(LemmaDiff / Word)	0.39 (0.22)	0.42 (0.23)	9.35	.002
M(NmdEnt_org / Sent)	0.01 (0.12)	0.06 (0.29)	9.05	.003
Max(NmdEnt_org / Sent)	0.06 (0.51)	0.12 (0.5)	8.91	.003
M(NmdEnt_org / Par)	0.10 (0.98)	0.14 (0.66)	8.87	.003
SD(Dep_expl / Sent)	0.00 (0.05)	0.02 (0.1)	8.62	.003
M(NmdEnt_loc / Sent)	0.08 (0.29)	0.13 (0.34)	8.60	.003
M(Dep_conj / Sent)	0.47 (0.64)	0.62 (0.86)	8.50	.004
M(NmdEnt_loc / Par)	0.40 (2.53)	0.51 (1.26)	8.42	.004
SD(Dep_fixed / Sent)	0.05 (0.18)	0.12 (0.32)	8.24	.004
M(Dep_obj / Sent)	0.39 (0.49)	0.50 (0.56)	8.17	.004
Max(Dep_expl / Sent)	0.01 (0.10)	0.04 (0.2)	8.16	.004
M(Dep_expl / Sent)	0.00 (0.06)	0.01 (0.08)	8.14	.004
M(Dep_expl / Par)	0.01 (0.13)	0.05 (0.22)	8.13	.004
Max(Dep_fixed / Sent)	0.15 (0.47)	0.25 (0.58)	7.93	.005
SD(Dep_appos / Sent)	0.07 (0.21)	0.13 (0.30)	7.76	.005
Max(NmdEnt_loc / Sent)	0.23 (0.72)	0.33 (0.73)	7.76	.005
M(Dep_fixed / Par)	0.19 (0.63)	0.27 (0.65)	7.67	.006
SD(Dep_iobj / Sent)	0.11 (0.25)	0.15 (0.26)	7.34	.007
SD(POS_pron / Sent)	0.44 (0.58)	0.56 (0.67)	7.02	.008
SD(Dep_advcl / Sent)	0.09 (0.21)	0.13 (0.24)	6.67	.010
SD(Dep_nsubj / Sent)	0.36 (0.47)	0.45 (0.52)	6.62	.010
M(Commas / Sent)	0.74 (0.80)	0.93 (1.01)	6.50	.011
SD(UnqPOS_pron / Sent)	0.42 (0.56)	0.52 (0.60)	6.11	.013
M(Repetitions / Sent)	0.43 (0.73)	0.64 (1.52)	5.74	.017
M(POS_adv / Sent)	0.55 (0.74)	0.65 (0.74)	5.72	.017
SD(Pron_snd / Sent)	0.06 (0.20)	0.11 (0.27)	5.51	.019
M(UnqPOS_adv / Sent)	0.55 (0.73)	0.65 (0.73)	5.49	.019
SD(NmdEnt_org / Sent)	0.01 (0.12)	0.05 (0.22)	5.39	.020
SD(Dep_csubj / Sent)	0.04 (0.15)	0.07 (0.18)	5.07	.024
SD(Dep_ccomp / Sent)	0.07 (0.19)	0.11 (0.23)	5.01	.025
Max(Dep_csubj / Sent)	0.10 (0.32)	0.16 (0.39)	4.85	.028
M(Dep_csubj / Sent)	0.04 (0.14)	0.05 (0.18)	4.80	.029

Index	A M (SD)	B M (SD)	χ2(1)	*p*
M(Dep_ccomp / Par)	0.27 (0.84)	0.36 (0.80)	4.55	.033
Max(Dep_ccomp / Sent)	0.19 (0.43)	0.26 (0.50)	4.35	.037
M(Dep_csubj / Par)	0.16 (0.59)	0.18 (0.46)	4.34	.037
Max(Dep_iobj / Sent)	0.28 (0.54)	0.35 (0.56)	4.28	.039
M(Dep_ccomp / Sent)	0.08 (0.23)	0.10 (0.23)	4.24	.039
M(Dep_iobj / Par)	0.50 (1.43)	0.47 (0.91)	3.95	.047
M(Dep_cc / Sent)	0.38 (0.50)	0.45 (0.55)	3.89	.049

Article history:
Received: 20 October 2021
Accepted: 06 February 2022

Bionotes:

Dragos CORLATESCU is a Teaching Assistant and a PhD student at the University Politehnica of Bucharest researching the field of Natural Language Processing from various perspectives. He has explored the areas of text analysis and classification, assessment of online communities, and chatbot development.
Contact information:
University Politehnica of Bucharest
Splaiul Independentei 313, Bucharest, 060042, Romania
e-mail: dragos.corlatescu@upb.ro
ORCID: 0000-0002-7994-9950

Stefan RUSETI is a Lecturer at the University Politehnica of Bucharest (UPB) with a PhD in Natural Language Processing. He has over 25 publications in the field, 6 of them in top ranked conferences, and an extensive experience in projects using NLP techniques and AI (deep neural networks) architectures.
Contact information:
University Politehnica of Bucharest
Splaiul Independentei 313, Bucharest, 060042, Romania
e-mail: stefan.ruseti@upb.ro
ORCID: 0000-0002-0380-6814

Mihai DASCALU is Full Professor at the University Politehnica of Bucharest (UPB) with a strong background in Computer Science applied in Education. He has extensive experience in national and international research projects with more than 200 published papers.
Contact information:
University Politehnica of Bucharest
Splaiul Independentei 313, Bucharest, 060042, Romania
e-mail: mihai.dascalu@upb.ro
ORCID: 0000-0002-4815-9227

Сведения об авторах:
Драгош КОРЛАТЕСКУ – ассистент и аспирант Политехнического университета Бухареста, работает в области обработки естественного языка, имеет опыт анализа и классификаци текстов, оценки онлайн-сообществ и разработки чат-ботов.
Контактная информация:
University Politehnica of Bucharest
Splaiul Independentei 313, Bucharest 060042, Romania
e-mail: dragos.corlatescu@upb.ro
ORCID: 0000-0002-7994-9950

Штефан РУСЕТИ – преподаватель Политехнического университета Бухареста (UPB) имеет степень доктора философии в области обработки естественного языка. Автор более 25 публикаций в этой области, 6 из них – на конференциях с высоким рейтингом, имеет опыт работы в проектах с использованием методов НЛП и архитектуры искусственного интеллекта (глубоких нейронных сетей).
Контактная информация:
University Politehnica of Bucharest
Splaiul Independentei 313, Bucharest 060042, Romania
e-mail: stefan.ruseti@upb.ro
ORCID: 0000-0002-0380-6814

Михай ДАСКАЛУ – профессор Политехнического университета Бухареста (UPB) с большим опытом работы в области компьютерных наук, применяемых в образовании. Участвовал во многих национальных и международных исследовательских проектах, автор более 200 опубликованных работ.
Контактная информация:
University Politehnica of Bucharest
Splaiul Independentei 313, Bucharest 060042, Romania
e-mail: mihai.dascalu@upb.ro
ORCID: 0000-0002-4815-9227

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 371–390
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30178

Research article

What neural networks know about linguistic complexity

Serge SHAROFF

University of Leeds, Leeds, Great Britain
✉s.sharoff@leeds.ac.uk

Abstract
Linguistic complexity is a complex phenomenon, as it manifests itself on different levels (complexity of texts to sentences to words to subword units), through different features (genres to syntax to semantics), and also via different tasks (language learning, translation training, specific needs of other kinds of audiences). Finally, the results of complexity analysis will differ for different languages, because of their typological properties, the cultural traditions associated with specific genres in these languages or just because of the properties of individual datasets used for analysis. This paper investigates these aspects of linguistic complexity through using artificial neural networks for predicting complexity and explaining the predictions. Neural networks optimise millions of parameters to produce empirically efficient prediction models while operating as a black box without determining which linguistic factors lead to a specific prediction. This paper shows how to link neural predictions of text difficulty to detectable properties of linguistic data, for example, to the frequency of conjunctions, discourse particles or subordinate clauses. The specific study concerns neural difficulty prediction models which have been trained to differentiate easier and more complex texts in different genres in English and Russian and have been probed for the linguistic properties which correlate with predictions. The study shows how the rate of nouns and the related complexity of noun phrases affect difficulty via statistical estimates of what the neural model predicts as easy and difficult texts. The study also analysed the interplay between difficulty and genres, as linguistic features often specialise for genres rather than for inherent difficulty, so that some associations between the features and difficulty are caused by differences in the relevant genres.
Keywords: *automatic text classification, deep learning, interpreting neural networks*

For citation:
Sharoff, Serge. 2022. What neural networks know about linguistic complexity. *Russian Journal of Linguistics* 26 (2). 371–390. https://doi.org/10.22363/2687-0088-30178

Научная статья

Что нейронные сети знают о лингвистической сложности

С.А. ШАРОВ

Университет Лидса, Лидс, Великобритания
✉s.sharoff@leeds.ac.uk

Аннотация
Лингвистическая сложность – это комплексное явление, поскольку оно проявляется на разных уровнях (от сложности текстов до предложений, от слов до подсловных единиц), через

разные особенности (от жанров до синтаксиса и семантики), а также через разные задачи (изучение языка, перевод, обучение, специфические потребности различных аудиторий). Наконец, результаты анализа сложности будут отличаться для разных языков из-за их типологических свойств, культурных традиций, связанных с конкретными жанрами в этих языках, или просто из-за свойств отдельных наборов данных, используемых для анализа. В данной статье эти аспекты лингвистической сложности исследуются с помощью искусственных нейронных сетей для прогнозирования сложности и объяснения данных прогнозов. Нейронные сети оптимизируют миллионы параметров для создания эмпирически эффективных моделей прогнозирования, работая как черный ящик, т.е. не определяя, какие лингвистические факторы приводят к конкретному решению. В статье показано, как связать нейронные прогнозы сложности текста с обнаруживаемыми свойствами лингвистических данных, например, с частотой союзов, дискурсивных частиц или придаточных предложений. Конкретное исследование касается нейронных моделей прогнозирования сложности, которые были обучены различать более простые и сложные тексты в разных жанрах на английском и русском языках, а также были исследованы на предмет лингвистических свойств, которые коррелируют с прогнозами. Представленное исследование показывает, что количество существительных и связанная с этим сложность именных групп влияют на сложность текста. Данная закономерность подтверждена статистически, а нейронная модель предсказывает сложность текста. В исследовании также проанализирована взаимосвязь сложности текста и жанра, поскольку лингвистические особенности часто связаны с жанром, а не с непосредственной сложностью текста, в связи с чем некоторые параметры взаимосвязи между функциями и сложностью детерминированы различиями в соответствующих жанрах.
Ключевые слова: *автоматическая классификация текста, глубокое обучение, интерпретация нейронных сетей*

Для цитирования:
Sharoff S. What neural networks know about linguistic complexity. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 371–390. https://doi.org/10.22363/2687-0088-30178

1. Introduction

Linguistic complexity is a complex phenomenon, as it manifests itself on different levels, through different features, and via different application tasks. In terms of levels of complexity analysis, it is natural to analyse complexity on the level of words, as some of them are naturally more difficult than others, which allows for a way of ranking them as is often done in Complex Word Identification (CWI) tasks. A different set of categories is needed to analyse complexity of sentences, which primarily depends on the networks of syntactic and semantic relations between words. Yet another level of complexity analysis concerns difficulty with respect to global text properties, which is primarily about capturing the flow of argumentation: even when individual sentences are easy to understand, the links between them might require a greater cognitive load.

Another aspect of complexity analysis concerns the features we use in our description of complexity. For words we can refer to their frequencies or their semantic features, such as abstractness, whereas morphosyntactic features are connected with the part-of-speech categories or the dependency relations. For text-level analysis we can use rhetorical relations as well as a typology of genres. In any case, each level of analysis (words, sentences or texts) is described computationally by a vector of such features with a fixed number of dimensions.

There is also a multitude of reasons why we are interested in the phenomenon of complexity. This determines what is considered to be simple or complex in each case. A typical example of applications of complexity analysis concerns language learning, which presupposes the existence of an audience of non-native speakers acquiring a foreign language either as children or adults. In this kind of application, we can tune our analysis for specific language teaching tasks, as some phenomena are less likely to cause problems in understanding, but more problems in production, or we can refer to a target audience, as different phenomena are likely to cause problems depending on the the learners' native language. Another example of applications concerns translation training, which is different from language learning, as the challenge for a trainee translator often consists in transferring various aspects of the source texts into their native language. A related case concerns analysis of complexity in the context of language acquisition for children learning their native language. Yet another example concerns specific needs of other kinds of audiences, such as production of texts for native speakers with various mental disabilities.

Finally, the results of complexity analysis will differ for different languages, because of their typological properties (such as greater complexity of syntactic relations between words vs greater morphological complexity of word forms); or the cultural traditions associated with specific genres in these languages, for example, emphasis on plain language in research papers in English vs traditionally accepted forms of academic discourse in Russian. It is also important to understand the properties of individual datasets used for analysis, as occasional confounding variables for the dataset, such as a limited range of genres or authors, might affect the replicability of the findings.

This paper investigates some of these aspects by focusing on word- and sentence-level analysis while also investigating the impact of genres. In terms of the task, the focus is on studying difficulties for adult learners for two languages, English and Russian, without a specification of their native language and with a specific focus on the language understanding task.

In terms of the computational methodology, the study uses artificial neural networks for predicting complexity. It deals with neural difficulty prediction models which have been trained to differentiate between easier and more complex texts in different genres in English and Russian. While neural networks produce empirically efficient prediction models by optimising millions of parameters, they operate as a black box without determining which linguistic factors lead to a specific prediction. Following the Bertology framework (Rogers et al. 2020), this paper shows how to link neural predictions of text difficulty to detectable properties of linguistic data, for example, to the frequency of conjunctions, discourse particles or subordinate clauses. More specifically, the linguistic features are primarily based on Douglas Biber's Multidimensional Analysis (Biber 1995), such as the rate of *that* deletion or public verbs, to explain predictions of fine-tuned transformer models, such as XLM-Roberta (Conneau et al. 2019).

2. Methodology

The study presented in this paper focuses on the fine-grained difficulty assessment, when difficulty analysis is transformed from the text level to the sentence level. The focus of this study is on the prediction of complexity with respect to teaching foreign languages, more specifically vide licet, automatic assessment of reading exercises from language learning textbooks. What varies in this study is a set of properties, namely the influence of genres, syntax and lexical semantics on the predictions.

2.1. Classification methods

From the computational viewpoint, the complexity prediction problem can be defined as a short-text classification task, which assigns a complexity label for a short text or a segment. Since difficulty naturally operates on a scale (some texts are considered as more difficult than others), this problem can be also defined as a regression task, which predicts a numeric difficulty value for a text. The study focuses on the classification task, because many statistical operations need categorical labels and because the original annotated corpora use a small fixed number of levels. While there is a range of methods for the short-text classification task, recent studies favoured fine-tuning pre-trained transformer models. The pre-training of neural networks aims at establishing their weights by the task of predicting missing words on large corpora, for example, Wikipedias in the case of BERT (Devlin et al. 2018) or Common Crawl in the case of XLM-Roberta transformer model (Conneau et al. 2019). In the end, the pre-trained representations can be shown to reflect general linguistic phenomena, such as agreement or semantic classes (Rogers et al. 2020). Fine-tuning on a target task (difficulty prediction in this case) adapts the weights of the pre-trained representations, so that the general phenomena can be linked to the target task.

In addition to building the difficulty prediction classifiers, other text parameters can be tested. More specifically, this study applied existing neural classifiers for genres to both training and testing corpora using a well-tested automatic genre annotation model (Sharoff 2021). This allows us to compare properties of texts of the same difficulty but in different genres, as well as texts in the same genre, but of different difficulty levels.

2.2. Human interpretation of neural predictions

Neural networks produce empirically efficient prediction models, especially the modern setup which is based on fine-tuning pre-trained transformer models, such as BERT. However, they act as a blackbox, as it is difficult to determine why a model with a given set of training parameters produced a specific prediction. Therefore, the NLP field recently has started developing a range of approaches under the name of Bertology to understand reasons for predictions (Rogers et al. 2020).

Bertology analysis of prediction difficulty developed in this study extends the framework from (Sharoff 2021), which uses Logistic Regression (LR) to detect the

linguistic features associated with (more accurate) predictions of a neural model. LR is a fast and transparent Machine Learning method, which is defined as:

$$ln\frac{p}{1-p} = w_0 + w_1x_1 + \ldots + w_nx_n$$

It fits a linear model to predict the log-odds ratio, where p is the probability of a text having a particular label, for example, Easy or Difficult, $x_1, \ldots, x_n$ are interpretable variables, e.g., the proportion of verbs or conjunctions. Since the model is linear, the relative contribution of each feature can be determined through its weight for detecting this function. To assist in comparing the weights, the variables have been standardised with respect to their values and dispersion prior to fitting the logistic regression, so that for each feature its mean is zero and its standard deviation is one. In the end, the feature weights can be directly compared. Another advantage of logistic regression over other machine learning methods is that it has been well investigated from the statistical viewpoint, thus allowing a number of tests to determine the significance of each feature. One of the approaches for testing the feature significance is based on the likelihood ratio test, which compares the likelihood of the data under the full model against the likelihood of the data under a model with one of the features removed (Hosmer Jr. et al. 2013). If the behaviour of the logistic regression model changes significantly when a feature is removed, the feature can be considered as more significant for this label. The lists below show the weights of features selected under the likelihood ratio test.

The linguistic features used in this study are based on the set introduced by Douglas Biber for describing register variation via Multi-Dimensional Analysis (Biber 1988). The features include the following categories:

Lexical features such as:

- public verbs = *acknowledge, admit, agree, assert, claim, complain, declare, deny…*
- time adverbials = *afterwards, again, earlier, early, eventually, formerly, immediately,…*
- amplifiers = *absolutely, altogether, completely, enormously, entirely,…*

Part-of-speech (POS) features such as:

- nominalisations
- prepositions
- past tense verbs.

Syntactic features such as:

- *be* as the main verb
- *that* deletions
- pied piping.

Text-level features such as:

- average word length
- average sentence length
- type/token ratio (TTR).

This set was designed specifically for English. However, some of its features are nearly universal, which could be exemplified with text-level features, even though their exact values are language-dependent. Many lexical features are comparable across languages if they can be translated reliably, public verbs is a good illustration. Many part-of-speech features can be used across a number of languages as well, particularly nominalisations, while many syntactic features are comparable only across a smaller set of closely related languages, for example, pied piping. Some functionally equivalent features are included into the list for Russian even when they are expressed in a different way in Russian. For instance, F18 (BYpassives according to (Biber 1988)) is expressed via passives with the agent in the instrumental case, but for consistency this feature still keeps the same name as in English. Similarly, detecting C12 (*do* as pro-verb in English) is based in Russian on detecting ellipsis in conditions similar to those used for detecting C12 in English. See the list in Appendix 1 for the full description of the features. Even though the set of features was introduced to describe register variation, it is sufficiently general to provide explanations for the difficulty levels.

Table 1. **CEFR-annotated datasets for English and Russian**

Level	English		Russian	
	Texts	**Segments**	**Texts**	**Segments**
A1	**0**	**0**	**178**	**1149**
A2/KET	**64**	**304**	**121**	**1707**
B1/PET	**60**	**516**	**134**	**2109**
B2/FCE	**71**	**1354**	**167**	**4022**
C1/CAE	**67**	**1606**	**120**	**1937**
C2/CPE	**69**	**1540**	**6**	**121**

2.3. Datasets

The training datasets came from the Cambridge Readability Dataset (Xia et al. 2016) for English and from the Rufola corpus (Laposhina et al. 2018) for Russian. In both cases, the source texts have been taken from existing textbooks marked with the CEFR levels by the developers of the respective corpora. Namely, the Cambridge Proficiency Tests have been mapped to the CEFR levels for English, while the levels of several textbooks have been unified into the CEFR scheme for Russian. In both cases, the corpora are annotated by the CEFR levels on the text level, which means that a text corresponds to a single reading exercise. Since the amount of data on the text level does not provide enough training samples for building reliable classifiers, each text in the respective datasets was split into smaller segments with the aim of training within a window of several sentences. The optimal window size was determined to be of three sentences (this window was expanded if the total length of three adjacent sentences was less than 15 words). The distribution of training data on the document level vs the chosen window level is given in Table 1.

Large-scale testing of the linguistic properties has been conducted with raw text corpora from the English and Russian portions of the Aranea family (Benko

2016), which were obtained by Web crawling and post-processing of websites in the respective languages. These corpora offer a reliable snapshot of how English and Russian are used in Web pages. In addition, the Nauka-Plus portion of the Taiga corpus (Shavrina & Shapovalova 2017) was used for testing in Russian, since it has been also annotated with difficulty levels, though the focus of its annotation was on assessing its difficulty for the native speakers of Russian. The reason for using Nauka-Plus in this study is to compare the automatic difficulty predictions aimed at the non-native speakers with the verified difficulty estimates for the native speakers.

Table 2. **Accuracy of XLM-Roberta for English and Russian**

		English			Russian	
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
A1				**0.72**	**0.75**	**0.74**
A2	**0.75**	**0.84**	**0.79**	**0.51**	**0.64**	**0.57**
B1	**0.58**	**0.66**	**0.62**	**0.50**	**0.66**	**0.57**
B2	**0.53**	**0.74**	**0.62**	**0.71**	**0.59**	**0.65**
C1	**0.54**	**0.53**	**0.53**	**0.58**	**0.47**	**0.52**
C2	**0.77**	**0.49**	**0.59**	**0.00**	**0.00**	**0.00**
macro avg	**0.70**	**0.62**	**0.63**	**0.50**	**0.52**	**0.51**
accuracy	**0.60**			**0.60**		
Binary case						
Easy	**0.89**	**0.98**	**0.93**	**0.90**	**0.98**	**0.94**
Difficult	**0.99**	**0.97**	**0.98**	**0.92**	**0.65**	**0.76**
macro avg	**0.94**	**0.97**	**0.96**	**0.91**	**0.82**	**0.85**
accuracy	**0.97**			**0.91**		

Table 3. **Confusion matrices**

	A2	**B1**	**B2**	**C1**	**C2**
A2	**256**	**37**	**10**	**0**	**1**
B1	**40**	**343**	**118**	**12**	**3**
B2	**18**	**129**	**1001**	**175**	**31**
C1	**4**	**60**	**505**	**845**	**192**
C2	**6**	**18**	**238**	**531**	**747**

The classifiers for difficulty were built by fine-tuning the XLM-Roberta transformer model (Conneau et al. 2019) from the HuggingFace library (Wolf et al. 2019) using the CUP and Rufola training sets respectively for English and Russian. Another set of classifiers for probing the neural predictions was built on the basis of the Multi-Dimensional Analysis features and the Logistic Regression model (see Section 2.2 below). Table 2 lists the cross-validation accuracy scores after fine-tuning on the respective training corpora. The overall accuracy of both models is 60%, but the Russian model is trailing behind with respect to the F1 score. Since C2 is a minority class for Russian (see Table 1), this class is not detected in cross-validation (its texts are all classified as C1), thus bringing the macro-average F1 score down. Overall, more difficult texts (C1 and C2) are not very common in the Russian training set, which makes the task of their detection more challenging in

comparison to English. Nevertheless, in the binary scenario of distinguishing between Easy (A1, A2, B1) and Difficult (C1 and C2) texts the accuracy reaches 91% for Russian and 97% for English, which is sufficient for our purposes.

3. Results

To simplify the presentation of the results, the study provides the contrast of Easy vs Difficult texts, i.e., those predicted at the lowest three levels (A1, A2 and B1) vs those at the top two level (C1 and C2) with the B2 level reserved as a boundary, since the errors of the classifiers overlap over this boundary. The reason for extending the scale of Easy texts to B1 comes from the lack of data for Web pages detected as suitable for A1 and A2 levels (the total number of such pages is less than 1% for either language), so what is presented as Easy in the analysis below comes mostly from pages classified as suitable for the B1 level.

Table 4. **Association of features with difficulty for English**

DIFFICULT		**EASY**	
A01.pastVerbs	**0.299**	**C07.2persProns**	**0.341**
J43.TTR	**0.229**	**K45.conjuncts**	**0.271**
P67.analNegn	**0.205**	**I39.preposn**	**0.206**
E14.nominalizations	**0.133**	**B04.placeAdverbials**	**0.160**
C06.1persProns	**-0.116**	**L54.predicModals**	**0.134**
G19.beAsMain	**-0.120**	**G19.beAsMain**	**0.120**
L54.predicModals	**-0.134**	**C06.1persProns**	**0.116**
B04.placeAdverbials	**-0.160**	**E14.nominalizations**	**-0.133**
I39.preposn	**-0.206**	**P67.analNegn**	**-0.205**
K45.conjuncts	**-0.271**	**J43.TTR**	**-0.228**
C07.2persProns	**-0.341**	**A01.pastVerbs**	**-0.300**

Table 5. **Association of features with difficulty for Russian**

DIFFICULT		**EASY**	
A03.presVerbs	**0.294**	**C07.2persProns**	**0.340**
I42.ADV	**0.292**	**J44.wordLength**	**0.332**
E14.nominalizations	**0.289**	**D13.whQuestions**	**0.024**
I39.preposn	**0.208**	**C08.3persProns**	**-0.077**
P67.analNegn	**0.207**	**C09.impersProns**	**-0.078**
H37.conditional	**0.098**	**H37.conditional**	**-0.132**
H38.otherSubord	**0.094**	**I39.preposn**	**-0.216**
B05.timeAdverbials	**0.094**	**A01.pastVerbs**	**-0.239**
C09.impersProns	**0.086**	**I42.ADV**	**-0.341**
C06.1persProns	**-0.205**	**P67.analNegn**	**-0.381**
C07.2persProns	**-0.242**	**A03.presVerbs**	**-0.390**

Tables 4 and 5 list associations of the positive and negative weights of the most significant features with respect to the predicted difficulty levels. Some features work in the same way in both languages. For example, the rate of the first and second person pronouns has the strongest positive association with easy texts and the strongest negative association with difficult texts. These pronouns indicate

personal interaction, which is often expressed in interactive spoken-like texts, even though the classifiers were applied to written language in HTML Web pages. The rate of first and second person pronouns is likely to be higher in discourse about areas of "immediate relevance" as expected for the A-level CEFR texts (Council of Europe 2001). Similarly, the greater rate of nominalisations and negations is consistently associated with difficult text across both languages. This quantitative evidence supports other linguistic studies concerning the extra complexity involved in processing negations in comparison to positive sentences (Doughty & Long 2008). Similarly, nominalisations and complex noun phrases have been linked to the conceptual difficulty of grammatical metaphors when actions, which are congruently expressed by verbs, get packed into noun phrases, for example, from *how glass cracks* into *the glass crack growth rate* (Halliday 1992).

Some difficulty indicators are language-specific. They can be often linked to prominent language-specific constructions. In particular, G19.beAsMain is associated with easy texts for English, as this construction offers a simple formulaic expression for relational predicates (*X is Y*), while other relational predicates, for example, *X involves Y*, are more likely to be found in more advanced writing. The same feature does not appear prominently in easier Russian texts, as the Russian equivalent of *to be* is not overtly expressed in the present tense and therefore it is not counted by the feature extraction mechanism.

It is interesting to note that the feature I39.preposn is associated with different directions of complexity in English and Russian. For English its greater rate indicates easier texts, while for Russian this is associated with more difficult ones. This can be explained by the typological differences between the two languages: what is expressed by the basic prepositions in English (*of, to* or *with*) is often rendered by the case endings in Russian (respectively, genitive, dative or instrumental). Therefore, a more active use of the prepositions in Russian correlates with more complex writing styles, when sentences need to include more information than the basic Subject-Verb-Object skeleton which introduces the main participants. At the same, more accessible writing styles in English need to use prepositions at a high rate, while this rate is reduced in more complex styles because of the more active use of other features, such as negations or noun compounds.

The adverbials as a syntactic function appear in Tables 4 and 5 in three different forms: as adverbs, which are detected as a POS category, and as either time adverbials or place adverbials, which are detected via lexical lists, for example, *behind* or *South*. Therefore, the rates of adverbials of different kinds affect difficulty in different ways. General adverbs tend to occur as modifiers of adjectives and verbs, thus leading to more elaborated constructions associated with more complex styles. However, time and place adverbials often occur in narratives, hence they are less likely to be associated with complex styles.

Some features do not offer an easy cross-lingual explanation, such as the greater rate of conjuncts in easier English texts or the greater rate of conditionals in more difficult Russian texts. Also, quite surprisingly, word length has a positive correlation with easier Web pages in Russian and has not been detected as a significant factor associated with difficulty in English.

Table 6. **Association of features with difficulty for Nauka-Plus**

DIFFICULT		EASY	
C10.demonstrProns	**0.542**	**N60.thatDeletion**	**0.461**
C08.3persProns	**0.406**	**J43.TTR**	**0.431**
I40.attrAdj	**0.375**	**I39.preposn**	**0.184**
E14.nominalizations	**0.343**	**B05.timeAdverbials**	**0.162**
I42.ADV	**0.298**	**D13.whQuestions**	**-0.010**
A03.presVerbs	**0.247**	**H38.otherSubord**	**-0.041**
C12.doAsProVerb	**-0.137**	**A03.presVerbs**	**-0.113**
P67.analNegn	**-0.154**	**K48.amplifiers**	**-0.120**
K45.conjuncts	**-0.178**	**E14.nominalizations**	**-0.300**
I39.preposn	**-0.185**	**C08.3persProns**	**-0.341**
B05.timeAdverbials	**-0.381**	**I40.attrAdj**	**-0.348**
J43.TTR	**-0.397**	**C10.demonstrProns**	**-0.392**

There is an apparent problem in interpreting the results of the Type-Token Ratio (TTR) score as reported in Table 6 for Nauka-Plus texts against the results reported in Table 4. The TTR rate (J43) in Table 4 is in line with previous studies, such as (Collins-Thompson & Callan 2004), when the higher TTR is associated with greater lexical diversity and hence with more difficult texts. At the same time, Table 6 for Nauka-Plus associates TTR with easier texts. It seems that the answer to this discrepancy comes from differences in the corpus composition in terms of topics, genres or other text properties. In this specific case, news reporting is the most common genre category in the Nauka Plus dataset (57%) with the second most common category being academic writing (30%), Table 9. As features vary across genres, the TTR is often considerably higher in news reporting as it often includes many personal names and locations, thus increasing their TTR without necessarily increasing their perceived difficulty. This can be illustrated by variation of the TTR across the genre categories in this dataset. For example, the Inter-Quartile Range (IQR) of TTR on the Nauka-Plus corpus is 0.5727 to 0.6727, with texts with the top quartile of the TTR values (i.e., above 0.6727) contain a higher proportion of news reporting (72%) vs academic writing (19%) in comparison to the entire corpus (57% vs 40%). Even relatively infrequent named entities do not necessarily contribute to the greater difficulty of their texts, for example, *Британское подразделение американской компании Локхид Мартин провело испытания модернизированной боевой машины пехоты Warrior* ('The British office of Lockheed Martin tested a upgraded version of their armoured carrier Warrior'). Another indicator of easy texts for Nauka Plus happens to be the higher rate of prepositions and time adverbials, which are also more typical for news reporting. This is another indication of the importance of genres to determining the difficulty features, as the preposition rate (I39) is also contrary to the observations from the general Web pages in Russian, which associate the higher rate of prepositions with more difficult texts.

Nauka Plus texts are closer to academic writing contain explications, which are treated as more difficult according to the annotators. From the viewpoint of the linguistic features, they contain more verbs in the present tense and more attributive

adjectives, while they tend to repeat relevant terms, thus leading to lower TTR, for example, *Burkholderia одновременно является патогенным паразитическим микроорганизмом, изменяющим геном амеб*… (‘At the same time Burkholderia is a pathogenic parasitic microorganism, which alters the amoeba genome…’) with words *Burkholderia, amoeba, genome, microorganism, pathogenic* repeated throughout the article.

Table 7. **Association of difficulty with communicative functions for English**

Difficult	#Texts	Functions	Easy	#Texts	Functions
23.15%	**945958**	**A12.promotion**	**35.93%**	**195245**	**A12.promotion**
17.50%	**715187**	**A16.information**	**17.85%**	**97005**	**A7.instruction**
16.97%	**693702**	**A1.argumentation**	**15.80%**	**85831**	**A8.newswire**
12.08%	**493616**	**A8.newswire**	**9.44%**	**51302**	**A16.information**
9.40%	**384344**	**A7.instruction**	**7.37%**	**40024**	**A11.personal**
6.56%	**268242**	**A11.personal**	**7.16%**	**38898**	**A1.argumentation**
5.10%	**208218**	**A17.reviewing**	**4.30%**	**23372**	**A17.reviewing**
4.26%	**174118**	**A14.academic**	**1.88%**	**10193**	**A9.legal**
3.88%	**158695**	**A9.legal**	**0.21%**	**1136**	**A4.fiction**
1.09%	**44571**	**A4.fiction**	**0.06%**	**349**	**A14.academic**

Table 8. **Association of difficulty with communicative functions for Russian**

Difficult	#Texts	Functions	Easy	#Texts	Functions
19.12%	**212072**	**A1.argumentation**	**29.28%**	**251923**	**A12.promotion**
15.37%	**170401**	**A7.instruction**	**19.68%**	**169320**	**A8.newswire**
15.34%	**170121**	**A12.promotion**	**12.35%**	**106272**	**A16.information**
14.64%	**162356**	**A8.newswire**	**11.77%**	**101265**	**A7.instruction**
13.26%	**147047**	**A16.information**	**9.08%**	**78111**	**A1.argumentation**
7.79%	**86435**	**A11.personal**	**6.07%**	**52224**	**A11.personal**
6.01%	**66696**	**A17.reviewing**	**5.36%**	**46098**	**A17.reviewing**
4.07%	**45123**	**A14.academic**	**3.92%**	**33734**	**A9.legal**
3.18%	**35264**	**A9.legal**	**1.91%**	**16460**	**A14.academic**
1.21%	**13396**	**A4.fiction**	**0.56%**	**4843**	**A4.fiction**

The close link between difficulty and genres observed in the Nauka-Plus corpus calls for experiments comparing predictions for these categories. Tables 7 and 8 present the association between genres (expressed in terms of generic communicative functions) and difficulty levels in the Aranea corpora for English and Russian. The tables highlight the cases when the proportion of genres predicted as Difficult or Easy is **higher** than for the opposite case. For example, the proportion of texts with the predicted function of A7.instruction is higher for Easy texts in English (17.85% vs 9.4% for Difficult texts in Table 7). Overall, the classifiers predict a greater proportion of promotional, news reporting, instructional and personal reporting texts as Easy across both languages. This matches the intuition of the language teachers who tend to include such texts in exercises. The Fiction category is an exception to this intuition as it is often treated as a prime example of texts useful for language learners with many exercises based on examples from novels. At the same time, this study finds that typical authentic examples of fiction (at least as found on the Web) are predicted as less suitable for the learners.

Table 9. **Distribution of genres in Nauka-Plus**

4463	**A8.newswire**
2295	**A14.academic**
319	**A12.promotion**
29	**A12.promotion/A8.newswire**
20	**A8.newswire/A14.academic**
16	**A1.argumentation**
16	**A8.newswire/A12.promotion**
13	**A14.academic/A18.newswire**
9	**A7.instruction**

Table 10. **Human annotations for difficulty Nauka-Plus vs predicted CEFR levels**

NP1:	**Human**	**CEFR**
1325	**L4**	**C1**
972	**L1**	**B1**
899	**L3**	**C1**
871	**L2**	**B1**
837	**L2**	**C1**

Despite the different aims of the human annotation of difficulty available in the Nauka-Plus corpus (aimed at the native Russian speakers) and the automatic difficulty predictions in terms of CEFR levels, the difficulty levels are well aligned (see Table 10). The most difficult texts according to the human annotation in Nauka-Plus receive the highest CEFR level predictions and vice versa, while the automatic classifier avoids making C2 and A-level predictions.

Table 11. **Positive and negative features for *easy* instructional and news texts**

A7.instructional		**A8.news**	
C07.2persProns	**0.5155**	**K55.publicVerbs**	**0.2913**
C06.1persProns	**0.1791**	**H35.causative**	**0.2666**
B04.placeAdverbials	**0.1702**	**H38.otherSubord**	**0.2214**
I39.preposn	**0.1603**	**N59.contractions**	**0.2192**
L54.predicModals	**0.1371**	**K47.generalHedges**	**0.2129**
N60.thatDeletion	**0.1341**	**D13.whQuestions**	**0.1841**
B05.timeAdverbials	**0.1028**	**A01.pastVerbs**	**0.1756**
L53.necessModals	**0.0638**	**C09.impersProns**	**0.1525**
H35.causative	**-0.0784**	**C08.3persProns**	**0.0521**
K56.privateVerbs	**-0.0902**	**F18.BYpassives**	**-0.1857**
H25.presPartClaus	**-0.0984**	**K48.amplifiers**	**-0.1864**
E14.nominalizations	**-0.1146**	**K50.discoursePart**	**-0.2290**
I42.ADV	**-0.1366**	**L54.predicModals**	**-0.2427**
C09.impersProns	**-0.1612**	**E16.Nouns**	**-0.2705**
A03.presVerbs	**-0.1678**	**K45.conjuncts**	**-0.3521**
E16.Nouns	**-0.2482**	**C07.2persProns**	**-0.4385**

A7.instruction and A8.news are among the communicative functions which are common in both Easy and Difficult parts of Aranea. Table 11 lists the linguistic features which are specific to easy texts **within** these genres. Some features resemble what is characteristic for Easy texts in English in general, such as the use of the first and second personal pronouns, as well as the prepositions and time and

place adverbials for instructions. As expected, the use of nouns, nominalisations, adverbs as modifiers, as well as more complex syntactic constructions in the form of subordinate clauses of different kinds, is associated with more difficult texts. At the same time, a novel feature specific to this genre concerns the use of modal verbs, either necessity or prediction modals, which can be associated with more complex writing styles in general, but in the case of instructions, the use of modals makes them clearer.

The two examples below illustrate instructional texts which are classified as respectively easy and difficult:

> **EASY** The Executive Hire Show takes place at The Ricoh Arena , Coventry . </p> Bus Public transport from train station to the Ricoh Arena : – Number 8 bus from Coventry Train Station to Coventry Transport Museum – Then catch the number 4 or number 5 from Coventry Transport Museum to Arena Park (Tesco) – Once you arrive at Arena Park there is an underpass which takes you into Car Park B of the Ricoh Arena . Follow signs for the Ricoh Arena main entrance from here . </p> Taxi For our local taxi service please visit www.mgmtaxi.co.uk or call 02476 375550 </p> Train Please note – The last train leaving Coventry Railway Station to London Euston is 23 : 31 …[1]
> **DIFFICULT** Introduction </p> The most important part of working with this particular linked dataset , and probably datasets in general , is understanding what the variables mean and how they are coded . This is aided by studying the codebook, where available, and by running frequency tables of categorical and ordinal variables and means / medians of continuous variables . The codebook describes (or should describe the name of each variable, what it is supposed to measure, and the number of levels or range of the values the variable takes on in the dataset. This will tell you, for example, if sex is coded as M and F, or 0 and 1, or 1 and 2, or 1, 2 and 9, etc. The codebook for the linked Census data tells you that the income variables actually refer to 1985 income, even though the Census was taken in June of 1986. It is important to keep this in mind when analyzing the data . </p> One-way or two-way frequency tables not only give information on how the variables are distributed , but also … [2]

Examples also show that the neural transformer model is able to detect the inherent difficulty of topics, for example, descriptions of a statistical procedure (Difficult) as compared to giving directions (Easy), because the latter topic is more expected in texts for learners of lower levels. However, this inherent difficulty is not reflected in the set of the Biber features, and therefore is not captured in probing experiments as reported in Tables 4 or 11.

As for distinguishing easy and difficult texts among the news reporting texts, TTR is not in this list, thus implying that this feature has less impact on the difficulty level within news items. The strongest indicator of difficult texts in this genre is K45.conjuncts, such as *in particular, instead, otherwise, similarly*, which are linked

[1] http://www.executivehireshow.co.uk/visiting/travel

[2] http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1244

to more complex reporting styles, also with fewer past tense verbs. The counter-intuitive link between the difficult news articles and the second person pronouns rate (which featured prominently for easy texts in Table 4) is related to incomplete cleaning of some of the Web pages, as the most frequent contexts for *you* in this collection are legalistic boilerplate privacy notes, such as *When you subscribe we will use the information you provide to send you these newsletters..*, which are not considered as simple by the classifier.

While the rate of nouns was not considered as a predictive feature for the full corpus, as it varies considerably across the genres, this was detected as a significant feature within the two genres in Table 11.

4. Related studies

Statistical methods for analysing text complexity can be traced to frequency studies aimed at designing systems of shorthand writing (Käding 1897), which was followed by traditional measures of readability, such as Lorge or Flesch-Kincaid measures, initially developed in the context of American adult education (Lorge 1944, DuBay 2004). There has also been a long line of research in statistical frequency distribution models, which can be linked to complexity (Juilland 1964, Orlov 1983, Baayen 2008).

With the rise of Machine Learning, novel methods for readability prediction appeared, initially based on extraction of features (Pitler & Nenkova 2008, Collins-Thompson 2014, Vajjala & Meurers 2014), such as those introduced by Biber, or on various frequency measures. In particular, it has been shown that unsupervised Principal Component Analysis arrives at the two principal dimensions with groups of features resembling lexical difficulty, for example, frequencies or word length, and syntactic difficulty, such POS codes (Sharoff et al. 2008). Other studies have also experimented with expanding the models from the document to the sentence level (Vajjala & Meurers 2014) with a specific aim of comparing sentences from the Simple English Wikipedia against aligned sentences from the standard English Wikipedia.

As in many other areas of computational linguistics, feature-less neural networks provided better efficiency in difficulty predictions (Nadeem & Ostendorf 2018), especially with the rise of pre-trained transformer models (Khallaf & Sharoff 2021), which outperform both the linguistic features and the traditional neural networks.

Other studies have also emphasised the influence of genres on the predictions of the classifiers. In particular, existing approaches for measuring text complexity tend to **overestimate** the complexity levels of informational texts while simultaneously **underestimating** the complexity levels of literary texts (Sheehan et al. 2013). The authors of that study had to design different difficulty models for each of the two kinds of texts.

This study uses the CUP and Rufola datasets for training the classifiers. There are also many other sources for building models to distinguish easy or difficult

texts. For English a commonly used choice is the WeeBit corpus (Vajjala & Meurers 2012), which consists of texts from the Weekly reader magazine and from the BBC Bite-Size website. The other source is the Core Standards for secondary education in the US context[3]. In all of these datasets, the aim of difficulty annotation assumes the audience of native learners aged 7—17. A related experiment investigated syntactic parameters for predicting difficulty of Russian academic texts (Solovyev et al. 2019). There are also various sources of texts with difficulty assessed for adult speakers, for example, the WikiHow corpus (Debnath & Roth 2021), which is based on Wiki texts edited for vagueness in instructions. Yet another source comes from other training scenarios, for example, from translation training, when texts are assessed with respect to the quality of their rendering by translation students. For instance, for translation into Russian (Kunilovskaya & Lapshinova-Koltunski 2019) or Chinese (Yuan & Sharoff 2020) the drop in quality or time spent on translation can be an indicator of difficulty.

5. Conclusions and further research

This paper presents a statistical study conducted on a large corpus to determine which features contribute to difficulty of English and Russian texts. This is based on a framework which combines a transformer-based neural prediction model operating as a blackbox and well-studied linguistic features providing a statistical explanation of how these features affect difficulty. For example, this study shows how the rate of nouns and the related complexity of noun phrases affects difficulty via statistical estimates of what the neural model predicts as easy and difficult texts (cf. Corlatescu et al., this issue).

The study also analysed the interplay between difficulty and genres, as linguistic features often specialise for genres rather than for inherent difficulty, so that some associations between the features and difficulty are caused by differences in the relevant genres. In particular, the Type-Token Ratio (TTR) is a good indicator of lexical diversity and it is usually higher with more difficult texts if both texts are in the same genre. At the same time, the study shows that the TTR of easy news reporting texts is likely to be higher than that of more difficult argumentative texts which make repeated references to the same key concepts.

From the practical viewpoint, the methods of this study help in automatic assessment of texts from the Web with the aim of extending the use of authentic texts in language teaching. The methods also help us to understand what makes authentic texts difficult and what might require their manual or automatic simplification. For example, despite the popularity of Fiction in language teaching applications, the study provides statistical evidence for the higher difficulty scores associated with fiction commonly found on the Web. This should not prevent tutors from using fiction for language teaching, as it can be beneficial for both engagement and pedagogic purposes, but this calls for more attention to choosing and simplifying such texts when necessary.

[3] http://www.corestandards.org/assets/Appendix_B.pdf

Further extensions planned for improving the neural difficulty detection models involve several lines of research. First, this study focused almost exclusively on reading exercises for language learners. We need more experiments on studying variations in the link between difficulty and linguistic features with respect to different difficulty assessment needs or the composition of the training datasets. Even within the area of studying language teaching and expressing difficulty via the CEFR levels, different datasets might have different approaches to what constitutes a B1 text, for example. Some texts are also included into a textbook for a specific level not because they fully correspond to a specific level, but because they can be used in other exercises for this level. For example, an authentic interview included into a B1 textbook might contain rare words or more complex grammatical constructions beyond expectations of typical B1 students, while it can be a good basis for a number of exercises for understanding how native speakers express their opinions. From the viewpoint of Machine Learning, an interview of this kind, even if legitimately included in the textbook, acts as noise for training neural prediction models. We need to experiment with various statistical tests to establish how annotation noise can lead to less reliable predictions and how to improve our prediction models (for example, see Paun et al. 2018).

Second, there is a rise in research on causal models (for example, Fytas et al. 2021), because when we have a classifier, it is important to know whether this decision has been made for the right reasons, rather than because of mere correlations in our training data. Recent causal interaction methods can explain some of the issues with interpretation of predictions reported above (Janizek et al. 2021).

Third, a related line of research involves assessment of the process of mapping CEFR levels of documents to the level of segments. The process of segmentation used in this study can lead to noise, because some 3-sentence segments coming from a textbook of a higher level can still be suitable for students on lower levels. This has already been noticed in the context of using simplified Wikipedia (Vajjala & Meurers 2014). A similar task exists in other areas, for example, turning models which predict the quality of sentence-level translations to models predicting word quality (Zhai et al. 2020).

Finally, we need to pay more attention to cognitive aspects of difficulty processing beyond simple scores, such as exemplified by the CEFR levels. For example, this involves adding an explicit model for processing named entities (NEs), such as people's names or locations. Anecdotal experience shows that language learners can often handle NEs, even if they are very rare, either because they are similar to how they are expressed in their native languages (see the example with *Lockheed Martin* above) or because they can understand the function of a personal name or a location even without knowing this particular entity. This needs to be quantified. NEs are also important in a different way, as neural models can be brittle to NE replacements. For example, replacing NEs in the co-reference task changes 85% of predictions (Balasubramanian et al. 2020).

REFERENCES

Baayen, Harald. 2008. *Analyzing Linguistic Data*. Cambridge University Press, Cambridge.

Balasubramanian, Sriram, Naman Jain, Gaurav Jindal, Abhijeet Awasthi & Sunita Sarawagi. 2020. What's in a name? Are BERT named entity representations just as good for any other name? *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Online. 205–214.

Benko, Vladimír. 2016. Two years of Aranea: Increasing counts and tuning the pipeline. *Proc LREC*. Portorož, Slovenia.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Collins-Thompson, Kevyn. 2014. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics* 165(2). 97–135.

Collins-Thompson, Kevyn & Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. *Proc. of HLT/NAACL*. Boston. 193–200.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Debnath, Alok & Michael Roth. 2021. A computational analysis of vagueness in revisions of instructional texts. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Online. 30–35.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doughty, Catherine, J. Michael & H. Long. 2008. *The Handbook of Second Language Acquisition* 27. John Wiley & Sons.

DuBay, William H. 2004. *The Principles of Readability*. Technical report, Impact Information.

Fytas, Panagiotis, Georgios Rizos & Lucia Specia. 2021. What makes a scientific paper be accepted for publication? *Proceedings of the First Workshop on Causal Inference and NLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic. 44–60.

Halliday, M.A.K. 1992. Language as system and language as instance: The corpus as a theoretical construct. In J. Svartvik (ed.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82 Stockholm* 65, 61–77. Walter de Gruyter.

Hosmer Jr, David W., Stanley Lemeshow & Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. John Wiley & Sons.

Janizek, Joseph D., Pascal Sturmfels & Su-In Lee. 2021. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research* 22(104). 1–54.

Juilland, Alphonse. 1964. *Frequency Dictionary of Spanish Words*. Mouton.

Käding, Friedrich Wilhelm (ed.). 1897. *Häufigkeitswörterbuch der Deutschen Sprache*. Selbstverlag.

Khallaf, Nouran & Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Kyiv, Ukraine (Virtual). 105–114.

Kunilovskaya, Maria & Ekaterina Lapshinova-Koltunski. 2019. Translationese features as indicators of quality in English-Russian human translation. *Proceedings of the*

Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019). Incoma Ltd., Shoumen, Bulgaria, Varna, Bulgaria. 47–56.

Laposhina, Antonina N., Tatyana Veselovskaya, Maria Lebedeva & Olga Kupreshchenko. 2018. Automated text readability assessment for Russian second language learners. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*.

Lorge, Irving. 1944. Predicting readability. *Teachers College Record*.

Nadeem, Farah & Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, Louisiana. 45–55.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR).* Technical report, Council of Europe, Strasbourg.

Orlov, Jurij. 1983. Ein modell der häufigkeitsstruktur des vokabulars. In H. Guiter & M. Arapov (eds.), *Studies on Zipf's law*, 154–233.

Paun, Silviu, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz & Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics* 6. 571–585.

Pitler, Emily & Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. *Proc EMNLP*. 186–195.

Rogers, Anna, Olga Kovaleva & Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* 8. 842–866.

Sharoff, Serge. 2021. Genre annotation for the web: Text-external and text-internal perspectives. *Register Studies* 3. 1–32.

Sharoff, Serge, Svitlana Kurella, & Anthony Hartley. 2008. Seeking needles in the Web haystack: Finding texts suitable for language learners. *Proc Teaching and Language Corpora Conference, TaLC 2008*. Lisbon.

Shavrina, Tatiana & Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: Taiga syntax tree corpus and parser. *CORPORA, International Conference*. Saint-Petersburg.

Sheehan, Kathleen M., Michael Flor & Diane Napolitano. 2013. A two-stage approach for generating unbiased estimates of text complexity. *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*. Association for Computational Linguistics, Atlanta, Georgia. 49–58.

Solovyev, Valery, Marina Solnyshkina, Vladimir Ivanov & Ildar Batyrshin. 2019. Prediction of reading difficulty in Russian academic texts. *Journal of Intelligent & Fuzzy System* 36(5). 4553–4563.

Straka, Milan & Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proc CoNLL 2017 Shared Task*. Association for Computational Linguistics, Vancouver, Canada. 88–99.

Vajjala, Sowmya & Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, Montréal, Canada. 163–173.

Vajjala, Sowmya & Detmar Meurers. 2014. 'Readability assessment for text simplification: From analysing documents to identifying sentential simplifications'. *ITL-International Journal of Applied Linguistics* 165(2). 194–222.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xia, Menglin, Ekaterina Kochmar & Ted Briscoe. 2016. Text readability assessment for second language learners. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, San Diego, CA. 12–22.

Yuan, Yu & Serge Sharoff. 2020. Sentence level human translation quality estimation with attention-based neural networks. *Proc LREC*, Marseilles.

Zhai, Yuming, Gabriel Illouz & Anne Vilnat. 2020. Detecting non-literal translations by fine-tuning cross-lingual pre-trained language models. *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online). 5944–5956.

Appendix 1. Linguistic features

The order of the linguistic features and their codes are taken from (Biber 1988). The conditions for detecting the features for English replicate the published procedures from (Biber 1988), many of them are expressed via lists of lexical items or via POS annotations, which in this study are provided by UDPIPE (Straka & Straková 2017). The Russian features are either based on translating the English word lists or on using identical or functionally similar constructions.

Code	Label	Condition
A01	past verbs	VERB, Tense=Past
A03	present verbs	VERB, Tense=Pres
B04	place adverbials	ADV, lex in (*aboard,above,abroad,across…*)
B05	time adverbials	ADV, lex in (*afterwards,again,earlier…*)
C06	first person pronouns	PRON, lex in (*I,we,me,us,my...*)
C07	second person pronouns	PRON, lex in (*you,your,yourself,yourselves*)
C08	third person pronouns	PRON, lex in (*she,he,they,her,him,them,his…*)
C09	impersonal pronouns	Conditions from (Biber 1988)
C10	demonstrative pronouns	Conditions from (Biber 1988)
C11	indefinite pronouns	PRON, lex in (*anybody,anyone,anything,everybody…*)
C12	*do* as pro-verb	Conditions from (Biber 1988)
D13	wh-questions	Conditions from (Biber 1988)
E14	nominalizations	lex ends with ('tion','ment','ness', 'ism')
E16	nouns	Conditions from (Biber 1988)
F18	passives with *by*	Conditions from (Biber 1988)
G19	*be* as main verb	Conditions from (Biber 1988)
H23	wh-clauses	Conditions from (Biber 1988)
H34	sentence relatives	Conditions from (Biber 1988)
H35	causatives	CONJ, lex in (*because*)
H36	concessives	CONJ, lex in (*although,though,tho*)
H37	conditionals	CONJ, lex in (*if, unless*)
H38	other subordination	Conditions from (Biber 1988)
I39	prepositions	ADP

Code	Label	Condition
I40	attributive adjectives	Conditions from (Biber 1988)
I41	predicative adjectives	Conditions from (Biber 1988)
I42	adverbs	ADV
J43	type-token ratio	Using 400 words as in (Biber 1988)
J44	word length	Average length of orthographic words
K45	conjuncts	Conditions from (Biber 1988)
K46	downtoners	lex in (*almost,barely,hardly,merely...*)
K47	general hedges	lex in (*maybe, at about, something like...*)
K48	amplifiers	lex in (*absolutely,altogether,completely,enormously…*)
K49	general emphatics	Conditions from (Biber 1988)
K50	discourse particles	Conditions from (Biber 1988)
K55	public verbs	VERB, lex in (*acknowledge,admit,agree…*)
K56	private verbs	VERB, lex in (*anticipate,assume,believe…*)
K57	suasive verbs	VERB, lex in (*agree,arrange,ask…*)
K58	seem/appear	VERB, lex in (*appear, seem*)
L52	possibility modals	VERB, lex in (*can,may,might,could*)
L53	necessity modals	VERB, lex in (*ought,should,must*)
L54	prediction modals	VERB, lex in (*shall,will,would*), excluding future tense
N59	contractions	Conditions from (Biber 1988)
N60	that deletion	Conditions from (Biber 1988)
P66	synthetic negation	Conditions from (Biber 1988)
P67	analytic negation	Conditions from (Biber 1988)

Article history:
Received: 20 October 2021
Accepted: 08 February 2022

Bionote:
Serge SHAROFF is a Researcher at the Centre for Translation Studies, University of Leeds, UK. His research interests include language technology, machine translation, corpus linguistics, genres and text classification.

Contact information:
Centre for Translation Studies, University of Leeds, Leeds, UK
e-mail: s.sharoff@leeds.ac.uk
ORCID: 0000-0002-4877-0210

Сведения об авторе:
Сергей Александрович ШАРОВ – научный сотрудник Центра переводоведения университета Лидса (Великобритания). В сферу его научных интересов входят языковые технологии, машинный перевод, классификация текстов и жанров.

Контактная информация:
Центр переводоведения, Университет Лидса, Лидс, Великобритания
e-mail: s.sharoff@leeds.ac.uk
ORCID: 0000-0002-4877-0210

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 391–408
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30122

Research article

A cognitive linguistic approach to analysis and correction of orthographic errors

Robert REYNOLDS[1,2] ✉, Laura JANDA[1] and Tore NESSET[1]

[1]*UiT The Arctic University of Norway, Tromsø, Norway*
[2]*Brigham Young University, Provo, Utah, USA*
✉robert_reynolds@byu.edu

Abstract
In this paper, we apply usage-based linguistic analysis to systematize the inventory of orthographic errors observed in the writing of non-native users of Russian. The data comes from a longitudinal corpus (560K tokens) of non-native academic writing. Traditional spellcheckers mark errors and suggest corrections, but do not attempt to model *why* errors are made. Our approach makes it possible to recognize not only the errors themselves, but also the conceptual causes of these errors, which lie in misunderstandings of Russian phonotactics and morphophonology and the way they are represented by orthographic conventions. With this linguistically-based system in place, we can propose targeted grammar explanations that improve users' command of Russian morphophonology rather than merely correcting errors. Based on errors attested in the non-native academic writing corpus, we introduce a taxonomy of errors, organized by pedagogical domains. Then, on the basis of this taxonomy, we create a set of mal-rules to expand an existing finite-state analyzer of Russian. The resulting morphological analyzer tags wordforms that fit our taxonomy with specific error tags. For each error tag, we also develop an accompanying grammar explanation to help users understand why and how to correct the diagnosed errors. Using our augmented analyzer, we build a webapp to allow users to type or paste a text and receive detailed feedback and correction on common Russian morphophonological and orthographic errors.
Keywords: *morphophonology, phonotactics, orthography, corpus, error taxonomy, webapp*

For citation:
Reynolds, Robert, Laura Janda &Tore Nesset. 2022. A cognitive linguistic approach to analysis and correction of orthographic errors. *Russian Journal of Linguistics* 26 (2). 391–408. https://doi.org/10.22363/2687-0088-30122

Научная статья

Лингвокогнитивный подход к классификации и исправлению орфографических ошибок

Роберт РЕЙНОЛЬДС[1,2] ✉, Лора ЯНДА[1], Торе НЕССЕТ[1]

[1] *Университет Тромсё — Арктический университет Норвегии, Тромсё, Норвегия*
[2] *Университет Бригама Янга, Прово, Юта, США*
✉robert_reynolds@byu.edu

Аннотация
В представленной статье мы предлагаем систематизацию орфографических ошибок неносителей русского языка на основе лингвистических и когнитивных критериев. Материалом исследования послужили данные лонгитюдного корпуса (560000 слов) работ на русском языке, написанных студентами-иностранцами. Традиционные автоматические средства проверки орфографии (spell checkers) выявляют ошибки и предлагают исправления, но не могут построить объяснительные когнитивные модели. Предлагаемый подход позволяет распознать не только сами ошибки, но и концептуальные причины этих ошибок, заключающиеся в непонимании фонотактики и морфофонологии русского языка, а также в способах их репрезентации орфографическими правилами. Этот способ позволяет обосновывать причины грамматических ошибок и рекомендовать правила, которые улучшают владение пользователями русской морфофонологией, а не просто исправляют ошибки. Принцип систематизации аннотированных ошибок в корпусе академического письма на неродном языке и таксономия ошибок ориентированы на преподавание. На основе представленной таксономии мы разработали набор правил (mal-rules), расширяющих функционал конечно-автоматного анализатора русского языка. Разработанный морфологический анализатор аннотирует словоформы специальными тегами ошибок. Для каждого тега ошибки мы предлагаем сопровождающее пояснение, чтобы помочь пользователям понять, почему и как исправить диагностированные ошибки. Используя наш расширенный анализатор, мы создаем веб-приложение, позволяющее пользователям набирать или вставлять текст, а также подробные комментарии и исправления распространенных морфофонологических и орфографических ошибок в русском языке.

Ключевые слова: *морфофонология, фонотактика, орфография, корпус, таксономия ошибок*

Для цитирования:
Reynolds R., Janda L., Nesset T. A cognitive linguistic approach to analysis and correction of orthographic errors. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 391–408. https://doi.org/10.22363/2687-0088-30122

1. Introduction

Traditional approaches to spell checking are sometimes inadequate for the needs of non-native users because they are optimized for native speakers. Not only is it assumed that the user is capable of choosing between suggested corrections, but the suggestions themselves are optimized for the kinds of errors that *native* speakers make. Even if a non-native user were able to select the correct form from the suggested corrections, it is entirely possible that the user would not understand *why* it is the correct form in contrast to the form they wrote. Furthermore, whereas spell checking for native speakers is mainly a matter of fixing one-off random

errors, non-native users need to acquire rules that they can apply in the future. The mistakes that non-native writers make tend to be systematic, and thereby can be analyzed linguistically and present excellent targeted learning opportunities.

The output of a spellchecker will frequently be either too broad (merely marking a word as misspelled) or too specific (suggesting an alternative for a single given misspelled word) to support the acquisition of useful generalizations. Our proposed tool, the Russian Mentor for Orthographic Rules (RuMOR) is designed to help non-native users connect each specific error to linguistic generalizations, orthographic rules, and examples. This design encourages the user to update their understanding of Russian linguistic and orthographic patterns so that they can avoid making similar errors in the future.

Section 2 reviews related research in the fields of morphological analysis, spelling correction, and intelligent tutoring systems. In Section 3, we describe our methodology, including the process of classifying errors in the RULEC (ENA, April 16, 2022)[1] corpus, modeling the errors in a finite-state framework using mal-rules, evaluating the model, and applying the model in a webapp for users. Section 4 contains a summary of our results and future research directions.

2. Related work

Our project is connected to research in a number of disparate fields, including Natural Language Processing (NLP), Intelligent Computer-Assisted Language Learning (ICALL), Russian Linguistics, and Second Language Acquisition (SLA).

2.1. Pedagogical foundations

Textbooks of Russian typically state spelling rules and contain explanations about pronunciation. However, the connection between this material and what it means for confident writing skills is underrepresented. In other words, students may learn that they should pronounce the letter е like an и when unstressed, or that the letter *и* sounds like *ы* when preceded by *ш* or *ж*. But students are not warned that these conventions will present challenges in spelling. Furthermore, these rules are typically not exercised in any systematic way and tend to remain peripheral from the students' perspective.

Traditional textbooks take an instruction-based perspective, with the idea of mere transfer of knowledge. A better model for pedagogy is learning by doing, whereby each student constructs their own knowledge network through active engagement. This framework, which is known as constructivism (Biggs 1999, Biggs & Tang 2011), promotes student-centered learning activities both within and outside the classroom. When a student of Russian makes a spelling error, RuMOR can capitalize on that event as an opportunity to engage students with targeted feedback on the relevant spelling and pronunciation conventions. A spelling error is something that is directly relevant to the student in the moment, thus opening up

[1] http://www.web-corpora.net/RLC/rulec

a "teachable moment", when the student is receptive to improvement of their skills. When used over time, RuMOR will engage each student with all of the typical errors that they need to focus on.

2.2. Morphological analysis

The Russian language has widespread fusional morphology, with each major word class having multiple inflection classes. Since the complexity of the morphological system is itself the source of many errors, a morphological analysis is frequently essential for determining what feedback will be most helpful to the user. Table 1 shows two authentic orthographic errors which, at the surface level, appear to be the same — mistakenly replacing *u* with *e* — but which are motivated by entirely different parts of the linguistic system.

Table 1. **Different underlying motivations for identical surface substitutions**

Correct form	Erroneous form	Substitution	Motivation
Марии 'Maria'	*Марие*	*u* → *e*	inflectional
умирает 'dies'	*умерает*	*u* → *e*	phonological

The erroneous *Марие* is morphologically motivated by the fact that the default Locative singular (and for feminine nouns like this one, Dative singular) ending is *-е*, but the writer has failed to take into account the exceptional rule that nouns in *-ия* take instead the ending *-и*. The incorrect spelling of *умерает* is phonologically motivated by the fact that the pronunciation of *е* is indistinguishable from that of *и* in unstressed syllables, and in all forms of this verb the stress is on the vowel *а*.

The output of traditional spellcheckers would be able to tell the user what substitution is needed to correct the error, but it would be inadequate for determining feedback that helps get at the root of the mistake. On the other hand, a morphological analyzer that is sensitive to the grammatical structure of words can model errors such that these two errors can be linked to distinct and appropriate feedback that is relevant to the different factors that led to the error.

Approaches to automatic morphological analysis of Russian have historically gravitated toward rule- and lexicon-based methods. One reason for this is the existence of the seemingly prescient Grammatical dictionary of Russian (Zaliznjak 1977), which specifies the inflectional patterns of more than 100 000 words. On the basis of this dictionary, computational linguists have produced many Russian morphological analyzers/taggers. These include RUSTWOL (Vilkki 1997, 2005), StarLing (ENA, April 17, 2022)[2] (Krylov & Starostin 2003), DiaLing (ENA, April 17, 2022)[3], Mystem (Nozhov, 2003)[4] (Segalovich 2003), pymorphy2 (ENA, April 17, 2022)[5] (Korobov 2015, Boxarov et al. 2013), and UDAR (ENA, April 17,

[2] http://starling.rinet.ru/downl.php
[3] http://www.aot.ru (In Russ.)
[4] https://yandex.ru/dev/mystem/
[5] https://yandex.ru/dev/mystem/

2022)[6]. Although all of these analyzers could theoretically be augmented or adapted to provide more informative feedback than a traditional spellchecker, UDAR is best suited to our needs for a number of reasons. First, it is free and open-source, which facilitates operating in an Open Research paradigm. Second, it includes specification of word stress position, which is crucial for predicting some kinds of spelling errors. Third, it is integrated with a Constraint Grammar, a framework designed to deal with inherent ambiguity, a property which errors are notorious for. Fourth, the finite-state paradigm enables extremely fast lookup times, avoiding procedural logic at runtime.

2.3. Spelling and grammar correction

Rozovskaya and Roth (2019) classified errors from the RULEC corpus (The Russian Learner Corpus of Academic Writing, Alsufieva et al. 2012), and found that spelling errors were by far the most frequent class of errors, accounting for 18.6% of non-native errors and 42.4% of heritage speaker errors. Since spelling errors are by definition limited to the modality of writing, it seems safe to say that most, if not all, of these errors are a direct reflection of *writing* proficiency, as opposed to general language proficiency. Therefore, significant improvement in spelling ability is one of the most straightforward paths to build writing confidence and proficiency.

In recent years, there has been a significant uptick in research on spelling correction for Russian (Sorokin 2017), including SpellRuEval, a competition on automatic spelling correction for Russian (Sorokin et al. 2016). However, so far these research projects have understandably been focused only on surface-level correction, without regard to the underlying linguistic sources of the errors. A natural result of this narrow focus is that grammatical input is generally not included because it is not helpful to these models. Whereas grammatical awareness is a sometimes crucial element of pedagogically oriented spelling correction, the official report from SpellRuEval states that adding morphological and semantic features to these models for traditional spelling correction yields little to no gains.

Research on automatic grammatical error correction has been dominated by studies of English, but Rozovskaya and Roth (2019, 2021) have recently extended this research to Russian as well, with impressive results for certain kinds of errors. Although their research path is promising, it falls short for our application in the same way that recent spelling correction does: the training data — and by extension the outputs of the models – do not contain hypotheses about *why* errors are made.

2.4. Intelligent Language Tutoring Systems (ILTS)

Intelligent Language Tutoring Systems (ILTS) use Natural Language Processing to provide individualized feedback to users without the need for human

[6] https://github.com/giellalt/lang-rus and https://github.com/reynoldsnlp/udar; UDAR is an abbreviated form of udarénie ‘word stress’, and it is also a recursive acronym: “UDAR Does Accented Russian.”

graders or tutors. Historically, research on ILTS has been focused on workbook-style exercises with tightly controlled context (Heift 2010, Nagata 2009, Amaral & Meurers 2011, Choi 2016; Meurers et al. 2019). In these systems, limiting the context allows the designers to anticipate what kinds of feedback are appropriate. The more controlled the context, the less sophisticated the language analysis needs to be. Conversely, providing feedback on every aspect of language with unlimited context in an ILTS would require something near artificial general intelligence.

One departure from the strategy of tightly controlling the context for feedback in ILTS is the Revita system (Kopotev et al. 2019), which allows users to upload their own texts in a number of languages, including Russian, and generate workbook exercises for that text. Notably, the feedback for incorrect responses is generally limited to connecting the mistake to another word in the sentence that governs the target word, or with which the target word should agree. Unlimited possibilities require limited feedback.

While the goal of RuMOR is also to provide feedback to any arbitrary text entered by the user, it is limited to spelling errors, which tend to be interpretable without reference to any surrounding context. Because the scope of the task is limited to only spelling errors, it is possible to provide detailed feedback with high confidence that the feedback will be germane.

Given the fact that all major Russian morphological analyzers are lexicon- and rule-based, the most natural approach to analyzing Russian produced by non-native speakers in an ILTS is through the use of mal-rules (cf. Sleeman 1982, Mathews 1992). Mal-rules are rules that are added to license structures that are not valid in the standard language, but are expected in non-native language production. For example, UDAR uses two-level orthographic and phonological rules[7] to generate standard Russian surface forms from an underlying representation. By modifying or deleting subsets of these rules, one can compile an analyzer that recognizes erroneous wordforms of the sort that non-native writers produce.

3. Methodology

In this section, we describe the methods used to 1) identify the classes of errors to model in our analyzer, 2) augment UDAR to label these errors, and 3) implement the analyzer in the RuMOR webapp.

3.1. Classifying RULEC errors

Russian morphology is more complex than that of many major world languages, and the size of the paradigms, as well as the large number of arcane exceptions, pose a significant challenge. Although RuMOR is not designed to teach inflectional morphology, there are a number of morphophonological phenomena, such as stem alternations, that directly lead to spelling mistakes. Orthographies tend

[7] Cf. Koskenniemi, Kimmo. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. Technical report, University of Helsinki, Department of General Linguistics.

to accrete idiosyncratic conventions that can be especially obscure to non-native writers, and Russian orthography is rife with challenges. Russian orthography can be characterized as morphophonemic, as it does not always reflect phonological phenomena, such as vowel reduction, consonant voicing assimilation and final-obstruent devoicing.

In order to determine which errors should be included in our model, we turned to the Russian Learner Corpus of Academic Writing (RULEC) (Alsufieva et al. 2012), currently the largest freely available corpus of Russian writing produced by non-native users. It consists of approximately 560 000 words, written by 15 non-native and 13 heritage writers, all residing in the United States. We analyzed the corpus using the `udar` (ENA, Arpril 17, 2022)[8] python package to output a list of all words not recognized by the analyzer. This method admittedly overlooks real-word errors, but we suspect that such errors are extremely infrequent in this corpus because opportunities for homophone errors in Russian are mostly limited to a few rare word pairs that are confusable due to final devoicing/voicing assimilation, such as *лук* 'onion' vs. *луг* 'meadow', both pronounced with final [k].

After generating this list of unrecognized tokens, we constructed a frequency distribution of errors and manually classified the tokens according to whether we believed the token was an actual error, or simply a valid token that UDAR did not recognize, such as the acronym *СПбГУ* 'Saint Petersburg State University'. For those tokens that we believe are spelling errors, we classified them linguistically according to the motivation behind the error, relying on our expertise as professional linguists and teachers. Each of these error tags is discussed in the following subsections.

The goal of RuMOR is to improve mastery of Russian orthography by making generalizations that users can apply in the future. In this sense, RuMOR has a different and more advanced linguistic goal than that of a spell-checker. Since RuMOR relies on linguistic analysis, it seizes upon spelling errors as teachable moments when it is most appropriate to deliver systematic explanations. Therefore, the tags are linguistically motivated rather than aimed at simple correction. Each tag can be considered an index to link the error to a relevant mini-lesson to help correct the error.

3.1.1. Overview of error tags

Table 2 contains a summary of the error tags currently included in our spelling model and webapp. The "Tag" column is the name of the tag, as implemented in UDAR. Many of the tag names merely describe the substitution that caused the error, so "a2o" means that the letter "*a*" was erroneously spelled as an "*o*". The "Linguistic label" column is a short pithy description of how to fix the error. More detailed descriptions of the error types are given in the "Tag explanation" column, and relevant examples of misspelled words are provided in the "Examples" column.

[8] https://github.com/reynoldsnlp/udar

Table 2. **Summary of error tags**

Tag	Linguistic label	Tag explanation	Example(s)
a2o	*о→а*	Misspelling (*о* should be *а*)	*озночает*
e2je	*е→э*	Misspelling (*е* should be *э*)	*ето*
FV	no fill vowel	Presence of unnecessary fleeting vowel	*отеца*
H2S	*ъ→ь*	Misspelling (*ь* should be *ъ*)	*подьезд*
i2j	*й→и*	Misspelling (*й* should be *и*)	*миллйард*
i2y	*ы→и*	Misspelling (*ы* should be *и*)	*блызко*
ii	*ие→ии*	*ие* should be *ии*	*Марие*
Ikn	*и→е/я/а*	Ikanje (*и* should be *е/я/а*)	*дитей*
j2i	*и→й*	Misspelling (*и* should be *й*)	*рабочии*
je2e	*э→е*	Misspelling (*э* should be *е*)	*проэкта*
NoFV	add fill vowel	Missing fleeting vowel	*окн*
NoGem	add double letter	Geminate letter is missing	*имено*
NoSS	add *ь*	Misspelling (*ь* is missing)	*болше*
o2a	*а→о*	Akanje (*а* should be *о*)	*каторый*
Pal	add softening	Missing palatalization at stem-ending interface	*землу*
sh2shch	*щ→ш*	Misspelling (*щ* should be *ш*)	*лучще*
shch2sh	*ш→щ*	Misspelling (*ш* should be *щ*)	*вообше*
ski	*ский→ски*	*по-~ский* instead of *по-~ски*	*по-русский*
SRo	*о→е*	Spelling Rule *о>е*	*нашой*
SRy	*ы→и*	Spelling Rule *ы>и*	*книгы*
y2i	*и→ы*	Misspelling (*и* should be *ы*)	*описивают*
prijti	*прийти*	Misspelling the stem of прийти	*прийду*
revIkn	*е/я/а→и*	Reversed Ikanje (*е*, *а*, *я* should be *и*)	*умерает*
Gem	no double letter	Should be just single, not geminate letter	*рассширить*

3.1.2. Fill vowels

Fill vowels (also known as “fleeting” or “mobile” vowels) are vowels that are only realized if there is no inflectional ending, or if the inflectional ending does not begin with a vowel. For example, *окно* ‘window.SG.NOM’ has an inflectional ending, so there is no fill vowel, but *окон* ‘window.PL.GEN’ has no inflectional ending so the fill vowel appears between the *к* and the *н*.

Fill vowel errors clearly demonstrate both the linguistic motivation for our project, as well as the methodological necessity of a morphological analyzer. There are generalizations that help predict which fill vowels appear in what contexts, but ultimately, they are lexically specified and must be memorized. A traditional spellchecker cannot identify that a particular letter omission or insertion is related to fill vowels, so it cannot direct users to remedial resources. Further, because the “rules” for fill vowels have many exceptions, it is essential to rely on a structured lexicon, such as that in UDAR, to model which errors are related to fill vowels.

We currently have two fill vowel (FV) error tags. The FV tag indicates the presence of a fill vowel that should not be present, and the NoFV tag indicates the absence of a fill vowel that should be present. Since users tend to think in terms of generating oblique forms from the lemma, these tags are far more likely to appear on oblique forms (e.g., erroneous *отеца* ‘father.SG.GEN.FV’ which should be

отца, and erroneous *окн* ‘window.PL.GEN.NoFV’ which should be *окон*, etc.) as opposed to the lemma, which users are most familiar with, (e.g., errors such as *отц* ‘father.SG.NOM.NoFV’ instead of correct *отец* or *оконо* ‘window.SG.NOM.FV’ instead of correct *окно* are quite rare). Our analyzer recognizes all of these forms.

3.1.3. Vowel reduction

Russian vowels are always spelled as they would be pronounced if they were stressed, despite the fact that the sounds of some vowels are very different when they are not stressed. What sounds like unstressed [i] might be spelled *и*, *е*, *а*, or *я*; and what sounds like unstressed [a] might be spelled *а* or *о*. Spelling unstressed vowels is therefore a major challenge, even for native Russian speakers. Native speakers can often solve this problem by remembering a related word or wordform where the given vowel is stressed. For example, to spell *река́* [rikˈa] ˈ‘river.SG.NOM’ a native speaker can think of a form of the word with different stress, such as *ре́ку* [r’ˈeku] ‘river.SG.ACC’. However, non-native users have more limited relevant knowledge to draw on, and vowel reduction is one of the most frequent causes for spelling errors in the RULEC corpus.

The pronunciation of an orthographic *о* as [a] is called “akanje” by linguists, and the associated spelling error is tagged o2a. The pronunciation of orthographic *е*, *а*, or *я* after palatalized consonants as [i] is called “ikanje”, and the associated spelling error is tagged “Ikn”. These are the most common error tags for vowel reduction. However, we were surprised to find that akanje and ikanje create enough confusion in the minds of users that they sometimes do the exact opposite (hypercorrection). The tag a2o identifies instances where an orthographic *а* is replaced by *о*, even though it is pronounced [a], as with the token *озночает* ‘signify.PRS.3P.SG.a2o’ (cf. correct *означает*). Similarly, the tag revIkn identifies instances where an orthographic *и* is replaced by *а*, *е*, or *я*, as with the token *умерает* ‘die.PRS.3P.SG.revIkn’ (cf. correct *умирает*).

3.1.4. Phonetic competence

Depending on a user's first language, some of the sounds of Russian are difficult to distinguish, so choosing between letters whose sounds seem indistinguishable is a common problem.

The first instance of confusion that we model is between the letters *ш* and *щ*, both representing voiceless fricatives that English-speaking users associate with “sh”. The prior is post-alveolar, and the latter is palatal. Whether because of the similarity of the orthographic symbols or the similarity of the sounds, non-native writers frequently substitute these letters for one another. The tag sh2shch identifies instances where *ш* has been replaced by *щ*, as with the erroneous token *лучще* ‘better.ADV.sh2shch’ (cf. correct *лучше*). Conversely, the tag shch2sh marks instances where *щ* has been replaced by *ш*, as in erroneous *вообше* ‘generally.ADV.shch2sh’ (cf. correct *вообще*).

Another phonetic difficulty is the distinction between the high central unrounded vowel [ɨ] and the high front vowel [i]. Although linguists do not agree on the phonemic status of [ɨ] and [i], they are represented in standard orthography by two separate letters, *ы* and *и*, respectively. Not only is the vowel [ɨ] difficult to pronounce for many non-native speakers, but it is not represented consistently in standard orthography. Although the vowel [ɨ] is mostly represented by the letter *ы*, in some contexts it is written as *и*, most notably when preceded by the letters *ж* or *ш*. The difficulty of phonetic competence, combined with orthographic inconsistency of [ɨ], leads to many spelling errors substituting these letters for one another. The tag y2i marks tokens where *ы* has been replaced by *и*, as in *описивают* 'describe.PRS.3P.PL.y2i' (cf. correct *описывают*). The i2y tag marks tokens with the inverse substitution, such as *блызко* 'close.ADV.i2y' (cf. correct *близко*).[9]

Two of our error tags are motivated by a misunderstanding of phonemic palatalization in Russian consonants. In modern usage, the soft sign *ь* indicates that the preceding consonant is palatalized, and the hard sign *ъ* indicates that the preceding consonant is not palatalized. Generally speaking, consonants are assumed to be hard, so the hard sign appears in only one context: between prefixes that end in a consonant, and stems that begin with *е*, *ё*, *ю*, or *я*, as in *подъезд* 'stairwell'. However, given the relative frequency of the visually similar soft sign *ь*, non-native writers frequently use the soft sign in place of the hard sign, as in *подьезд* 'stairwell.H2S' (cf. correct *подъезд*). Similarly, for users that have not acquired palatalization in their language, the role of the soft sign *ь* is difficult to grasp. This leads to its frequent omission, as in *болше* 'bigger/more.NoSS' (cf. correct *больше*).

A prominent feature of Russian phonology is consonant palatalization (commonly referred to as hardness vs. softness). Russian orthography marks consonant hardness or softness by two parallel sets of vowel letters (and the symbols *ь* and *ъ*), so that hard consonants are followed by one set, and soft consonants by the other. When inflecting words, users are prone to change the hardness or softness of the stem-final consonant by using a vowel from the wrong set. In particular, it is most common to change soft consonants to hard consonants. Errors of this type are indicated with the tag Pal, as in the error *землу* 'earth.ACC.Pal' (cf. correct *землю*).

3.1.5. Alphabetic confusion

Some spelling errors are either evidence of misunderstanding of the sounds or roles associated with a given letter, or interference from the alphabet of the user's first language. These errors differ from those in Section 3.1.4 (Phonetic competence) in that the users are proficient at producing and perceiving these sounds, but simply fail to associate the sounds with their corresponding symbols. The first pair of such letters is the vowel letter *и* [i] and the consonant letter *й* [j].

[9] Note that the i2y tag and the SRy tag are complementary. The i2y tag applies anywhere that the SRy tag does not.

Examples of these errors include *рабочии* ‘worker.SG.NOM.j2i’ (cf. correct *рабочий*) and *миллйард* ‘billion.SG.NOM.i2j’ (cf. correct *миллиард*).

Another pair of letters that are easily confused are *е* [je] and *э* [e]. The letter *э* only occurs in a small number of high-frequency types, almost exclusively word-initially. Examples of these errors include *ето* ‘this.e2je’ (cf. correct *это*) and *проэкта* ‘project.SG.GEN.je2e’ (cf. correct *проекта*).

3.1.6. Spelling Rules

A small set of consonant letters have restrictions on which vowel letters are allowed to follow them, in some cases motivated by phonological restrictions at the time of orthographic standardization. The relevant consonants are the so-called hushers (*ж*, *ч*, *ш*, and *щ*), velars (*г*, *к*, and *х*), and the letter *ц*. These spelling rules are generally mentioned by Russian textbooks because they are especially relevant for inflectional endings. However, in many cases textbooks merely state these rules rather than attempting to actively engage students in acquiring them. As a result, such rules tend to remain abstract and students get little opportunity to work out their implications.

The first spelling rule is that after the so-called hushers and *ц*, an unstressed letter *о* is replaced by the letter *е*. Violations of this rule are indicated with the tag SRo, as in the error *нашой* ‘our.FEM.SG.GEN.SRo’ (cf. correct *нашей*).

Another spelling restriction is that after velars or hushers, the letter *ы* is replaced by *и*. Unfortunately, for two of the hushers, this restriction is no longer a valid reflection of modern phonology, since *ж* and *ш* are now non-palatalized consonants. Because of this, not only is the rule sometimes difficult to remember and apply, but it is also phonetically misleading. Violations of this spelling rule are indicated with the tag SRy, as in the error *душы* ‘soul.PL.NOM.SRy’ (cf. correct *души*).

The third spelling rule is one that is not explicitly discussed in any textbooks that we are aware of but is nonetheless a cause for confusion for many non-native speakers. The letter *ц* can be followed by either *ы* or *и*, depending on whether it is in the stem or the inflectional ending. In stems, *ц* is followed by *и* (e.g., *цирк* ‘circus’),[10] and in endings *ц* is followed by *ы*. Violations of this rule are indicated with the tag SRc, as in the error *цыфровой* ‘digital.SRc’ (cf. correct *цифровой*).

3.1.7. По-___ски

Many adjectives ending in *-ский* can be converted to adverbs by adding the hyphenated prefix “*по-*“ and removing the final *й*. For example, *русский* ‘Russian’ becomes *по-русски* ‘in Russian’. Non-native writers frequently forget to remove the final *й*. This error is indicated by the tag ski, as in the error *по-русский* ‘Russian.ski’.

[10] There are a handful of exceptions to this rule, including *цыплёнок* 'chick', *цыган* 'gypsy', *на цыпочках* 'on tiptoe'.

3.1.8. Morphological errors

Another common error is particular to stems ending in an underlying /ij/, whose lemmas orthographically end in *-ий*, *-ие*, and *-ия*, such as *критерий* 'criterion', *здание* 'building', and *Мария* 'Maria'. For such stems, any paradigmatic cell that would otherwise end in *-е* ends in *-и* instead. For all three classes, this includes the locative (i.e. prepositional) case and for feminine nouns, the dative case. Errors regarding this principle are indicated with the tag ii, as in *о критерие* 'about the criterion.LOC.ii' (cf. correct *о критерии*).

3.1.9. Geminates

As in many languages, it is difficult for writers to know which letters are duplicated. Errors that include geminate letters where they should not be are indicated using the tag Gem, as in *колличество* 'quantity.Gem' (cf. correct *количество*).[11] Errors that do not include geminate letters where they should be are indicated with the tag NoGem, as in *искуство* 'art.NoGem' (cf. correct *искусство*).

3.1.10. Прийти

The stem of the lexeme *прийти* 'to come' causes problems for native and non-native speakers alike. The *й* appears in the infinitive *прийти*, but not the indicative: *пришла* 'come.PST.FEM', *придет* 'come.NONPST.3P.SG'. This may feel unexpected when compared with some other prefixed forms of *идти* 'go' which do have *й* in the non-past: *зайдет* 'drop by.NONPST.3P.SG', *пройдет* 'pass.NONPST.3P.SG'. Errors related to this lexeme are indicated with the tag prijti, as in *прийду* 'come.NONPST.1P.SG.prijti' (cf. correct *приду*).

3.2. Automatic error diagnosis: extending UDAR

Each of the sources of errors discussed in Section 3.1 can be formalized in rules defining each of the error types discussed in the previous section. As mentioned in section 2.4, rules that license non-normative words or structures are referred to as mal-rules (cf., e.g., Sleeman 1982, Matthews 1992 and references therein). In this section, we provide an abbreviated overview of the mechanics of applying our mal-rules to UDAR.

UDAR is a finite-state transducer, built using three formalisms: the lexc language for creating the finite-state lexical network; the twolc language for realizing orthographic and morphophonological rules on surface forms; and vislcg3

[11] The insertion of geminates is problematic for practical reasons. The corresponding mal-rule would apply to virtually every letter of every word in the analyzer, exploding the amount of storage/memory required for the analyzer. Although theoretically possible, the Gem tag is usually omitted for practical reasons.

for writing a Constraint Grammar to resolve morphosyntactic ambiguity on the basis of surrounding context.[12] Our mal-rules are applied in one of two ways. First, rules that are sensitive to underlying morphophonological structure–such as ii, FV, NoFV, Pal, and SRo–are implemented as alternative twolc rules.[13] Rules that can be modeled as simple character substitution are implemented as XFST regular expression replace rules.[14] In either case, the process for adding a tag to the transducer is the following.

First, a standard transducer is compiled, using UDAR's original rules. Then, for each tag, the mal-rule is applied to make an error transducer. The standard transducer is subtracted from the error transducer so that only wordforms that were affected by the mal-rule remain. Then, the error tag is added to all forms in the error transducer, and the resulting transducer is added to the standard transducer by disjunction. (ENA, April 17, 2022)[15]. In this way, all of UDAR's original contents are preserved, and all additions are tagged with the appropriate error tags.

To the extent possible, errors are accumulated, one after the other, so that words with more than one kind of error can be recognized. However, several of the rules feed into one another, or could even reverse one another. For example, if e2je were added on top of je2e, the resulting surface form would be identical to the correct form, but would be tagged for both errors. Therefore, the errors were grouped by contexts, and all errors affecting the same context are added in parallel. In this way, errors in different context-groups can stack on one another, but errors in the same context-group do not.

3.2.1. Evaluation

We analyzed the entire RULEC corpus using our augmented analyzer, compiled a list of all types that are tagged as errors, and compared the output of the analyzer with our manual labels. We found that for our target errors, the analyzer has perfect recall, meaning that every token that was manually labelled with one of our target error tags was also labeled by the augmented analyzer as such. However, not all of the errors identified in the corpus fit into these categories. Out of 279 manually labeled error types, our analyzer labeled 124 (44.4%). Out of 999 manually labeled error tokens, our analyzer labeled 467 (46.7%).

[12] The lexc and twolc source files can be compiled using either Xerox Finite-State Tools (XFST) (Beesley and Karttunen 2003) or Helsinki Finite-State Transducer Technology (HFST) (Linden et al. 2011).

[13] For a detailed explanation of how the twolc rules in UDAR function, see chapter 2 of Reynolds, Robert. 2016. *Russian natural language processing for computer-assisted language learning: capturing the benefits of deep morphological analysis in real-life applications.* Ph.D. thesis, UiT – The Arctic University of Norway.

[14] For a detailed explanation of XFST regular expressions, see Beesley and Kartunnen (2003).

[15] The Makefile that builds the error transducer can be found at https://github.com/giellalt/lang-rus/blob/8839887e986ae15a255e3396f08d394e8efac363/src/Makefile_L2

3.3. RuMOR webapp

RuMOR is a free and open-source webapp allowing users to get interactive feedback on Russian spelling errors.(ENA, April 17, 2022)[16] RuMOR was built as a mobile-first webapp, so that it can be used comfortably on desktops, laptops, and mobile devices. Currently, two interface languages are available: English and Norwegian. A screenshot of the app is shown in Figure 1.

The user is prompted to type or paste a text, and upon submitting the text, words identified by our augmented analyzer as spelling errors are turned into clickable links. Tokens are considered errors only if all possible readings are errors, so our system does not currently attempt to handle real-word errors. For example, in Figure 1, the token *эй* 'hey' is obviously intended to be *ей* 'she.DAT', but because the analyzer outputs at least one non-error reading, it is not treated as an error by RuMOR.[17]

When an error is clicked, all possible readings are shown in a pane to the side of the text. For each reading, we display the dictionary form, the type of error that would lead to the attested token, and the corrected form (which is shown by clicking or hovering). The readings are sorted by lemma frequency, so the most likely reading is listed first. In Figure 1, the token *Ана* is selected, and four possible readings are displayed: *она* 'she.o2a', *оно* 'she.o2a', *Анна* 'Anna.NoGem', and Аня 'Anya.Pal'.

When the user clicks on any of the error tags, the error explanation is shown in the next column. These explanations are intended to be as short as possible while still giving enough explanation and examples to be reasonably complete. The explanations are open-source, and hosted separately at (ENA, April 17, 2022)[18]

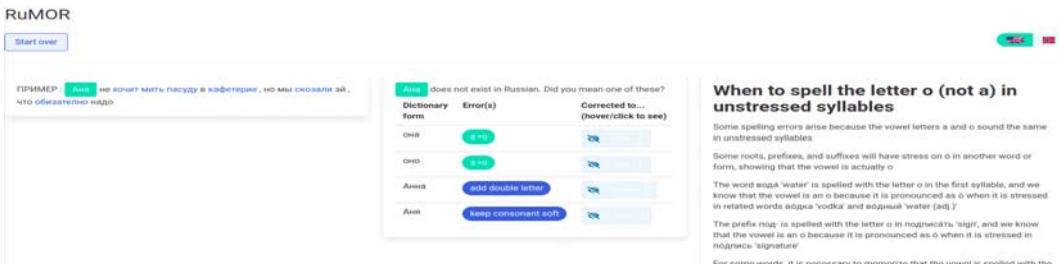

Figure 1. Screenshot of the RuMOR webapp

4. Conclusions and future work

This article has introduced RuMOR, a free, open-source, interactive webapp for identifying, diagnosing, correcting, and explaining a variety of common spelling

[16] The source code for the webapp is available at https://github.com/reynoldsnlp/rus_L2_flask. At the time of writing, the app is accessible at https://icall.byu.edu/rumor.

[17] Although this particular example would be difficult to disambiguate, some real-word errors can be resolved by Constraint Grammar rules which would remove some real-word readings on the basis of the surrounding context.

[18] https://github.com/reynoldsnlp/rus_grammar_explanations.

errors, based on linguistic analysis. The webapp uses a modified version of the UDAR analyzer, which we augmented using mal-rules. The validity of our model was maximized by deriving error tags from real-world errors identified in the RULEC corpus. To our knowledge, this is the first such application for Russian that attempts to provide comparable targeted feedback to any arbitrary running text.

This linguistic approach is especially well-suited to error annotation, but also facilitates text normalization. As demonstrated in the webapp, UDAR can automatically generate the corrected wordform.

Another potential application of our error-augmented analyzer is automatic corpus annotation. Until now, corpora of Russian texts produced by non-native speakers have relied almost exclusively on human annotators to analyze and classify errors. Our analyzer can make this process faster and more consistent by giving annotators a preliminary linguistic analysis of orthographic errors to review.

Future work will focus on adding more classes of errors attested in corpora. These errors include conjugation errors, especially related to stem alternations and inflection class selection. Hapaxes in RULEC were excluded from the present study, but we know that there are some error types represented among them that deserve to be included in our error model. For example, users whose first language uses the Latin alphabet frequently misuse alphabetic false friends, i.e., letters that appear the same as Latin letters, but which represent different sounds. In addition to expanding our spelling error model, we also intend to expand UDAR's existing Constraint Grammar to add syntactic error labels.

Finally, although it is tempting to assume that RuMOR is an effective tool, it is crucial to understand how such tools are actually used, and what effect they have on motivation and proficiency outcomes. We hope to perform evaluations and experiments to understand the outcomes of this project.

REFERENCES

Amaral, Luiz & Detmar Meurers.2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL* 23(1). 4–24.

Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite State Morphology*. Stanford, CA: CSLI Publications.

Biggs, John & Catherine Tang. 2011. *Teaching for Quality Learning at University*. Maidenhead, UK: Open University Press.

Biggs, John. 1999. What the student does: Teaching for enhanced learning. *Higher Education & Development* 18 (1). 57–75.

Bocharov, Victor, Svetlana Alexeeva, Dmitry Granovsky, E. Protopopova, Anastasia Bodrova, Svetlana Volskaya, I.V. Krylova & A.S. Chuchunkov. 2013. Crowdsourcing morphological annotations. *In Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialog"* 1. http://opencorpora.org/doc/articles/2013 Dialog.pdf (accessed 20.04.2022).

Choi, Inn-Chull. 2016. Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning* 29. 334–364.

Heift, Trude. 2010. Developing an Intelligent Language Tutor. *CALICO Journal* 27(3). 443–459.

Kopotev, Mixail, Sardana Ivanova, Anisia Katinskaia & Roman Yangarber. 2019. Corpus-based language teaching tool. *Trudy Mežđunarodnii Konferencii «KORPUSNAYA LINGVISTIKA–2019»*. 30–39. (In Russ.)

Korobov, Mikhail. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. *In Proceedings of AIST'2015*. 320–332. New York: Springer.

Krylov, Sergej & Sergej Starostin. 2003. Upcoming tasks for morphological analysis and generation in the integrated information environment STARLING. *In Proceedings of the International Conference "Dialog 2003"*. https://www.dialog-21.ru/media/2655/krylov.pdf (In Russ.) (accessed 20.04.22).

Linden, Krister, Erik Axelson, Sam Hardwick & Tommi A. Pirinen. 2011. HFST– framework for compiling and applying morphologies. In Cerstin Mahlow & Michael Pietrowski (eds.), *Systems and frameworks for computational morphology,* 100 of Communications in Computer and Information Science, 67–85. New York: Springer.

Matthews, Clive. 1992. Going AI: Foundations of ICALL. *Computer Assisted Language Learning* 5(1). 13–31.

Matthews, Clive. 1992. Going AI: Foundations of ICALL. *Computer Assisted Language Learning* 5(1). 13–31.

Meurers, Detmar, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz &Ramon Ziai.2019. Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics* 39.

Nagata, Noriko. 2009. Robo-Sensei's NLP-Based Error detection and feedback generation. *CALICO Journal* 26(3). 562–579.

Rozovskaya, Alla & Dan Roth. 2019. Grammar Error Correction in Morphologically Rich Languages: The Case of Russian. *Transactions of the Association for Computational Linguistics* 7. 1–17. https://doi.org/10.1162/tacl_a_00251

Rozovskaya, Alla & Dan Roth. 2021. How Good (really) are Grammatical Error Correction Systems? *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*: Main Volume. 2686–2698.

Segalovich, Ilya. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. *In International Conference on Machine Learning; Models, Technologies and Applications*. 273–280.

Sleeman, Derek. 1982. Inferring (mal) rules from pupil's protocols. *In Proceedings of the 5th European Conference on Artificial Intelligence (ECAI)*. 160–164. Orsay, France.

Vilkki, Liisa. 2005. RUSTWOL: A tool for automatic Russian word form recognition. In Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund & Anssi Yli-Jyrä (eds.), *Inquiries into words, constraints and contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday*, 151–162. Stanford, CA: CSLI Publications.

Vilkki, Liisa. 1997. *RUSTWOL: A System for Automatic Recognition of Russian Words.* Technical report, Lingsoft, Inc.

Vilkki, Liisa. 2005. RUSTWOL: A tool for automatic Russian word form recognition. In Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H., and Yli-Jyrä, A. (eds.), *Inquiries into Words, Constraints and Contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday*, 151–162. CSLI Publications.

Dictionaries

Zaliznjak, Andrej A. 1977. Grammatical dictionary of the Russian language: Inflection: Approx 100 000 words. *Russkij Jazyk.* (In Russ.)

Article history:
Received: 20 October 2021
Accepted: 21 January 2022

Bionotes:

Robert REYNOLDS is employed as Assistant Research Professor in the Office of Digital Humanities at Brigham Young University. He holds a PhD in Russian Language Technology from UiT The Arctic University of Norway. His research interests include Intelligent Computer-Assisted Language Learning (ICALL), Natural Language Processing for low-resource languages, automatic analysis of text complexity/readability, automatic reading proficiency assessment using eye-tracking, structure of Russian, and morphological complexity.

Contact Information:
Brigham Young University
Brigham Young University Provo, UT 84602
e-mail: robert_reynolds@byu.edu
ORCID: 0000-0003-0306-087X

Laura A. JANDA is Professor of Russian in the Department of Language and Culture at UiT The Arctic University of Norway. She holds a PhD in Slavic Linguistics from UCLA (1984). She pursues research in the framework of cognitive linguistics applied mostly to the analysis of grammatical categories and constructions in Russian using corpus data. She also works on the development of research-based electronic resources for learners of Russian.

Contact Information:
UiT The Arctic University of Norway
UiT Norges arktiske universitet
Postboks 6050 Langnes 9037 Tromsø
e-mail: laura.janda@uit.no
ORCID: 0000-0001-5047-1909

Tore NESSET is Professor of Russian linguistics in the Department of Language and Culture at UiT The Arctic University of Norway. He received his doctoral degree from the University of Oslo in 1997. His research interests include corpus and cognitive linguistics which he applies to the study of Russian and Norwegian. He also works on historical linguistics and is the author of the widely used textbook *How Russian Came to Be the Way It Is* (2015).

Contact Information:
UiT The Arctic University of Norway
UiT Norges arktiske universitet
Postboks 6050 Langnes 9037 Tromsø
e-mail: tore.nesset@uit.no
ORCID: 0000-0003-1308-3506

Сведения об авторах:
Роберт РЕЙНОЛЬДС – доцент-исследователь в Отделе цифровых гуманитарных наук Университета Бригама Янга. Имеет докторскую степень по языковым технологиям в русском языке, полученную в Арктическом университете Норвегии. Его исследовательские интересы включают обучение языку с помощью интеллектуальных компьютерных технологий (ICALL), обработку естественного языка для малоресурсных языков, автоматический анализ сложности/читабельности текста, автоматическую оценку навыков чтения с помощью айтрекинга, структуру русского языка и морфологическую сложность языков.
Контактная информация:
Brigham Young University
Brigham Young University Provo, UT 84602
e-mail: robert_reynolds@byu.edu
ORCID: 0000-0003-0306-087X

Лора А. ЯНДА – профессор кафедры языка и культуры Арктического университета Норвегии, степень доктора наук получила в Калифорнийском университета в Лос-Анджелесе (1984), специалист по славянскому языкознанию. Сфера интересов включает когнитивную и корпусную лингвистику, грамматические категории русского языка, а также создание электронных ресурсов исследовательского типа для изучающих русский язык.
Контактная информация:
UiT The Arctic University of Norway
UiT Norges arktiske universitet
Postboks 6050 Langnes 9037 Tromsø
e-mail: laura.janda@uit.no
ORCID: 0000-0001-5047-1909

Type НЕССЕТ – профессор кафедры языка и культуры Арктического университета Норвегии. Докторскую степень получил в Университете Осло в 1997 году. Его исследовательские интересы включают корпусную и когнитивную лингвистику применительно к русскому и норвежскому языкам. Он также работает в области исторической лингвистики и является автором широко известного учебника *How Russian Came to Be the Way It Is* (2015).
Контактная информация:
UiT The Arctic University of Norway
UiT Norges arktiske universitet
Postboks 6050 Langnes 9037 Tromsø
e-mail: tore.nesset@uit.no
ORCID: 0000-0003-1308-3506

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 409–425
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30118

Research article

Collection and evaluation of lexical complexity data for Russian language using crowdsourcing

Aleksei V. ABRAMOV ✉, Vladimir V. IVANOV

Kazan Federal University, *Kazan, Russia*
✉AlVAbramov@stud.kpfu.ru

Abstract
Estimating word complexity with binary or continuous scores is a challenging task that has been studied for several domains and natural languages. Commonly this task is referred to as Complex Word Identification (CWI) or Lexical Complexity Prediction (LCP). Correct evaluation of word complexity can be an important step in many Lexical Simplification pipelines. Earlier works have usually presented methodologies of lexical complexity estimation with several restrictions: hand-crafted features correlated with word complexity, performed feature engineering to describe target words with features such as number of hypernyms, count of consonants, Named Entity tag, and evaluations with carefully selected target audiences. Modern works investigated the use of transforner-based models that afford extracting features from surrounding context as well. However, the majority of papers have been devoted to pipelines for the English language and few translated them to other languages such as German, French, and Spanish. In this paper we present a dataset of lexical complexity in context based on the Russian Synodal Bible collected using a crowdsourcing platform. We describe a methodology for collecting the data using a 5-point Likert scale for annotation, present descriptive statistics and compare results with analogous work for the English language. We evaluate a linear regression model as a baseline for predicting word complexity on handcrafted features, fastText and ELMo embeddings of target words. The result is a corpus consisting of 931 distinct words that used in 3,364 different contexts.
Keywords: *Lexical complexity, Russian language, annotation, corpora, Bible*

For citation:
Abramov, Aleksei V. & Vladimir V. Ivanov. 2022. Collection and evaluation of lexical complexity data for Russian language using crowdsourcing. *Russian Journal of Linguistics* 26 (2). 409–425. https://doi.org/10.22363/2687-0088-30118

Научная статья

Сбор и оценка лексической сложности данных для русского языка с помощью краудсорсинга

А.В. АБРАМОВ✉, В.В. ИВАНОВ

Казанский (Приволжский) федеральный университет, *Казань, Россия*
✉AlVAbramov@stud.kpfu.ru

Аннотация
Оценка сложности слова с помощью бинарной или непрерывной метки является сложной задачей, изучение которой проводилось для различных доменов и естественных языков. Обычно данная задача обозначается как идентификация сложных слов или прогнозирование лексической сложности. Корректная оценка сложности слова может выступать важным этапом в алгоритмах лексического упрощения слов. Представленные в ранних работах методологии прогнозирования лексической сложности нередко предлагались с рядом ограничений: авторы использовали вручную созданные признаки, которые коррелируют со сложностью слов; проводили детальную генерацию признаков для описания целевых слов, таких как количество согласных, гиперонимов, метки именованных сущностей; тщательно выбирали целевую аудиторию для оценки. В более современных работах рассматривалось применение моделей, основанных на архитектуре Transformer, для извлечения признаков из контекста. Однако большинство представленных работ было посвящено алгоритмам оценки для английского языка, и лишь небольшая часть переносила их на другие языки, такие как немецкий, французский и испанский. В данной работе мы представляем набор данных для оценки лексической сложности слова, основанный на Синодальном переводе Библии и собранный с помощью краудсорсинговой платформы. Мы описываем методологию сбора и оценки данных с помощью шкалы Лайкерта с 5 градациями; приводим описательную статистику и сравниваем ее с аналогичной статистикой для английского языка. Мы оцениваем качество работы линейной регрессии как базового алгоритма на ряде признаков: вручную созданных; векторных представлениях слов fastText и ELMo, вычисленных на основе целевых слов. Результатом является корпус, содержащий 931 словоформу, которые встречались в 3364 различных контекстах.
Ключевые слова: *лексическая сложность, русский язык, разметка, корпус, Библия*

Для цитирования:
Abramov A.V., Ivanov V.V. Collection and evaluation lexical complexity data for Russian language using crowdsourcing. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 409–425. https://doi.org/10.22363/2687-0088-30118

1. Introduction

This paper introduces a new dataset for lexical complexity prediction in Russian. Automatic predicting of lexical complexity can be useful in many areas such as readability assessment and text simplification (Dale 1948: 37–54, Devlin 1998). Typically, this task is formulated as mapping a word in a context with a complexity score on a certain scale. For instance, a selected word in a sentence may be assigned a binary label (complex/non-complex), or a score on the Likert scale (from 1 to 5). In recent works, this task has been studied in both multiple-domain settings, where lexical complexity depends on a subject domain of a text (e.g., biblical text, biomedical articles and proceedings of the European Parliament) and

cross-lingual settings (e.g., English, German, Spanish) (Yimam 2018: 66–78). Basic parameters that can affect lexical complexity include a variety of lexical features, including word length, frequency features, character N-grams, and word embeddings[1]. The features that represent words as vectors can be used for fitting a machine learning model to the existing labeled dataset. A general approach of application machine learning models (such as Random Forest, Neural Network or Support Vector Machines) in Computational Linguistics and Natural Language Processing can be found in numerous monographs, including, but not restricted to (Manning & Schutze 1999, Nitin & Damerau 2010, Clark 2013).

Moreover, with the advances in machine learning and natural language processing (Delvin et al. 2018), pre-trained neural language models can be applied in the task of lexical complexity prediction in context (Shardlow 2021: 1–16). A comprehensive overview of computational linguistics methods applied in complexology can be found in (Solovyev et al. 2022). However, labeled datasets are still needed to fine-tune such models. At the same time, a task for multilingual lexical complexity prediction was studied for a limited number of languages. For instance, cross-lingual features for complexity prediction are studied at the level texts in (Morozov et al. 2022), while an neural approach is analyzed in (Sharoff 2022) Therefore, the main aim of the paper is to leverage existing methodology for the development of a Russian dataset for lexical complexity prediction.

We follow the methodology proposed in (Shardlow 2020: 57–62) which uses crowdsourcing to collect data. We investigate the statistical properties of the dataset to compare it with the English counterpart. The dataset contains 931 distinct words that occurred within 3,364 different contexts. Finally, we carried out a series of experiments for predicting lexical complexity with a simple linear function that uses lexical parameters of words as input and outputs a complexity score (so called linear regression model). The results of the model are close to the results of the same model trained on the English dataset.

2. Related works

In this section, we review the studies of lexical complexity prediction (LCP) focusing on two aspects: (i) dataset construction and (ii) baseline models evaluation. Since 2016, to evaluate methods for the lexical analysis, three shared tasks have been organized (Paetzold 2016: 560–569, Yimam 2018: 66–78, Shardlow 2021: 1–16). The first two initiatives address a very close problem of Complex Word Identification (CWI-2016 and CWI-2018), the latter one deals with the LCP task. In CWI-2016, the goal was to detect a complex English word in a context wherein a word is considered complex if it is difficult to understand for at least one of the annotators – non-native speakers. The training dataset had 2,237 instances, each labeled by 20 annotators, and the test dataset had 88,221 instances. Each word was assigned a binary label, naturally leading to a

[1] Word embedding is a representation of a word in the form of a numerical vector.

classification task. The participants experimented with lexical and statistical features available from the external sources, including Simple Wikipedia, as well as word embeddings. The feature sets served as an input to classifiers leveraging existing machine learning models. The post evaluation done in (Zampieri 2017: 59–63) has shown that the majority of the participating systems performed poorly mostly because of the data annotation flaws and the small size of the training dataset. In CWI-2018, the organizers proposed a new dataset aiming at both multilingual (English, German, French and Spanish) and multi-domain evaluation. In addition to the classification task, the participants of the CWI-2018 were able to solve another task, predicting a probability of the given target word in its particular context being complex (a regression problem).

The LCP-2021 dataset features an augmented version of CompLex, a multi-domain English dataset with texts from annotated using a 5-point Likert scale (1–5) (Shardlow 2020: 57–62) texts represent from three sources/domains: the Bible, Europarl (European Parliament), and biomedicine. The dataset covers 10,800 instances spanning three domains and containing unigrams and bigrams as targets for complexity prediction. The task was to predict the complexity value of words in a context (same tokens may appear in different contexts; on average each token has around 2 contexts). The LCP-2021 Shared task has two sub-tasks: predicting the complexity score for single words; and predicting the complexity score for multi-word expressions. For both subtasks the same performance measures were used to evaluate quality: correlations between human assessments and system results (here, the authors used two measures: Pearson's and Spearman's coefficients that show how well machine ranking corresponds to the human ranking of words)[2], and mean absolute and mean squared errors (MAE and MSE that correspond to average deviation between a score assigned by a machine and a score estimated from human judgements, respectively). The top-performing system (Yaseen 2021: 661–666), which applied modern models, where features are weighted token and context representations derived from very large neural networks that are pre-trained on multi-billion token text corpora, i.e., BERT (Delvin et al. 2018) and RoBERTa (Liu et al. 2019), reported 0.7886 Pearson Correlation in Task 1. However, there are 0.0182 points of Pearson's Correlation separating the systems at ranks 1 and 10. The LCP-2021 dataset has only English contexts, therefore this evaluation has not covered any other language except English. In the present paper, we develop a dataset for Russian using a methodology from (Shardlow 2020: 57–62) as closely as possible, because this can ease further multilingual and multi-domain lexical complexity evaluations. The methodology for data collection includes selecting target words and multi-word expressions (MWE) using predetermined frequency bands to ensure that targets are distributed across different ranges of low to high frequency. Automatic part-of-speech (POS) tagging was used for selecting nouns and MWEs that match certain patterns. In data labeling respect, the methodology leveraged a 5-point Likert scale with the

[2] Pearson's Correlation coefficient was used for final ranking of the results.

following descriptors: “Very easy”, “Easy”, “Neutral”, “Difficult”, “Very Difficult”. Each instance was annotated by 20 workers (annotators from English speaking countries) of a crowdsourcing platform. The labels for each word were transformed into a complexity score on a scale [0,1]. The resulting dataset had the average complexity for words equal to 0.395, with a standard deviation of 0.115. The subset of instances that were extracted from the Bible had the lowest average complexity score (0.387).

In the remaining part of this section, we review studies of CWI and LCP tasks. Most of the works have detailed descriptions of technical details that are more relevant in computer science than in linguistics. Therefore, we decided to focus on the review on the features (or parameters) that different methods use to model word complexity.

In (Yimam et al. 2017), the authors use four different language-independent sets of features: Length and frequency features, Syntactic features, Word embeddings features and Topic features. The authors used three length features: the number of vowels, the number of syllables, and the number of characters in a word; and three sets of frequency features: frequency of the word in Wikipedia, frequency of the word in the Google Web 1T 5-Grams, and frequency of the word in a context. A proper normalization for all the length and frequency features was performed. As syntactic features, the authors use the part of speech (POS) tags of words in different languages and map them into universal POS tags. As word embedding features, they use word2vec representations of content words (both complex and simple), in addition to cosine similarity between the vector representations of the word and its context. To compute topic-related features, the authors use a topic modeling technique LDA (Blei 2003) capable of representing each context as a distribution over topics, which in their turn are represented as distribution over words. The authors used 100 topics and computed cosine similarity between the word-topic vector and the document vector. The best classifiers trained on the described sets of features outperformed baseline results; however, feature analysis was not the primary goal in (Yimam 2017).

In (Kajiwara & Komachi 2018), the authors present their system that participated in the CWI-2018. They experimented with length features (Number of characters and Number of words for MWE) as well as with frequency features extracted from several corpora (Wikipedia, WikiNews, Lang-8). The authors evaluated the importance of features using ablation study on a classification task and found that the frequency features can yield better performance in comparison to probabilistic features extracted from the same corpora. The Lang-8 corpus seems to be more useful for their system than Wikipedia.

In (Aroyehum et al. 2018), the authors compared two approaches: feature engineering and a deep neural network. Both approaches achieved comparable performance on the English test set. The features sets used for training can be divided into several groups: Morphological Features, Syntactic and Lexical Features, Psycholinguistic Features, Word Embedding Distances that served as Features.

In (Malmasi et al. 2016), the authors of the LTG system focused on the use of contextual language model features and the application of ensemble classification methods. Both versions of their systems achieved good performance (second and third place in CWI-2016). They leveraged a core set of features based on estimating n-gram probabilities using web-scale language models from the Microsoft Web N-Gram Service. These probabilities fall into three groups: Word Probability (how likely it is that the word is present in the corpus), Conditional Probability (how likely it is that the word is present in the corpus given the immediate previous word), and Joint Probability (how likely it is that the pairs and triples of words are in the corpus). All of these probabilities help a system modeling the context in which a word appears. In later works on CWI and LCP such information was represented using word embeddings. In addition, the authors use the length of a word as a feature.

A number of novel features were proposed in (Aprosio 2020). Their approach based on the user's native language identifies complex terms by automatically detecting cognates and false friends, using distributional similarity computed from fastText (Bojanowski 2017: 135–146) word embeddings. Similar types of features are used in (Zaharia 2020). To calculate similarity measures between words, the authors apply a technique presented in (Conneau 2017) to learn a linear mapping of two vector spaces that represent monolingual fastText word embeddings (e.g., between Spanish and German) into the same vector space.

The MacSaar (Zampieri 2016) system presented in CWI-2016 based on a simple idea – observing Zipfian frequency distributions computed from text corpus – helps to determine whether a word is complex or simple. The authors calculate the *Zipfian frequency* feature by taking the inverse of the rank of a word. Additionally, word length, normalized sum probability of the character trigrams in a word, sentence length and sum probability of the character trigrams of the sentence were used in their experiments.

In 2021, a number of models and features were evaluated in the new LCP-2021 Shared task in (Shardlow 2021: 1–16). First, we should mention that top-performing systems for lexical complexity prediction used context by means of contextualized pre-trained language models. Those systems, as mentioned above, use deep learning models that make use of the Transformer architecture which in recent years has disrupted the field of natural language processing (Vaswani 2017). During pre-training, such language models are forced to use context in order to reconstruct missing words in a large corpus (usually, multi-billion tokens corpora). In the LCP-2021, the participants used BERT-based models: BERT (Delvin 2018), RoBERTa (Liu 2019), ELECTRA (Clark 2020), ALBERT (Zhenzhong 2019), DeBERTa (He 2020) to encode (i.e., to represent in the form of vectors) both a target word and the input context of the word. Other systems used a variety of features, including lexical frequency and length features, psycholinguistic features that represent human perception of words, semantic features from WordNet to represent word ambiguity or abstractness. The third group of systems combined the

deep learning models with the models trained on engineered feature sets. An extensive exploration of sentence and word features are presented in (Mosquera 2021), where the author investigates feature engineering methods for predicting the complexity of English words in a context using regression models. A substantial set of 51 features was studied, including Word and Lemma lengths, Syllable count, Morpheme length (a number of morphemes for the target word), Google frequency (the frequency of the target word based on Google ngram corpus), two Wikipedia-based word frequencies (one based on the target word occurrences and the other based on the number of documents in Wikipedia where the target word appears), Complexity score taken from a complexity lexicon (Maddela & Xu 2018), Zipf frequency, two Kucera-Francis frequencies: for a target word and for the target word lemma, binary features (is_stopword and is_acronym), Average age of acquisition, Average concreteness, Word and Lemma frequencies in COCA, WordNet-related features (Number word senses, synonyms, hypernyms, hyponyms), Minimum and maximum distances to the root hypernym in WordNet for the target word, Number of Greek or Latin affixes, Year of appearance (the first year when the target and its preceding word appeared in the Google Books Ngram Dataset), as well as a number of SUBTLEX-based features and various readability scores (such as SMOG index, Dale-Chall index, Gunning-Fog, etc.). A list of top ten important features includes age of acquisition, Dale-Chall index, Zipf frequency, average concreteness and lemma frequency.

3. Methodology

Following the methodology proposed for the English language in (Shardlow 2020: 57–62), we chose a Russian parallel translation of the Bible from (Christodouloupoulos 2015: 375–395), based on the Russian Synodal Bible, as the initial corpus. For annotation we selected nouns listed in the Frequency dictionary of modern Russian language (Lyashevskaya 2009), that fall within the following frequency intervals (ipm, instances per million): (2-4), (5-10), (11-50), (51-250), (251-500), (501-1400), (1401-3100). Such restrictions on the choice of part of speech and specific frequency intervals provide us with a basis for a fair comparison with the original methodology. The selection of suitable nouns was performed in such a way that the number of words in each frequency interval was approximately the same for the first four intervals and decreased with the growth of frequency for the rest. We selected 931 distinct words that occurred within 3,364 different contexts. Each word was provided with a surrounding context, such as a Bible verse.

The assessors were asked to estimate the lexical complexity of a highlighted word in a given context using five-level Likert scale with the following items:

1. Very easy: the meaning of the highlighted word is clear;
2. Easy: the meaning of the highlighted word is obvious and the context supplements it;

3. Average[3]: the meaning of the highlighted word is familiar, but it becomes clear only after taking into account the surrounding context;

4. Difficult: the meaning of the highlighted word is not evident, but might be understood after considering the context;

5. Very difficult: the meaning of the highlighted word is unclear or the word itself is unfamiliar.

Compared to the data labeling procedure described for CompLex, we decided to present a more detailed description for each item of the scale, particularly, in terms of impact of the context on the understanding of the word meaning. A detailed explanation for each item could simplify the lexical complexity evaluation for the assessors and the subsequent analysis of the answers.

The words and their surrounding contexts were grouped into samples as in the following sample: "*Их конец – погибель, их бог –* **чрево***, и слава их – в сраме, они мыслят о земном*" ("*Whose end is destruction, whose God is their* **belly***, and whose glory is in their shame, who mind earthly things*"), where the target word is bold type and its context is marked with italics. The collected samples were shuffled and divided into batches of 10 samples each to ensure that every batch had samples with different lexical complexity. Additionally, we split batches into 12 task pools with 30 batches each, except for the last one with 7 batches. Every batch was shown to 10 distinct annotators, so that every word with a corresponding context was evaluated 10 times. We selected assessors from Russia, Ukraine, Belarus and Kazakhstan to introduce speakers with different language skills. A more detailed information about their native language and experience of using Russian could be useful, but unfortunately we were not able to collect such data from the crowdsourcing platform (Yandex.Toloka). To filter assessors with reliable assessments and to gather various opinions, we used the following automatic rules:

- Limited daily earnings: if the assessor completed five tasks per day, he (she) would be suspended for 24 hours;
- The number of skipped assignments: if the assessor skipped more than two assignments in a row, he (she) would be banned for three days;
- Captcha: if at least three out of five last captchas were not recognized, the assessor would be banned for seven days;
- Limit on response time: if at least two out of five latest assignments were completed in less than 15 seconds, the assessor would be banned for seven days;
- Majority vote: if more than five out of the last ten assignments were completed with responses different from the majority (minimum five similar responses), the assessor would be banned for seven days.

We selected the top 10% of the available assessors and paid 10 cents for each evaluated batch. All the gathered evaluations were transformed into [0,1] range and averaged per sample. Examples of words in a context, corresponding complexities and score variance are listed in Table 1 above.

[3] In Russian we use the descriptor "Средняя сложность" (moderate, medium) that better corresponds to the original descriptor "Neutral".

Table 1. **Samples from corpus; target words are in *bold type***

Samples	Complexity	Variance
При **выходе** их из Иудейской синагоги язычники просили их говорить о том же в следующую субботу. (And when the Jews **were gone out** of the synagogue, the Gentiles besought that these words might be preached to them the next sabbath).	0.075	0.11
Моисей весьма огорчился и сказал Господу: не обращай ***взора*** Твоего на приношение их; я не взял ни у одного из них осла и не сделал зла ни одному из них. (And Moses was very wroth, and said unto the Lord, ***Respect*** not thou their offering: I have not taken one ass from them, neither have I hurt one of them).	0.28	0.175
Никакое гнилое слово да не исходит из уст ваших, а только доброе для ***назидания*** в вере, дабы оно доставляло благодать слушающим. (Let no corrupt communication proceed out of your mouth, but that which is good to the use of ***edifying***, that it may minister grace unto the hearers).	0.4	0.19
Услышав об этом, все бывшие в башне Сихемской ушли в башню ***капища*** Ваал-Верифа. (And when all the men of the tower of Shechem heard that, they entered into an hold of the ***house of the god*** Berith).	0.63	0.26

It took 60.4 seconds on average to annotate one batch of samples and 135 assessors on average to complete the task pool. Each assessor annotated 2.19 batches of samples. We did not use any training or control tasks due to the following reasons: 1) the evaluation of lexical complexity is subjective and depends on various factors, such as education, occupation, overall erudition, age (some modern words might be more familiar to a younger audience), and language proficiency (in our research we also included annotations gathered from non-native speakers); thus we cannot reliably provide "correct" answers for tasks to estimate one's accuracy; 2) the use of averaged or majority's answers as ground truth could narrow down the amount of available assessors to those who have similar views on lexical complexities of different words; therefore, we would not be able to estimate the true distribution of lexical complexities performed by people with different background.

4. Analysis

We conducted the distribution analysis of the obtained lexical complexities by estimating their distribution and connection with the word frequency. Figure 1 contains histograms of lexical complexity scores from (Shardlow 2020: 57–62) and our work.

It can be observed that there was a median complexity score equal to 0.225 (denoted as a vertical blue dashed line), wherein the majority of given evaluations are equal to either "Very easy" or "Easy", according to the aforementioned scale. This is consistent with the well known dependency between lexical complexity and word frequency; uncommon words tend to have a higher complexity; therefore, truly rare and difficult words are harder to obtain and less likely to fit in our frequency ranges.

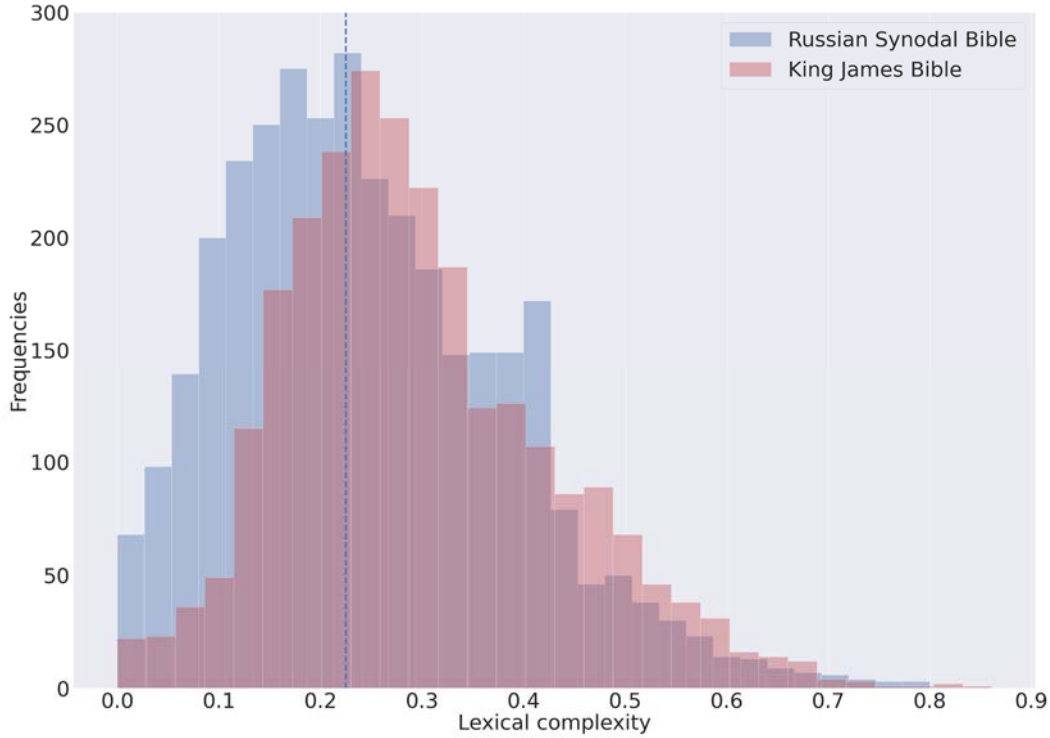

Figure 1. Lexical complexity scores distribution for words selected from King James Bible and Russian Synodal Bible

We also noticed a non-linear dependency between lexical complexity and word frequency; an estimated lexical complexity remains mostly the same almost among all frequency ranges, but starts to rise when close to the lowest frequencies. Indeed, the Pearson correlation between lexical complexity and word frequency is moderately low ($r_f = -0.32$), albeit significant. Additionally, we observed a weak positive correlation between lexical complexity and word length: $r_w = 0.14$. A similar dependency can be observed for the CompLex dataset with a weak negative correlation between lexical complexity and word frequency ($r_f = -0.24$) and slightly stronger positive correlation between lexical complexity and word length ($r_w = 0.28$).

Figures 2 and 3 contain randomly sampled subsets of the corpuses that illustrate such phenomena; the x-axis is depicted in log-scale, lexical complexities are averaged per lemma. This shows that lexical complexity depends on other word features as well, such as length, number of syllables, morphological structure, context, meaning ambiguity, etc.

We also noticed a dependency between word's frequency and variance of lexical complexity scores from different annotators. Figure 3 illustrates this observation, i.e., the variance of complexity scores within certain frequency ranges decreases as range boundaries increase. These results can be explained by the following reason – the less frequent (and, hence, more complex) a word is, the fewer annotators are familiar with it, which translates into higher complexity scores from

annotators who are unfamiliar with the meaning of the word or unable to derive it from the context. But we did not observe the same dependency for the CompLex dataset as shown in Figure 4.

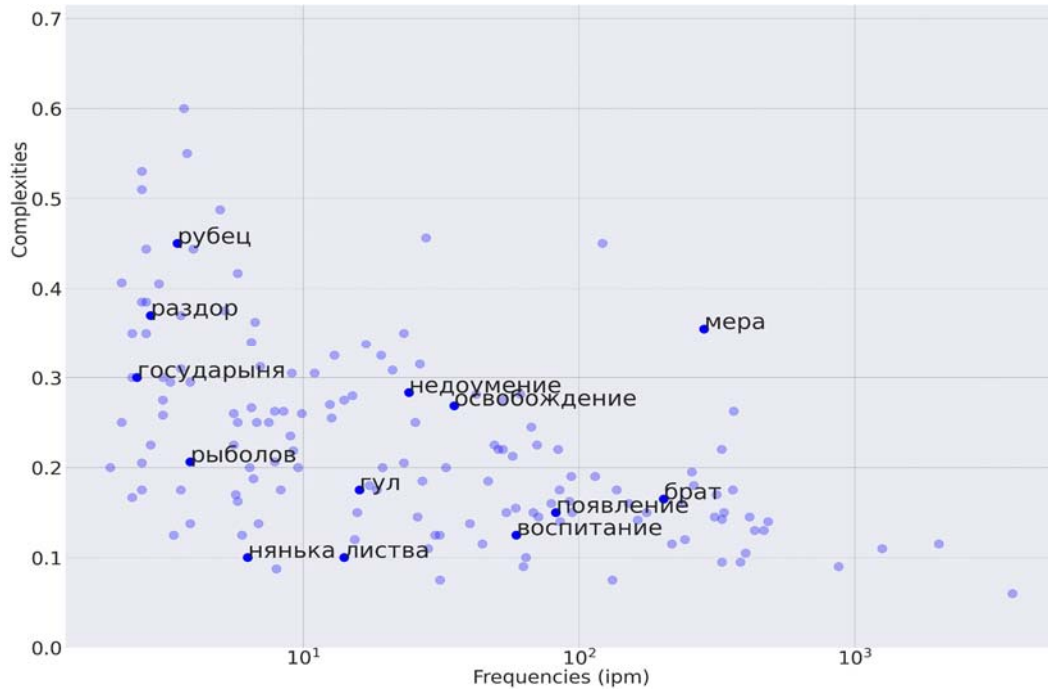

Figure 2. Dependency between word frequency and lexical complexity for words from Russian Synodal Bible

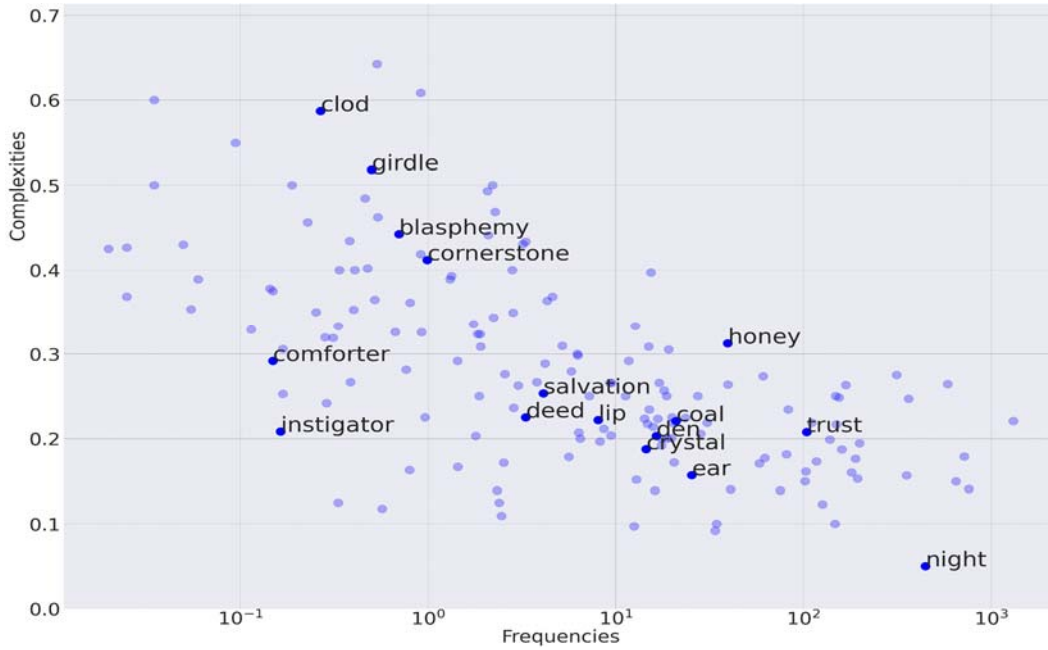

Figure 3. Dependency between word frequency and lexical complexity for words from King James Bible

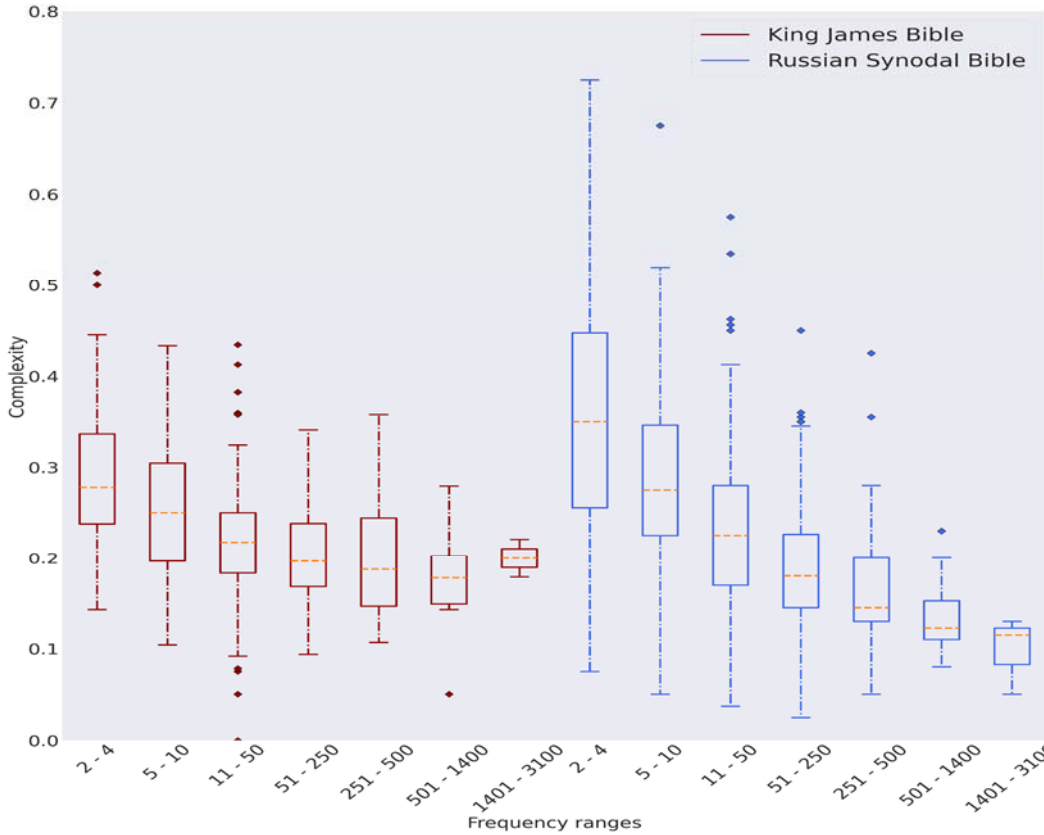

Figure 4. Box plots showing the distribution of lexical complexity scores of words grouped by their frequency from King James Bible (*left*) and Russian Synodal Bible (*right*)

Table 2. **The result of linear regression on handcrafted features (HC), Fasttext and ELMo embeddings and concatenated features**

	Handcrafted	Fasttext	ELMo	Fasttext+HC	ELMo+HC
MAE	0.102	0.084	0.099	0.084	0.099
Pearson correlation	0.342	0.614	0.498	0.619	0.501

Table 3. **Pearson correlation between handcrafted features**

	Frequency	Word length	Number of syllables
Frequency	1	-0.206	-0.172
Word length	-0.206	1	0.819
Number of syllables	-0.172	0.819	1

5. Experiments

To investigate how simple features, such as word frequency and its length, affect lexical complexity, we created a simple baseline. For comparison with CompLex, we selected a linear regression as our model with the following three features: (i) word length, (ii) word frequency according to Lyashevskaya's dictionary and (iii) number of syllables. These features were mentioned by (Shardlow 2020: 57–62) as hand-crafted (HC) features. On target words from the

Complex, linear regression achieved Mean Absolute Error of 0.0888 (for the same set of features). In Table 3, we provide the pairwise correlations between the three HC features in our dataset. In addition, we fitted linear regression using fastText and ELMo embeddings – N-dimensional vector representations of words trained on a large unannotated corpus (Bojanowski 2017: 135–146, Peters 2018: 2227–2237).

FastText embeddings were pretrained on the joint Russian Wikipedia and Lenta.ru corpora; ELMo embeddings were pretrained on the Russian WMT-News corpora; both were taken from DeepPavlov repository (Burtsev M. 2018: 122–127). In our case, we used 300-dimensional embeddings from fastText and 1024-dimensional embeddings from ELMo. For a complete comparison we concatenated embeddings and handcrafted features and applied linear regression as well. As evaluation metrics, we selected Mean Absolute Error and Pearson correlation. Final results were averaged using 10-fold cross-validation.

6. Discussion

The main novel contribution of the work is a new dataset for word-level complexity evaluation in Russian. At present we are not aware of any other resources with a comparable size or coverage. We claim that the dataset also has a comparable quality to its English counterpart. This claim can be supported by a comparison of the complexity scores distribution and the experiments we carried out with the baseline models for lexical complexity prediction. Indeed, this was expected because we applied the same principles to collect and label the data, which led to very similar results. For instance, Figures 4 and 5 illustrate similar behavior for variance of complexity scores, which decay when the word frequency grows. Moreover, experiments with the linear regression model trained on the similar feature sets show similar results (Table 2): on the English dataset MAE value for hand-crafted features was 0.089, while for Russian it is 0.100; training with word embeddings as features provides almost identical results.

Despite these positive findings, we need to mention a few substantial differences between Russian and English datasets. First, complexity score histograms for Russian and English are shifted relative to each other (see Fig. 1); overall, the Russian version contains simpler words. Second, the correlation between word frequency and complexity in the Russian dataset (–0.32) differs from its English counterpart, wherein the correlation coefficient is slightly weaker (–0.24). This histogram shift and the discrepancy in correlation coefficients can be explained by the fact that the King James Bible was published long before the Russian Synodal edition of the Bible and contains more deprecated words and expressions compared to the Russian edition. Hence, the Russian data have simpler labels than the English data.

Our dataset has a few limitations, including a coverage restricted to a single domain (Bible texts) and only single words, without multi-word expressions. We are aiming to overcome the first limitation in our future work, as the methodology that we made use of is already well-studied and has proved to be successful. The

second limitation (lack of MWEs) seems to be important, but less urgent. The LCP-2021 evaluation shows that thye prediction of single word complexity in a context is harder than the MWE complexity prediction.

7. Conclusion

In this paper, we presented a novel dataset for predicting lexical complexity in the Russian language. The dataset has 931 distinct words that occurred within 3,364 different contexts. It was labeled using a crowdsourcing platform (Yandex Toloka). During data collection and labeling we followed a well-studied methodology previously applied in English. We compared our dataset with its English counterpart by two means: 1) we analyzed statistical properties of both datasets; 2) we trained a linear regression model on Russian data and compared its outcomes to its English analog. We found a few discrepancies between datasets which are viewed as potential targets of our further investigation. In our future experiments with the dataset, we expect to develop better models and study extensive feature sets for predicting lexical complexity, which might be important in a broader context of text and discourse complexity studies, as well as the development of automatic complexity analyzers (Solnyshkina et al. 2022).

Acknowledgments

This paper has been supported by the Russian Science Foundation, grant # 22-21-00334, https://rscf.ru/project/22-21-00334/.

REFERENCES

Aprosio, Alessio P., Stefano Menini & Sara Tonelli. 2020. Adaptive complex word identification through false friend detection. *In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization.* 192–200. https://doi.org/10.1145/3340631.3394857

Aroyehun, Segun Taofeek, Jason Angel, Daniel Alejandro Pérez Alvarez & Alexander Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. *In Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications.* 322–327. https://doi.org/10.18653/v1/W18-0538

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3. 993–1022.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5. 135–146.

Burtsev, Mikhail, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva & Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. *Proceedings of ACL 2018, System Demonstrations*. 122–127. https://doi.org/10.18653/v1/P18-4021

Christodouloupoulos, Christos & Mark Steedman. 2015. A massively parallel corpus: The bible in 100 languages. *Language Resources and Evaluation* 2(49). 375–395. https://doi.org/10.1007/s10579-014-9287-y

Clark, Alexander, Chris Fox & Shalom Lappin (eds.). 2013. *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le & Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *In Proceedings of the International Conference on Learning Representations.*

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer & Hervé Jégou. 2017. Word translation without parallel data. *In Proceedings of the International Conference on Learning Representations.*

Dale, Edgar & Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin* 27. 37–54.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (1). 4171–4186. https://doi.org/10.18653/v1/N19-1423

Devlin, Siobhan & John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*. 161–173.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao & Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *In Proceedings of the International Conference on Learning Representations.*

Kajiwara, Tomoyuki & Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. *In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 195–199.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma & Radu Soricut.2019. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.*

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019).

Lyashevskaya, Olga N. & Sergey A. Sharoff. 2009. *The Frequency Dictionary of Modern Russian Language.* Moscow: Azbukovnik. (In Russ.)

Maddela, Mounica & Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* 3749–3760. https://doi.org/10.18653/v1/D18-1410

Malmasi, Shervin, Mark Dras & Marcos Zampieri. 2016. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 996–1000. https://doi.org/10.18653/v1/S16-1154

Manning, Christopher & Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing.* MIT press.

Morozov, Dmitry, Anna Glazkova & Boris Iomdin. 2022. Text Complexity and Linguistic Features: their correlation in English and Russian. *Russian Journal of Linguistics* 26 (2). 425-447.

Mosquera, Alejandro. 2021. Alejandro Mosquera at SemEval-2021 Task 1: Exploring Sentence and Word Features for Lexical Complexity Prediction. *In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).* 554–559. https://doi.org/10.18653/v1/2021.semeval-1.68

Nitin, Indurkhya & Fred J. Damerau (eds.). 2010. *Handbook of Natural Language Processing*. 2nd edn. Boca Raton: CRC Press.

Paetzold, Gustavo & Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 560–569. https://doi.org/10.18653/v1/S16-1085

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1. 2227–2237. https://doi.org/10.18653/v1/N18-1202

Shardlow, Matthew, Michael Cooper & Marcos Zampieri. 2020. CompLex – A New corpus for lexical complexity prediction from Likert Scale Data. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI).* 57–62.

Shardlow, Matthew, Richard Evans, Gustavo Henrique Paetzold & Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).* 1–16. https://doi.org/10.18653/v1/2021.semeval-1.1

Sharoff, Serge. 2022. What neural networks know about linguistic complexity? *Russian Journal of Linguistics.* 26(2). 370–389.

Solnyshkina, Marina, Mcnamara Danielle & Zamaletdinov Radif. 2022. Natural language processing and discourse complexity studies. *Russian Journal of Linguistics.* 26(2). 317–341.

Solovyev, Valery, Marina Solnyshkina & Mcnamara Danielle. 2022. Computational linguistics and Discourse complexology. *Russian Journal of Linguistics.* 26(2). 275–316.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems.* 5998–6008.

Yaseen, Tuqa Bani, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh & Malak Abdullah. 2021. JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models. *In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021).* 661–666. https://doi.org/10.18653/v1/2021.semeval-1.85

Yimam, Seid Muhie, Sanja Stajner, Martin Riedl & Chris Biemann. 2017. Multilingual and cross-lingual complex word identification. *In Proceedings of the International Conference Recent Advances in Natural Language Processing.* 813–822. https://doi.org/10.26615/978-954-452-049-6_104

Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack & Marcos Zampieri. 2018. A report on the complex word identification shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA).* 66–78. https://doi.org/10.18653/v1/W18-0507

Zaharia, George-Eduard, Dumitru-Clementin Cercel & Mihai Dascalu. 2020. Cross-lingual transfer learning for complex word identification. *In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI).* 384–390. https://doi.org/10.1109/ICTAI50040.2020.00067

Zampieri, Marcos, Liling Tan & Josef van Genabith. 2016. Macsaar at semeval-2016 task 11: Zipfian and character features for complex word identification. *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* 1001–1005. https://doi.org/10.18653/v1/S16-1155

Zampieri, Marcos, Shervin Malmasi, Gustavo Paetzold & Lucia Specia. 2017. Complex word identification: Challenges in Data Annotation and System Performance. *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017).* 59–63.

Article history:
Received: 27 October 2021
Accepted: 21 January 2022

Bionotes:
Aleksei V. ABRAMOV is a PhD student at Kazan (Volga region) Federal University (Russia). His research interests comprise applied linguistics, use of computer technologies in the creation of corpora and methods of assessment texts and words complexity.
Contact information:
Kazan (Volga region) Federal University
18 Kremliovskaya str., Kazan, 420008, Russia
e-mail: AlVAbramov@stud.kpfu.ru
ORCID: 0000-0002-5509-9680

Vladimir V. IVANOV is Assistant Professor at Innopolis University (Russia). His scientific interests lie in the field of applied computational linguistics (text complexity analysis, information extraction from text), development and application of knowledge bases and knowledge graphs, as well as the application of machine learning methods in software engineering.
Contact information:
Innopolis University
1 Universitetskaya st., Innopolis, 420500, Russia
e-mail: v.ivanov@innopolis.ru
ORCID: 0000-0003-3289-8188

Сведения об авторах:
Алексей Валерьевич АБРАМОВ – аспирант Казанского (Приволжского) федерального университета. К сферам его научных интересов относятся прикладная лингвистика и применение компьютерных технологий в создании корпусов и методов, посвященных оценке сложности понимания отдельных слов и текстов.
Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: AlVAbramov@stud.kpfu.ru
ORCID: 0000-0002-5509-9680

Владимир Владимирович ИВАНОВ – доцент Университета Иннополис (Россия). Его научные интересы включают прикладную компьютерную лингвистику (анализ сложности текста, извлечение информации из текста), разработку и применение баз знаний и графов знаний, применение методов машинного обучения в разработке программного обеспечения.
Контактная информация:
Университет Иннополис
Россия, 420500, Иннополис, ул. Университетская, д. 1
e-mail: v.ivanov@innopolis.ru
ORCID: 0000-0003-3289-8188

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 426–428
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30132

Research article

Text complexity and linguistic features: Their correlation in English and Russian

Dmitry A. MOROZOV[1] ✉, Anna V. GLAZKOVA[2] and Boris L. IOMDIN[3]

[1]Novosibirsk State University, Novosibirsk, Russia
[2]University of Tyumen, Tyumen, Russia
[3]Vinogradov Russian Language Institute, RAS, Moscow, Russia
✉morozowdm@gmail.com

Abstract
Text complexity assessment is a challenging task requiring various linguistic aspects to be taken into consideration. The complexity level of the text should correspond to the reader's competence. A too complicated text could be incomprehensible, whereas a too simple one could be boring. For many years, simple features were used to assess readability, e.g. average length of words and sentences or vocabulary variety. Thanks to the development of natural language processing methods, the set of text parameters used for evaluating readability has expanded significantly. In recent years, many articles have been published the authors of which investigated the contribution of various lexical, morphological, and syntactic features to the readability level. Nevertheless, as the methods and corpora are quite diverse, it may be hard to draw general conclusions as to the effectiveness of linguistic information for evaluating text complexity due to the diversity of methods and corpora. Moreover, a cross-lingual impact of different features on various datasets has not been investigated. The purpose of this study is to conduct a large-scale comparison of features of different nature. We experimentally assessed seven commonly used feature types (readability, traditional features, morphological features, punctuation, syntax frequency, and topic modeling) on six corpora for text complexity assessment in English and Russian employing four common machine learning models: logistic regression, random forest, convolutional neural network and feedforward neural network. One of the corpora, the corpus of fiction literature read by Russian school students, was constructed for the experiment using a large-scale survey to ensure the objectivity of the labeling. We showed which feature types can significantly improve the performance and analyzed their impact according to the dataset characteristics, language, and data source.
Keywords: *text complexity, machine learning, neural network, corpus linguistics*

For citation:
Morozov, Dmitry A., Anna V. Glazkova & Boris L. Iomdin. 2022. Text complexity and linguistic features: Their correlation in English and Russian. *Russian Journal of Linguistics* 26 (2). 426–448. https://doi.org/10.22363/2687-0088-30132

Научная статья

Сложность текста и лингвистические признаки: как они соотносятся в русском и английском языках

Д.А. МОРОЗОВ[1] ✉, А.В. ГЛАЗКОВА[2], Б.Л. ИОМДИН[3]

[1]Новосибирский государственный университет, Новосибирск, Россия
[2]Тюменский государственный университет, Тюмень, Россия
[3]Институт русского языка им. В. В. Виноградова РАН, Москва, Россия
✉morozowdm@gmail.com

Аннотация
Автоматическая оценка читабельности текста — актуальная и непростая задача, которая требует учёта разнообразных лингвистических факторов. Сложность текста должна соответствовать уровню читателя: слишком сложный текст останется непонятым, слишком простой будет скучным. Исторически для оценки читабельности использовались простые характеристики: средняя длина слов и предложений, разнообразие лексики. Благодаря развитию методов обработки естественного языка набор используемых для оценки параметров текста существенно расширился. За последние годы было опубликовано множество работ, в которых исследовался вклад в сложность текста различных лексических, морфологических, синтаксических признаков. Тем не менее, поскольку использованные методы и корпусы довольно разнообразны, затруднительно делать общие выводы об эффективности различных лингвистических характеристик текста. Более того, не было проведено сравнение влияния признаков для различных языков. Целью настоящего исследования является проведение масштабного сравнения признаков различного характера. Мы экспериментально сравнили семь часто используемых типов признаков (индексы читабельности, традиционные, морфологические, синтаксические, пунктуационные, частотные признаки и тематическое моделирование) на материале трёх русскоязычных и трёх англоязычных корпусов, с использованием четырех распространённых алгоритмов машинного обучения: логистической регрессии, случайного леса, свёрточной нейронной сети и нейронной сети с прямой связью. Один из корпусов — корпус художественной литературы, читаемой российскими школьниками, — был создан для этого эксперимента с помощью масштабного опроса для обеспечения объективности разметки. Мы показали, какие типы признаков могут значительно повысить качество прогнозирования, и проанализировали их влияние в зависимости от характеристик корпуса, его языка и источника текстов.

Ключевые слова: *сложность текста, машинное обучение, нейронные сети, корпусная лингвистика*

Для цитирования:
Morozov D.A., Glazkova A.V., Iomdin B.L. Text complexity and linguistic features: Their correlation in English and Russian. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 426–448. https://doi.org/10.22363/2687-0088-30132

1. Introduction

Text complexity is crucial for the comprehension process. Texts that are too difficult may be hard to understand. In contrast, unnecessarily simple texts may conflict with the reader's level of communicative skills. Hence, text complexity assessment is an essential task that represents a major challenge for developing natural language processing (NLP) tools. Text complexity can be expressed in different ways, ranging from quantitative characteristics to semantic complexity represented by text topics. Numerous studies have been published on evaluating various features for text complexity assessment. The reported results were obtained from text corpora of widely differing sizes and domains. Moreover, the authors used different machine learning (ML) models and text representation techniques (Feng et al. 2010, Ivanov et al. 2018, Cantos & Almela 2019, Isaeva & Sorokin 2020, Deutsch et al. 2020, Glazkova et al. 2021, Martinc et al. 2021). This makes it complicated to achieve an objective evaluation of the impact of different types of features.

The goal of this paper is to perform an extensive evaluation of seven feature types for text complexity assessment that were frequently used in research on the subject. The results allow us to make the text complexity analysis process more defined and transparent. These findings have a broad spectrum of potential applications in education and recommendation systems. We used the following feature types: readability, traditional features, morphological features, punctuation, syntax, frequency, and topic modeling. The features were evaluated on three Russian and three English text complexity corpora and four ML models in order to answer the following research questions (RQ).

- **RQ1:** How do different types of features affect the performance of baselines?
- **RQ2:** Is the impact of these feature types similar in English and Russian?
- **RQ3:** Do feature-enriched models outperform fine-tuned state-of-the-art language models?

This paper is organized as follows. Section 2 contains a brief review of related works. In Section 3, we introduce datasets and models utilized and provide some short background on the feature types we use. Section 4 presents and discusses the results. Section 5 concludes the paper.

2. Related Work

2.1. Readability: methods and approaches

The earliest approaches to automatic readability assessment, developed in the second half of the 20th century, were intuitive, severely limited by the small number

of existing natural language processing tools and the lack of computing power. Most of these readability indices represented linear combinations of simple features, such as average word or sentence length, the proportion of words in the text with a large number of syllables, and the proportion of words included in special lists of "simple" and "complex" words. A more detailed overview of these algorithms is given, for example, by Cantos and Almela (2019).

These algorithms have become widespread in practical tasks, despite their simplicity and seeming naivety. They are still in use in some spheres, including government requirements for insurance, for example, in some US states such as Connecticut (Chapter 699a. Readable language in insurance policies). At the same time, it is quite clear that such simple mechanisms cannot give a reliable result, especially in relation to fiction and poetry.

2.2. New possibilities: more features, more sophisticated models

The rapid development of NLP tools, including neural networks, has made it possible to significantly expand the set of features and to improve the quality of the text complexity assessment. Since the algorithm that solves this problem can be widely used in application-oriented studies, many authors have analyzed the impact of features of different nature (e.g. Feng et al. 2010, Ivanov et al. 2018, Cantos & Almela 2019, Isaeva & Sorokin 2020, Deutsch et al. 2020, Glazkova et al. 2021).

The most intuitive way to noticeably improve the quality of the prediction, which does not require serious modifications, is to use a combination of classical algorithms. Cantos and Almela (2019) analyzed this approach on a corpus containing excerpts from English-as-a-Foreign-Language textbooks. The presented classifier is based on features from Flesch–Kincaid readability test (Kincaid et al. 1975), Coleman–Liau index (Coleman & Liau 1975), Automated readability (ARI) index (Senter & Smith 1967), SMOG grade (McLaughlin 1969) and some other. The precision of the constructed algorithm significantly outperforms separate approaches.

At the same time, significant gains can be achieved using more abstract and complex characteristics. Feng et al. (2010) analyzed the impact of various categories of features on the complexity of the text, such as the number and density of named entities, semantic chains, referential relations, language modeling, syntactic dependencies, and morphology. Ivanov et al. (2018) considered 24 various features. such as average sentence and word lengths, word frequencies, morphological, and syntactic features on the Russian corpus.

The robustness of different features across various corpora with texts of different languages, styles, and genres is also a challenging question. This issue was partly solved by Isaeva and Sorokin (2020), who studied three groups of features, namely, average lengths plus frequencies, morphological, and syntactic ones. The experiments on three corpora of educational texts showed that there is a core of features that are crucial for all texts: the average number of syllables per word, the proportion of verbs in active voice among all words, the proportion of personal

pronouns among all words, and the average syntax tree depth. Deutsch et al. (2020) considered a few combinations of deep learning models with linguistically motivated features in order to determine how much such a combination will improve the quality of predictions.

As in many other areas of natural language processing, state-of-the-art results can be achieved by fine-tuning Bidirectional Encoder Representations from Transformers (BERT) presented by Devlin et al. (2019) and similar models. Martinc et al. (2021) studied unsupervised and supervised approaches, comparing BERT, Hierarchical attention networks, and Bidirectional Long short-term memory networks. The experiments were conducted on a few English and Slovenian corpora. The results suggested that BERT can be used as a high-level baseline for our research.

2.3. Russian datasets

The study of readability for the Russian language is developing more slowly than, for example, for English. This is largely due to the lack of text corpora with the labeled age of readers or readability level. The creation of a corpus of texts readable for students requires a large number of preliminary surveys of respondents of school age. This can be avoided by using the texts of school textbooks. For example, Ivanov et al. (2018) presented Russian Readability Corpus based on the texts of school textbooks on Social Studies for grades 5–11, later expanded by Isaeva and Sorokin (2020). An alternative way would be to use lists of Recommended Reading as done by Iomdin and Morozov (2021). Presumably, many school students read such texts. However, by their nature, such corpora are far from ideal: they contain texts that, in the opinion of adult compilers, should be read at the appropriate age, but they in no way guarantee the fact of reading, much less understanding of these texts. Such lists often include various historical texts that are incomprehensible at the level of an average student without additional explanations. In the above-mentioned list of recommended literature, there are such complex texts from the point of view of the reader as “The Lay of Igor's Campaign” and “The Tale of Bygone Years”[1]. Thus, despite the high labor intensity, a large-scale survey of schoolchildren seems to be the optimal way to select texts for the corpus, which makes it possible to achieve the best representativeness.

3. Datasets, features and models

3.1. Books Read by Students corpus

In real practice, text complexity is often identified with the age of the reader, usually a student. Thus, readability indices often predict a grade of school or college in which the text would be understandable (e.g., Kincaid et al. 1975). This makes it

[1] It is interesting that due to the inability to take into account the obsolescence of the vocabulary, traditional indices assign these texts a very low level of complexity (Iomdin & Morozov 2021).

possible to exclude from consideration adult readers, whose reading experience can be incredibly variable and, as a result, can be estimated only with a very large margin of error. In our study, we decided to use this approximation and assess the age of a reader instead of the abstract complexity level.

To create a corpus with a labeled reading age, two stages of experiments have been conducted. At the first stage, respondents of preschool and school-age (or their parents, acting on their behalf) were asked to name a few of the books they had most recently read, together with their authors. The survey involved 1176 respondents under the age of 16, of which more than 1000 were of school age. In order to complete the list of presumably popular books, a similar additional survey was conducted at the linguistic session at the Sirius Educational Center (Russia), the target audience of the survey was mainly students of grades 9–11.

At the second stage, another group of respondents was asked to select the books they read from the list of those most frequently mentioned during the first stage, 93 books in total. The books were distributed over several subsurveys in such a way that each respondent was asked about no more than two books from the series. The experiment involved 1120 respondents, each of whom answered questions about 15 or 16 books.

Due to the insufficient number of respondents of primary school age and in order to exclude the peculiarities of the individual experience of readers, we decided to combine the ages of students into 5 categories: 1–2, 3–4, 5–7, 8–9, and 10–11 grades. The proportion of participants who read the text was calculated for each book and each age category. After that, the youngest category containing at least 25% of respondents who had read the book was assigned a book label. For example, “In Search of the Castaways” (“Les Enfants du capitaine Grant”) was marked as read by 0% of respondents from the category 1–2, 13% of respondents from the category 3–4, and 70% of respondents in the category 5–7, thus, the category 5–7 was assigned.

As a result, 75 books were included into the *Books Read by Students* corpus **(BR)**. A complete list is given in Appendix A. 18 texts from the second stage of the experiment did not receive any age label and were not included in the corpus. In such a situation, the level of complexity cannot be established even with a sharp increase in the proportion of readers when moving between age categories. A brief overview of the collected corpus is presented in Table 1.

Table 1. **Brief statistics of the Books Read by Students corpus**

Category	All	1–2	3–4	5–7	8–9	10–11
Total texts	75	31	14	8	14	8
Total words	4077579	888984	881578	693448	1060079	553490
Total unique lemmas	50930	24513	24616	24185	30670	24645
Avg sentence length	9.41	8.37	9.58	11.19	9.55	10.44
Avg word length	4.94	4.88	5.02	5.06	4.91	4.83
Avg unique lemmas per text	6864.05	3431.6	6033.6	7375.6	6757.0	7375.6

3.2. Alternative datasets

As mentioned above, for the Russian language, there are few corpora with complexity labels. We decided to compare BR with two of them: *Fiction Previews* **(Fic)** presented by Glazkova et al. (2021) and *Recommended Literature* **(RL),** which we constructed in the previous study from the list of recommended literature for schoolchildren created by the Russian Ministry of Education (Iomdin & Morozov 2021). All collected texts were divided into fragments of 70 sentences each. This allowed us to considerably increase the size of datasets without significant loss of labeling quality (Isaeva & Sorokin 2020).

For English, there are a few corpora with complexity labels; we used three of them. *Common Core State Standards* **(CC)**[2] is a corpus designed to represent text complexity levels for each grade in the USA. *OneStopEnglish* **(OSE)** corpus was specially created for training readability models (Vajjala & Lucic 2018). It consists of 189 text samples, each in three complexity versions. *CommonLit* (**CL**) corpus was presented at a Kaggle competition[3]. The main difference of this corpus from the rest ones is continuous labeling set instead of classes. An overview of the datasets is shown in Table 2.

Table 2. **Some statistics of the datasets.**
BR — Books Read by Students, RL — Recommended Literature, Fic — Fiction Previews, CC — Common Core State Standards, CL — CommonLit, OSE — OneStopEnglish

Characteristics	BR	RL	Fic	CC	CL	OSE
Total texts	5795	9230	58184	219	2834	567
Total categories	5	3	2	6	1	3
Total words	2897003	4888290	26252666	84014	491944	381137
Total unique words	55577	103875	304731	10007	24449	13611
Avg words/text	984.75	1053.28	918.64	450.46	199.65	757.82
Avg words/sentence	13.92	14.95	13.12	22.24	24.94	22.04
Avg sentences/text	70	70	70	23.26	9.46	34.98

3.3. Linguistic Features

According to the related works, we identified seven types of features, which can be used to assess the text complexity: 1) readability indices, e.g., the Flesch–Kincaid readability test and the SMOG grade; 2) traditional features, e.g., the average word length and type/token ratio; 3) morphological feature, e.g., the proportion of nouns and verbs; 4) punctuation, e.g., the number of semicolons; 5) syntactic features, e.g., the average syntactic tree depth and number of clauses; 6) frequencies, e.g., the percentage of tokens included in the list of the most frequent words; and 7) topic modeling. In total, we collected 128 features for English and 126 for Russian of types 1–6. Additionally, we evaluated 100 topics using Latent Dirichlet allocation (Blei et al. 2003). To the best of our knowledge, such a wide

[2] http://www.corestandards.org/
[3] https://www.kaggle.com/c/commonlitreadabilityprize

range of features was considered for the first time in relation to Russian text complexity models. A full list of features can be found in Appendix B.

For evaluation we used the following libraries: readability (Readability 0.3.1 2019), pymorphy2 (Korobov 2015), nltk (Loper & Bird 2002), gensim (Rehurek & Sojka 2010), spacy (Honnibal & Montani 2017), deeppavlov (Burtsev et al. 2018), and API of readability.io. The source code for our methods is available at GitHub (Readability 2021).

3.4. Models

We used the following machine learning algorithms as baselines:

1 Linear Support Vector Classifier (LSVC). LSVC was built with the l2 penalty and the squared hinge loss. We fitted LSVC on bag-of-words (BoW) text representations with a maximum length of 10000. Scikit-learn (Pedregosa et al. 2011) was used for implementation.

2 Random Forest (RF). We used 100 estimators and the Gini impurity to measure the quality of a split. The implementation details are the same as those for LSVC.

3 Feedforward Neural Network (FNN). The hyperparameters used are identified in Table 3. We employed the Adam optimizer (Kingma & Ba 2015). The model was implemented using Keras (Chollet et al. 2015). Each model was trained with early stopping for a maximum of 100 epochs and patience of 20. We utilized Sentence Transformers text representations obtained using the all-mpnet-base-v2 model (Reimers & Gurevych 2019) for the English corpora and the distiluse-base-multilingual-cased model (Reimers & Gurevych 2020) for the Russian ones.

4 Convolutional Neural Network (CNN). The training details are the same as for FNN. The model was implemented using FastText embeddings for English (Mikolov et al. 2018) and Russian (Kutuzov & Kuzmenko 2016).

Table 3. **Hyperparameters for neural baselines**

Hyperparameters	FNN	CNN
Number of convolutional layers	-	2
Number of pooling layers	-	2
Number of convolutional filters	-	256
Filter size	-	256
Number of fully connected layers	3	1
Size of fully connected layers	1024	32
Activation (hidden layers)	Tanh	relu
Activation (output layer)	softmax (classification) linear (regression)	
Dropout	0.5	

We randomly shuffled all the Russian corpora and the CL dataset and split them into train and test sets in the ratio of 3:1. The splitting was conducted in such a way that all fragments of one book belonged either to the train set or to the test

one. Due to the small number of documents in OSE and CC corpora, we used five-fold cross-validation on these datasets to obtain more reliable results. For all of the models above, we systematically evaluated each type of linguistic features applying the Min-Max technique for normalization.

To compare the scores obtained with the results of a few state-of-the-art models, we used BERT-base and RuBERT (Kuratov & Arkhipov 2019) for English and Russian corpora respectively. Each model was fine-tuned for 3 epochs using Transformers (Wolf et al. 2020) with the learning rate of 2e-5 using the AdamW optimizer (Loshchilov & Hutter 2018). We set batch size to 4 and maximum sequence size to 512. To validate our models during the development phase, we divided labelled data using the train and validation split in the ratio 90:10.

We used mean absolute error (MAE) and weighted F1-score to compare the results. MAE is calculated as an arithmetic average of the absolute errors $e_i = y_i - x_i$, where y_i is the prediction, x_i is the true value, n is the number of values:

$$MAE = \frac{\sum_{i=1}^{n} e_i}{n}. \quad (1)$$

The weighted F1-score calculates the standard F1-score for each label, and finds their average, weighted by the number of true instances for each label. The formula of the standard F1-score is:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, \quad (2)$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative predictions respectively.

4. Results and Discussion

We report the results in terms of the MAE (for the CL corpus) and weighted F1-score (for the other corpora) in Table 4. The gray highlight presents the values that outperform the baseline, the values that outperform the baseline by at least 1% are underlined. The best results are shown in bold. Appendix C contains the overall results expressed by several common metrics.

Based on the results, we can identify four performance categories, see Table 5, that describe the impact of various linguistic features **(RQ1)**. In most cases, the considered features improved the model performance. Meanwhile, it was only morphological features that gave a positive impact in most classifiers for all corpora. Readability features exceeded the baseline on most models for most datasets except the BR corpus. Punctuation, traditional, and syntactic features showed a performance growth at least for two models on each corpus. Frequency and topic modeling features produced mixed results. On the one hand, topic modeling features improved the performance of all classifiers on two corpora. Nevertheless, the score on the OSE corpus increased for only RF classifier. This could be because the corpus contains parallel versions of the same papers. Although frequency features improved the performance in some cases, they demonstrated

higher MAE in most classifiers on the CL dataset. Probably, it reflects the fact that short texts normally lack word frequency and context information because of word sparsity (Yan et al. 2013, Xun et al. 2016).

Table 4. **F1 (%) and MAE for each type of features. For F1 the best result is the highest, for MAE — the lowest. BERT refers to the BERT-base in English corpora and to RuBERT in the Russian ones**

Model	BR	RL	Fic	CC	CL	OSE
BERT	45.23	62.74	80.96	42.18	**0.453**	70.99
LSVC	32.31	63.16	76.66	28.22	0.673	70.41
RF	30.94	48.21	78.87	30.03	0.627	68.21
FNN	34.22	63.26	66.34	33.73	0.533	54.00
CNN	39.82	58.12	80.12	33.60	0.593	70.64
+ readability						
LSVC	32.12	63.16	76.67	32.43	0.663	70.49
RF	29.19	49.89	78.45	27.77	0.599	70.11
FNN	40.89	63.62	68.23	37.56	0.502	56.07
CNN	45.90	61.35	80.52	35.89	0.590	68.59
+ traditional						
LSVC	33.15	62.67	77.14	29.3	0.666	69.89
RF	30.03	46.53	78.26	28.57	0.609	**73.01**
FNN	32.12	**69.76**	70.51	34.7	0.482	58.76
CNN	44.32	65.19	80.68	**45.98**	0.604	64.82
+ morphological						
LSVC	32.55	63.22	77.03	31.99	0.662	71.75
RF	30.36	46.63	76.20	29.56	0.611	70.67
FNN	35.63	69.12	72.04	37.42	0.504	62.00
CNN	42.29	68.63	80.75	37.12	0.573	69.02
+ punctuation						
LSVC	32.26	62.87	76.73	30.44	0.664	70.41
RF	30.30	47.25	78.20	28.39	0.629	68.92
FNN	35.21	66.54	68.70	32.51	0.505	55.79
CNN	40.74	67.95	80.86	43.68	0.580	64.33
+ syntactic						
LSVC	32.66	61.91	76.88	29.27	0.674	70.54
RF	28.84	46.70	77.41	33.97	0.617	72.59
FNN	32.10	69.41	68.31	36.48	0.476	56.68
CNN	45.49	65.35	81.01	36.19	0.592	58.71
+ frequency						
LSVC	32.52	63.07	76.84	33.08	0.662	71.34
RF	30.01	45.87	77.76	26.02	0.640	67.63
FNN	31.45	67.46	67.58	35.33	0.729	63.01
CNN	**46.97**	65.08	**81.11**	38.65	0.597	56.38
+ topic modeling						
LSVC	35.36	62.14	76.92	29.97	0.669	67.00
RF	34.09	49.44	77.65	27.15	0.623	66.10
FNN	38.85	62.01	77.30	34.08	0.516	59.46
CNN	43.93	65.78	80.91	41.28	0.588	64.95

Table 5. **Features types with positive impact for N classifiers on each corpus. 1 — readability indices, 2 — traditional features, 3 — morphological features, 4 — punctuation, 5 — syntactic features, 6 — frequencies, 7 — topic modeling.**

Improvement	BR	RL	Fic	CC	CL	OSE
N=4	7	-	-	5	1, 3, 7	-
N=3	3	1, 3	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 6, 7	2, 4, 5	1, 3, 5
N=2	1, 2, 4, 5, 6	2, 4, 5, 6, 7	-	4	-	2, 4, 6
N=1	-	-	-	-	6	7

Tables 6 and 7 illustrate the performance growth as a percentage averaged over all classifiers for Russian and English corpora **(RQ2)**. The averaged results demonstrate that the models trained on Russian texts benefit more from topic modeling and frequency features in comparison with the models trained on English corpora. On the other hand, the results on the CC corpus indicate that this superiority is rather due to text length than language properties. Readability and punctuation features present similar results for both languages. Although morphological, traditional, and syntactic features show better performance on English texts, the results on specific corpora are strongly determined by the source of texts and the type of markup. Thus, any influence of syntactic features for the OSE corpus could not be identified during our experiments. However, there was a noticeable increase for the CC corpus containing fiction texts that characterized English as an analytic language. Overall, these results indicate that the impact of all feature types is mainly attributable to specific circumstances of a corpus. This enables one to use transfer-learning algorithms for cross-lingual analysis of text corpora having similar characteristics.

Table 6. **Averaged performance growth for Russian corpora, %**

Features	BR	RL	Fic	Avg Rus
Readability	7.13	2.4	0.71	3.41
Traditional	1.21	4.54	1.71	2.49
Morphological	2.3	6.04	1.62	3.32
Punctuation	0.75	4.91	0.93	2.2
Syntactic	0.58	4.26	0.63	1.83
Frequency	1.88	3.4	0.48	1.92
Topic modeling	10.87	3.04	4.07	5.99

Table 7. **Averaged performance growth for English corpora, %**

Features	CC	CL	OSE	Avg Eng
Readability	6.39	3.05	0.96	3.47
Traditional	9.67	3.04	-1.33	4.74
Morphological	8.3	3.19	4.51	5.33
Punctuation	7.2	2.06	-1.14	2.71
Syntactic	8.18	3.04	-1.33	3.3
Frequency	5.91	-9.54	-0.76	-1.46

The performance of the models trained on feature combinations per dataset is presented in Table 8. The results are given only for those models the performance

of which was increased by two and more types of features. We enriched the baseline models with the concatenation of all features that showed a positive impact for the relevant models and datasets. The combination of features increased the F1 of RF on the OSE corpus outperforming all the results obtained for this dataset. This result is marked with an asterisk (*). Moreover, FNN trained on feature combinations showed the best result among all the feature-enriched models on the CL corpus. Taken together, the results presented in Table 4 and Table 8 demonstrate that feature-enriched models outperformed BERT on five out of the six corpora (**RQ3**). In some cases, significant increases were obtained, including 7.02% for the RL corpus and 3.8% for the CC corpus. By contrast, the performance of feature-enriched models depends on the features used and data specifics. Simultaneously, in some cases, models trained on feature combination showed a worse result, than those trained on the one type of features.

Table 8. **F1 (%) and MAE for feature combinations**

Model	BR	RL	Fic	CC	CL	OSE
LSVC	34.50	-	78.09	33.12	0.633	71.44
RF	-	49.38	-	-	0.568	76.44*
FNN	40.88	62.99	78.70	39.71	0.466	74.24
CNN	43.85	65.29	81.06	43.58	0.541	-
BERT	45.23	62.74	80.96	42.18	0.453	70.99

5. Conclusion

We have presented the first comparative analysis of the impact of seven types of linguistic features on the performance of text complexity models. We provided the results of large-scale experiments on six text corpora. Each feature type was evaluated in four representative ML models. Our research demonstrated the advantage of some features over others. For example, morphological features improved the performance of our models in almost all cases. At the same time, topic modeling features did not show any positive impact on the corpus containing parallel versions of the same papers. We also identified performance categories based on the scores obtained and estimated the impact of feature combinations. According to our study, the results depend more on the specific characteristics of the dataset than on language. This provides an opportunity for exploring cross-lingual transfer learning and multilingual models for text complexity assessment. Finally, experimental results on five out of the six corpora showed that feature-enriched models can achieve significant improvements in comparison with the state-of-the-art ones. Here, future research may focus on evaluating more complex semantic and narrative features, such as plot characteristics and the features related to named entity analysis, on including BERT-based features, and on explaining text complexity in terms of each feature type.

Acknowledgements

The article was funded by RFBR, project number 19-29-14224.

REFERENCES

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3. 993–1022. https://doi.org/10.1016/B978-0-12-411519-4.00006-9

Burtsev, Mikhail, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva & Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*. 122–127. https://doi.org/10.18653/v1/P18-4021

Cantos, Pascual & Ángela Almela. 2019. Readability indices for the assessment of textbooks: A feasibility study in the context of EFL. *Vigo International Journal of Applied Linguistics* 16. 31–52. https://doi.org/10.35869/VIAL.V0I16.92

Chollet, Francois. 2015. Keras. Github. https://github.com/fchollet/keras (accessed 31.01.2022).

Coleman, Meri & Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60(2). 283.

Dale, Edgar & Jeanne S. Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin* 27. 11–20, 37–54.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186. Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Deutsch, Tovly, Masoud Jasbi & Stuart Shieber. 2020. Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics*. 1–17. https://doi.org/10.18653/v1/2020.bea-1.1

Feng, Lijun, Martin Jansche, Matt Huenerfauth & Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*. 276–284.

Glazkova, Anna, Yury Egorov & Maksim Glazkov. 2021. A comparative study of feature types for age-based text classification. *Analysis of Images, Social Networks and Texts*. 120–134. Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-72610-2_9

Honnibal, Matthew & Ines Montani. 2017. spaCy 2:Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Iomdin, Boris L. & Dmitry A. Morozov. 2021. Who Can Understand "Dunno"? Automatic Assessment of Text Complexity in Children's Literature. *Russian Speech* 5. 55–68. https://doi.org/10.31857/S013161170017239-1

Isaeva, Ulyana & Alexey Sorokin. 2020. Investigating the robustness of reading difficulty models for Russian educational texts. In *AIST 2020: Recent Trends in Analysis of Images, Social Networks and Texts*. 65–77. https://doi.org/10.1007/978-3-030-71214-3_6

Ivanov, Vladimir, Marina Solnyshkina & Valery Solovyev. 2018. Efficiency of text readability features in Russian academic texts, In *Komp'yuternaya Lingvistika I Intellektual'nye Tehnologii*. 284–293.

Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers & Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Naval Technical Training Command Millington TN Research Branch. https://doi.org/10.21236/ada006655

Kingma, Diederik P. & Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Korobov, Mikhail. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In *International Conference on Analysis of Images, Social Networks and Texts*. 320–332. Springer. https://doi.org/10.1007/978-3-319-26123-2_31

Kuratov, Yuri & Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. *Komp'uuternaya Lingvistika i Intellektual'nye Tehnologii*. 333–339.

Kutuzov, Andrey & Elizaveta Kuzmenko. 2016. Web-vectors: A toolkit for building web interfaces for vector semantic models. In *International Conference on Analysis of Images, Social Networks and Texts*. 155–161. Springer. https://doi.org/10.1007/978-3-319-52920-2 15

Leech, Geoffrey, Paul Rayson & Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* Routledge.

Loper, Edward & Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. 63–70.

Loshchilov, Ilya & Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lyashevskaya, Olga & Serge Sharoff. 2009. *The Frequency Dictionary of the Modern Russian Language (Based on the Materials of the Russian National Corpus).* Moscow: Azbukovnik.

Martinc, Matej, Senja Pollak & Marko Robnik-Sikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics* 47. 1–39. https://doi.org/10.1162/coli_a_00398

McLaughlin, G. Harry. 1969. Smog grading – a new readability formula. *Journal of reading* 12(8). 639–646.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch & Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12. 2825–2830.

Rehurek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Reimers, Nils & Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 3982–3992. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Reimers, Nils & Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 4512–4525. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365

Senter, R. J. & E. A. Smith. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*. 1–14.

Solnyshkina, Marina, Vladimir Ivanov & Valery Solovyev. 2018. Readability formula for Russian texts: A modified version. In *Mexican International Conference on Artificial Intelligence*. 132–145. Springer. https://doi.org/10.1007/978-3-030-04497-8_11

Templin, Mildred C. 1957. *Certain Language Skills in Children; Their Development and Interrelationships*. Minneapolis: University of Minnesota Press.

Vajjala, Sowmya & Ivana Lucic. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 297–304. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0535

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xun, Guangxu, Vishrawas Gopalakrishnan, Fenglong Ma, Yaliang Li, Jing Gao & Aidong Zhang. 2016. Topic discovery for short texts using word embeddings. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 1299–1304. IEEE.

Yan, Xiaohui, Jiafeng Guo, Yanyan Lan & Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*. 1445–1456. https://doi.org/10.1145/2488388.2488514

Internet sources

Chapter 699a. Readable language in insurance policies. URL: https://www.cga.ct.gov/current/pub/chap_699a.htm#sec_38a-29 (accessed 29.05.2022).

Readability. 2021. URL: https://github.com/morozowdmitry/readability (accessed 29.05.2022).

Readability 0.3.1. 2019. URL: https://pypi.org/project/readability/ (accessed 29.05.2022).

Appendix A. Books Read by Students corpus

If the original text was published in a language other than English, the translated title is followed by the title in the original language.

Table 9. **Books included into the Books Read by Students corpus**

Category	Title (original title)	Author(s)
1-2	Winnie-the-Pooh	Alan Alexander Milne
1-2	The Wizard of the Emerald City (Волшебник Изумрудного города)	Alexander Volkov
1-2	Urfin Jus and his Wooden Soldiers (Урфин Джус и его деревянные солдаты)	Alexander Volkov
1-2	The Smart Dog Sonya (Умная собачка Соня)	Andrey Usachev
1-2	Grandma and Eight Children in the Forest (Mormor og de åtte ungene i skogen)	Anne-Catharina Vestly
1-2	Eight Children and a Truck (Åtte små, to store og en lastebil)	Anne-Catharina Vestly
1-2	Pippi Longstocking (Pippi Långstrump)	Astrid Lindgren
1-2	Emil of Lönneberga (Emil i Lönneberga)	Astrid Lindgren
1-2	The Lion, the Witch and the Wardrobe	Clive Staples Lewis
1-2	Harry Potter and the Half-Blood Prince	Joanne Rowling

Category	Title (original title)	Author(s)
1-2	Harry Potter and the Chamber of Secrets	Joanne Rowling
1-2	Harry Potter and the Philosopher's Stone	Joanne Rowling
1-2	Alice's Adventures in Wonderland	Lewis Carroll
1-2	Waffle Hearts (Vaffelhjarte)	Maria Parr
1-2	Dunno in Sun City (Незнайка в Солнечном городе)	Nikolay Nosov
1-2	The Adventures of Dunno and his Friends (Приключения Незнайки и его друзей)	Nikolay Nosov
1-2	The Little Witch (Die kleine Hexe)	Otfried Preußler
1-2	Treasure Island	Robert Louis Stevenson
1-2	The Wonderful Adventures of Nils (Nils Holgerssons underbara resa genom Sverige)	Selma Lagerlöf
1-2	Pancakes for Findus (Pannkakstårtan)	Sven Nordqvist
1-2	When Findus was Little and Disappeared (När Findus var liten och försvann)	Sven Nordqvist
1-2	The Mechanical Santa (Tomtemaskinen)	Sven Nordqvist
1-2	Findus and the Fox (Rävjakten)	Sven Nordqvist
1-2	Findus Plants Meatballs (Kackel i grönsakslandet)	Sven Nordqvist
1-2	Findus Goes Fishing (Stackars Pettson)	Sven Nordqvist
1-2	Findus Goes Camping (Pettson tältar)	Sven Nordqvist
1-2	Findus at Christmas (Pettson får julbesök)	Sven Nordqvist
1-2	Findus Moves Out (Findus flyttar ut)	Sven Nordqvist
1-2	The Rooster's Minute (Tuppens minut)	Sven Nordqvist
1-2	Finn Family Moomintroll (Trollkarlens hatt)	Tove Jansson
1-2	The Adventures of Dennis (Денискины рассказы)	Viktor Dragunsky
3-4	Seven Underground Kings (Семь подземных королей)	Alexander Volkov
3-4	Ronia, the Robber's Daughter (Ronja rövardotter)	Astrid Lindgren
3-4	Robinson Crusoe	Daniel Defoe
3-4	Pollyanna	Eleanor H. Porter
3-4	Three Jolly Fellows (Naksitrallid)	Eno Raud
3-4	Harry Potter and the Deathly Hallows	Joanne Rowling
3-4	Harry Potter and the Goblet of Fire	Joanne Rowling
3-4	Harry Potter and the Prisoner of Azkaban	Joanne Rowling
3-4	The Mysterious Island	Jules Verne
3-4	One hundred years ahead (Сто лет тому вперёд)	Kir Bulychev
3-4	The Adventures of Tom Sawyer	Mark Twain
3-4	The Little Water Sprite (Der kleine Wassermann)	Otfried Preußler
3-4	The Little Ghost (Das kleine Gespenst)	Otfried Preußler
3-4	Comet in Moominland (Kometjakten)	Tove Jansson
5-6-7	The Three Musketeers (Les Trois Mousquetaires)	Alexandre Dumas
5-6-7	The Captain's Daughter (Капитанская дочка)	Alexander Pushkin
5-6-7	The Adventure of the Final Problem	Arthur Conan Doyle
5-6-7	The Hound of the Baskervilles	Arthur Conan Doyle
5-6-7	A Study in Scarlet	Arthur Conan Doyle
5-6-7	The Hobbit, or There and Back Again	John Ronald Reuel Tolkien
5-6-7	Harry Potter and the Order of the Phoenix	Joanne Rowling
5-6-7	In Search of the Castaways (Les Enfants du capitaine Grant)	Jules Verne
8-9	The Time Is Always Good (Время всегда хорошее)	Andrey Zhvalevsky and Evgeniya Pasternak

Category	Title (original title)	Author(s)
8-9	Monday Begins on Saturday (Понедельник начинается в субботу)	Arkady and Boris Strugatsky
8-9	The Lost World	Arthur Conan Doyle
8-9	His Last Bow	Arthur Conan Doyle
8-9	The Sign of the Four	Arthur Conan Doyle
8-9	The Adventure of the Empty House	Arthur Conan Doyle
8-9	The Adventure of the Six Napoleons	Arthur Conan Doyle
8-9	Ivanhoe	Walter Scott
8-9	The Two Captains (Два капитана)	Veniamin Kaverin
8-9	The Invisible Man	H.G. Wells
8-9	The Lord of the Rings	John Ronald Reuel Tolkien
8-9	George's Secret Key to the Universe	Lucy Hawking, Stephen Hawking, Christophe Galfard
8-9	The Master and Margarita (Мастер и Маргарита)	Mikhail Bulgakov
8-9	Dandelion Wine	Ray Bradbury
10-11	The Catcher in the Rye	J. D. Salinger
10-11	1984	George Orwell
10-11	Fathers and Sons (Отцы и дети)	Ivan Turgenev
10-11	Brave New World	Aldous Huxley
10-11	Fahrenheit 451	Ray Bradbury
10-11	Lord of the Flies	William Golding
10-11	Crime and Punishment (Преступление и наказание)	Fyodor Dostoevsky
10-11	To Kill a Mockingbird	Harper Lee

Appendix B. Evaluated Features

Readability indices

1 Flesch–Kincaid readability test (Kincaid et al. 1975).
2 Coleman–Liau index (Coleman and Liau 1975).
3 Automated readability (ARI) index (Senter and Smith 1967).
4 SMOG grade (McLaughlin 1969).
5 Dale-Chall index (Dale and Chall 1948).

Traditional features

1 Average and mean sentence length.
2 Average and mean word length.
3 Long words (>4 syllables) proportion.
4 Type/token ratio (TTR) (Templin 1957).
5 NAV: TTR for Nouns only plus TTR for Adjectives only divided by TTR for Verbs only (Solnyshkina et al. 2018).

Morphological features

1 Percentages of lexical categories.
2 Percentage of grammatical cases.
3 Proportion of animated nouns.
4 Proportion of grammatical aspects for verbs.
5 Proportion of grammatical tenses for verbs.
6 Proportion of transitive verbs.

Punctuation

1 Punctuation/token ratio.
2 Semicolons/token ratio.

Syntactic features

Three features were extracted from each of the following characteristics: average, mean, and maximum.

1 Syntactic tree depth.
2 Distance between a node and its descendant.
3 Number of clauses.
4 Number of adverbial clause modifiers.
5 Number of adnominal clauses.
6 Number of clausal complements.
7 Number of open clausal complements.
8 Number of nominal modifiers.
9 Length of nominal modifiers sequence.

Frequencies

For evaluating frequencies of Russian and English words we used dictionaries based on Russian National Corpus (Lyashevskaya & Sharoff 2009) and British National Corpus (Leech et al. 2001) respectively.

1 Average and mean frequency.
2 Proportion of words, which are in the list of the most 100/200/…/1000 popular words, and similar features for nouns, verbs, adverbs, and adjectives separately.

Appendix C. Overall Results

In the tables below we use the following notation: F — F1-score weighted, P — precision weighted, R — recall weighted, MAE — mean absolute error, MSE — mean squared error.

MSE measures the average of the squares of the errors:

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{n}, \quad (3)$$

where Y_i is the vector of true values, $\hat{Y}_i$ is the vector of predicted values.

Russian corpora

Table 10. **Results for the Books Read By Students corpus**

Model	F		P		R	
BERT	45.23		54.06		41.32	
LSVC	32.31		35.74		34.28	
RF	30.94		32.73		37.18	
FNN	34.22		39.06		31.75	
CNN	39.82		57.34		33.66	
	F	**P**	**R**	**F**	**P**	**R**
	+ readability			**+ traditional**		
LSVC	32.12	35.5	34.2	33.15	36.79	35.2

RF	29.19	26.87	36.04	30.03	32.49	36.34
FNN	40.89	61.23	31.83	32.12	44.37	27.08
CNN	45.9	66.18	37.8	44.32	64.88	36.27
	+ morphological			**+ punctuation**		
LSVC	32.55	37.52	36.5	32.26	35.79	34.35
RF	30.36	37.52	36.5	30.3	37.94	36.57
FNN	35.63	42.75	31.68	35.21	39.54	33.05
CNN	42.29	55.72	37.26	40.74	57.25	33.44
	+ syntactic			**+ frequency**		
LSVC	32.66	36.02	34.66	32.52	35.77	34.28
RF	28.84	31.26	34.74	30.01	32.14	35.88
FNN	32.1	40.95	28.46	31.45	37.54	28.39
CNN	45.49	67.47	36.57	46.97	69.57	38.41
	+ topic modeling			**Combined**		
LSVC	35.36	38.63	36.88	34.5	37.36	35.88
RF	34.09	37.74	38.18	-	-	-
FNN	38.85	45.77	35.96	40.88	55.03	35.96
CNN	43.93	62.93	36.65	43.85	62.93	37.18

Table 11. **Results for the Recommended Literature corpus**

Model	**F**		**P**		**R**	
BERT	62.74		65.71		61.86	
LSVC	63.16		63.54		64.98	
RF	48.21		61.8		59.92	
FNN	63.26		79.19		53.76	
CNN	58.12		58.23		58.99	
	F	**P**	**R**	**F**	**P**	**R**
	+ readability			**+ traditional**		
LSVC	63.16	63.22	64.89	62.67	62.83	64.56
RF	49.89	63.88	60.68	46.53	55.2	58.82
FNN	63.62	81.66	52.91	69.76	93.52	56.03
CNN	61.35	66.33	59.49	65.19	66.22	64.64
	+ morphological			**+ punctuation**		
LSVC	63.22	63.11	64.98	62.87	63.07	64.73
RF	46.63	58.54	59.07	47.25	62.9	59.58
FNN	69.12	92.34	55.78	66.54	87.1	54.43
CNN	68.63	72.84	66.58	67.95	71.19	66.33
	+ syntactic			**+ frequency**		
LSVC	61.91	61.88	63.88	63.07	62.93	64.64
RF	46.7	57.58	58.9	45.87	57.89	58.65
FNN	69.41	93.01	55.78	67.46	89.35	54.6
CNN	65.35	69.58	63.21	65.08	66.22	64.64
	+ topic modeling			**Combined**		
LSVC	62.14	62.71	64.22	-	-	-
RF	49.44	65.98	60.68	49.38	62.93	60.34
FNN	62.01	65.98	59.66	62.99	68.99	58.99
CNN	65.78	67.24	64.89	65.29	68.18	63.54

Table 12. **Results for the Fiction Previews corpus**

Model	**F**		**P**		**R**	
BERT	80.96		81.83		80.82	
LSVC	76.66		77.89		76.87	
RF	78.87		79.67		78.99	
FNN	66.34		72.31		65.01	
CNN	80.12		80.87		80.04	
	F	**P**	**R**	**F**	**P**	**R**
	+ readability			**+ traditional**		
LSVC	76.67	77.84	76.87	77.14	78.29	77.34
RF	78.45	78.85	78.51	78.26	78.86	78.36
FNN	68.23	72.54	67.36	70.51	70.61	70.49
CNN	80.52	81.9	80.37	80.68	81.74	80.56
	+ morphological			**+ punctuation**		
LSVC	77.03	78.27	77.24	76.73	77.94	76.94
RF	76.2	77.16	76.38	78.2	78.93	78.32
FNN	72.04	72.09	72.04	68.7	68.75	68.69
CNN	80.75	81.73	80.65	80.86	81.84	80.75
	+ syntactic			**+ frequency**		
LSVC	76.88	78.08	77.09	76.84	78	77.04
RF	77.41	78.21	77.54	77.76	78.4	77.86
FNN	68.31	68.41	68.29	67.58	67.59	67.57
CNN	81.01	81.97	80.9	81.11	82.08	81.01
	+ topic modeling			**Combined**		
LSVC	76.92	78.18	77.12	78.09	79.3	78.27
RF	77.65	78	77.71	-	-	-
FNN	77.3	78.28	77.17	78.7	79.06	78.66
CNN	80.91	82.07	80.78	81.06	82.17	80.93

English corpora

Table 13. **Results for the Common Core State Standards corpus**

Model	**F**		**P**		**R**	
BERT	42.18		64.57		33.77	
LSVC	28.22		30.13		30.61	
RF	30.03		30.38		34.65	
FNN	33.73		37.93		32.9	
CNN	33.6		58.04		26.92	
	F	**P**	**R**	**F**	**P**	**R**
	+ readability			**+ traditional**		
LSVC	32.43	33.55	35.59	29.3	31.37	31.5
RF	27.77	26.95	31.95	28.57	28.68	32.88
FNN	37.56	42.34	36.53	34.7	38.48	34.28
CNN	35.89	56.83	29.25	45.98	78.2	36.12
	+ morphological			**+ punctuation**		
LSVC	31.99	35.29	33.33	30.44	32.07	33.32
RF	29.56	29.53	34.26	28.39	27.23	34.65
FNN	37.42	46.15	34.7	32.51	37.2	32

CNN	37.12	57.32	30.62	43.68	60.51	37.95
	+ syntactic			**+ frequency**		
LSVC	29.27	29.45	31.95	33.08	35.74	34.67
RF	33.97	34.75	38.33	26.02	23.17	31.55
FNN	36.48	41.42	35.64	35.33	40.79	34.27
CNN	36.19	62.18	28.3	38.65	54.04	32.45
	+ topic modeling			**Combined**		
LSVC	29.97	31.38	32.42	33.12	35.21	34.67
RF	27.15	29.19	30.15	-	-	-
FNN	34.08	38.34	32.91	39.71	47.55	37.94
CNN	41.28	65.93	33.85	43.58	44.09	39.44

Table 14. **Results for the CommonLit corpus**

Model	**MAE**		**MSE**	
BERT	0.4532		0.3159	
LSVC	0.6728		0.695	
RF	0.6266		0.6199	
FNN	0.533		0.4421	
CNN	0.5926		0.555	
	MAE	**MSE**	**MAE**	**MSE**
	+ readability		**+ traditional**	
LSVC	0.6627	0.6742	0.6664	0.6819
RF	0.5986	0.5743	0.609	0.5831
FNN	0.5024	0.4045	0.4823	0.3832
CNN	0.5896	0.5496	0.6041	0.5813
	+morphological		**+ punctuation**	
LSVC	0.6621	0.6775	0.664	0.6785
RF	0.6113	0.5917	0.6288	0.6204
FNN	0.5042	0.4002	0.5053	0.4102
CNN	0.5728	0.5269	0.5803	0.5307
	+ syntactic		**+ frequency**	
LSVC	0.6741	0.6924	0.6619	0.6703
RF	0.6167	0.5853	0.6401	0.643
FNN	0.4759	0.3705	0.7293	0.7627
CNN	0.5923	0.5566	0.5973	0.5602
	+ topic modeling		**combined**	
LSVC	0.6686	0.6861	0.6334	0.6166
RF	0.623	0.5986	0.568	0.5174
FNN	0.5156	0.4149	0.4658	0.3542
CNN	0.5882	0.5403	0.5408	0.4726

Table 15. **Results for the OneStopEnglish corpus**

Model	**F**	**P**	**R**
BERT	70.99	78.15	69.34
LSVC	70.41	72.15	72.03
RF	68.21	70.44	69.85
FNN	54	56.34	52.83
CNN	70.64	84.44	65.23

	F	P	R	F	P	R
	+ readability			**+ traditional**		
LSVC	70.49	72.17	72.02	69.89	71.76	71.69
RF	70.11	71.63	71.83	73.01	74.89	74.45
FNN	56.07	59.02	54.59	58.76	62.86	57.18
CNN	68.59	76.29	67.37	64.82	77.32	60.71
	+ morphological			**+ punctuation**		
LSVC	71.75	73.65	73.39	70.41	72.15	72.03
RF	70.67	72.22	72.25	68.92	70.24	70.4
FNN	62	65.37	60.19	55.79	57.56	54.8
CNN	69.02	78.87	66.33	64.33	75.55	60.33
	+ syntactic			**+ frequency**		
LSVC	70.54	72.61	72.37	71.34	73.1	73.04
RF	72.59	73.67	73.82	67.63	68.8	69.89
FNN	56.68	77.87	49.85	63.01	65.63	61.68
CNN	58.71	73.85	54.88	56.38	68.41	53.15
	+ topic modeling			**Combined**		
LSVC	67	68.9	69.14	71.44	72.96	73.07
RF	66.1	68.1	66.45	76.44	77.18	77.37
FNN	59.46	61.84	58.38	74.24	75.71	74.17
CNN	64.95	76.98	62.17	-	-	-

Article history:
Received: 20 October 2021
Accepted: 21 Jenuary 2022

Bionotes:

Dmitry A. MOROZOV is Junior Researcher at the Laboratory of Applied Digital Technologies, International Mathematical Center of Novosibirsk State University, Novosibirsk, Russia, and Developer at the Russian National Corpus. His spheres of interest include corpus linguistics, discrete math, math modeling, and machine learning.
Contact information:
Novosibirsk National Research State University
1 Pirogova, Novosibirsk, 630090, Russia
e-mail: morozowdm@gmail.com
ORCID: 0000-0003-4464-1355

Anna V. GLAZKOVA is Doctor of Sc. (Technology), Associate Professor of the Department of Software at the Institute of Mathematics and Computer Science of Tyumen University, Russia. Her current research interests include natural language processing and text mining, with a focus on text classification and deep learning.
Contact information:
University of Tyumen
6 Volodarsky, Tyumen, 625003, Russia
e-mail: a.v.glazkova@utmn.ru
ORCID: 0000-0001-8409-6457

Boris L. IOMDIN holds a Ph.D. in Philology and is a Leading Researcher at Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia. He is one of the authors and editors of the Active Dictionary of the Russian Language and a co-author of the project "Semantic Complexity of Words".

Contact information:
Vinogradov Russian Language Institute of the Russian Academy of Sciences
18/2 Volkhonka, Moscow, 119019, Russia
e-mail: iomdin@ruslang.ru
ORCID: 0000-0002-1767-5480

Сведения об авторах:
Дмитрий Алексеевич МОРОЗОВ – младший научный сотрудник Лаборатории прикладных цифровых технологий Международного математического центра НГУ, Новосибирск, Россия. Сотрудник Национального корпуса русского языка. Область научных интересов: корпусная лингвистика, дискретная математика, математическое моделирование, машинное обучение.

Контактная информация:
Новосибирский национальный исследовательский государственный университет
Россия, 630090, Новосибирск, ул. Пирогова, д. 1
e-mail: morozowdm@gmail.com
ORCID: 0000-0003-4464-1355

Анна Валерьевна ГЛАЗКОВА – кандидат технических наук, доцент кафедры программного обеспечения Института математики и компьютерных наук Тюменского государственного университета, Тюмень, Россия. Область научных интересов: обработка естественного языка, интеллектуальный анализ текста, классификация текстов, глубокое обучение.

Контактная информация:
Тюменский государственный университет
Россия, 625003, Тюмень, ул. Володарского, д. 6
e-mail: a.v.glazkova@utmn.ru
ORCID: 0000-0001-8409-6457

Борис Леонидович ИОМДИН – кандидат филологических наук, ведущий научный сотрудник Института русского языка им. В.В. Виноградова Российской академии наук, Москва, Россия. Один из авторов и редакторов Активного словаря русского языка, соавтор проекта «Семантическая сложность слов».

Контактная информация:
Институт русского языка им. В. В. Виноградова РАН
Россия, 119019, Москва, ул. Волхонка, д. 18/2
e-mail: iomdin@ruslang.ru
ORCID: 0000-0002-1767-5480

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 449–470
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30140

Research article

Discourse complexity in the light of eye-tracking: a pilot Russian language study

Svetlana TOLDOVA[1] ✉, Natalia SLIOUSSAR[1,2] and Anastasia BONCH-OSMOLOVSKAYA[1]

[1]National Research University Higher School of Economics, Moscow, Russia
[2]Saint Petersburg State University, Sait-Petersburg, Russia
✉toldova@yandex.ru

Abstract
The paper explores the influence of discourse structure on text complexity. We assume that certain types of discourse units are easier to read than others, due to their explicit discourse structure, which makes their informational input more accessible. As a data source, we use the dataset from the MECO corpus, which contains eye movement data for 12 Russian texts read by 35 native speakers. We demonstrate that the approach relying on elementary discourse units (EDUs) can be felicitously used in the analysis of eye movement data, since fixation patterns on EDUs are similar to those on whole sentences. Our analysis has identified EDU outliers, which show shorter time of first fixation than estimated. We arranged these outliers into several groups associated with different discourse structures. First, these are statements with nominal predicates that set exposition of the text or macroproposition and, following those, EDUs that elaborate on the previous statement and signal the beginning of the narrative. Second, they are EDUs that serve as the middle component of a listing or a group of coordinated clauses or phrases. The final group represents EDUs that are part of an opposition, contrast or comparison. Discourse analysis based on EDUs has never been applied to eye movement data, so our project opens many avenues for further research of complexity of discourse structure.
Keywords: *discourse, text complexity, eye movement, EDU, MECO corpus*

For citation:
Toldova, Svetlana, Natalia Slioussar & Anastasia Bonch-Osmolovskaya. 2022. Discourse complexity in the light of eye-tracking: a pilot Russian language study. *Russian Journal of Linguistics* 26 (2). 449–470. https://doi.org/10.22363/2687-0088-30140

Научная статья

Дискурсивная сложность в свете данных о движениях глаз при чтении: пилотное исследование на материале русского языка

С.Ю. ТОЛДОВА[1] ✉, Н.А. СЛЮСАРЬ[1,2], А.А. БОНЧ-ОСМОЛОВСКАЯ[1]

[1]Национально исследовательский университет «Высшая школа экономики», Москва, Россия
[2]Санкт-Петербургский университет, Санкт-Петербург, Россия
✉toldova@yandex.ru

Аннотация
В статье исследуется влияние структуры дискурса на сложность текста. Предполагается, что некоторые типы дискурсивных единиц читаются легче, чем другие, благодаря выраженной дискурсивной структуре, которая делает содержащуюся в них информацию более доступной для обработки. В качестве источника данных мы используем набор данных из корпуса MECO, который содержит данные о движении глаз для 12 русских текстов, прочитанных 35 носителями языка. В статье демонстрируется, что подход, основанный на элементарных единицах дискурса (ЭДЕ), может быть успешно использован для анализа данных о движении глаз, поскольку паттерны фиксации на ЭДЕ схожи с паттернами фиксации на целых предложениях. Проведеный анализ выявил выбросы ЭДЕ, которые показывают более короткое время первой фиксации, чем предполагалось. Они были разделены на несколько групп, связанных с различными структурами дискурса. Во-первых, это высказывания с номинативными предикатами, задающими экспозицию текста или макропропозицию, и следующие за ними ЭДЕ, развивающие предыдущее высказывание и сигнализирующие о начале повествования. Во-вторых, это ЭДЕ, которые служат средним компонентом перечисления или группы согласованных клаузул или фраз. Последняя группа представляет ЭДЕ, которые являются частью оппозиции, контраста или сравнения. Анализ дискурса на основе ЭДЕ никогда не применялся к данным движения глаз, поэтому наш проект открывает новые перспективы для дальнейшего исследования сложности структуры дискурса.
Ключевые слова: *дискурс, сложность текста, движение глаз, ЭДЕ, корпус MECO*

Для цитирования:
Toldova S.Yu., Slioussar N.A., Bonch-Osmolovskaya A.A. Discourse complexity in the light of eye-tracking: a pilot Russian language study. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 449–470. https://doi.org/10.22363/2687-0088-30140

1. Introduction

The paper explores the influence of discourse structure on text complexity. We assume that certain types of discourse units are easier to read than others, due to their explicit discourse structure, which makes their informational input more accessible. As a data source, we rely on the Multilingual Eye-movement Corpus, or MECO (Kuperman et al. 2022a, b). The first release of the corpus contains eye movement data from speakers of 12 languages reading 12 texts in their native language and 12 texts in English, as well as answering comprehension questions and passing several tests. We use a dataset with 12 Russian texts read by 35 native speakers.

Registering eye-movements as they unfold in real time, or eye-tracking, was shown to be a very precise and ecologically valid method to study reading (e.g. Rayner 1998, Rayner et al. 2012). However, so far there are not so many studies that analyze the influence of discourse structure on reading behavior. In the present paper, the discourse analysis method based on identifying elementary discourse units (EDUs) (Grosz & Sidner 1986, Mann & Thompson 1987, Polanyi 1988) is applied to eye-tracking data. Firstly, we demonstrate that using this method is very effective: fixation patterns on EDUs are similar to those on whole sentences. Secondly, we identify EDUs that are read significantly faster than expected (based on the estimates taking such parameters as word length into account). Then we analyze them qualitatively: we show that they form several groups associated with different discourse structures.

The first group includes statements with nominal predicates that set the exposition of the text or a macroproposition and, following those, EDUs that elaborate on the previous statement and signal the beginning of the narrative. The second group contains EDUs that serve as the middle component of a listing or a group of coordinated clauses or phrases. The third group includes EDUs that are part of an opposition, contrast or comparison. The main goal of our project is exploratory. We envision it as a pilot study that opens up many avenues for further research in the field of discourse structure complexity.

2. Background

2.1. Eye tracking studies

Let us start with several basic facts about human vision. We have high visual acuity only in the very center of the visual field. This area is called the fovea. Therefore, when we are reading, our eyes fixate on a word for a fraction of a second and then quickly move to the next word. During these movements, or saccades, no visual information is processed – this happens only during fixations. Some words may require more than one fixation, especially longer and less frequent ones, while the others may be skipped altogether. Short functional words are skipped regularly. About 10 – 15% of the saccades are regressive (Rayner 1998) – we return to what we have just read and then move forward again.

Eye trackers record this complex pattern of fixations and saccades, and provide the researcher with many measures potentially reflecting different processing stages. These measures are usually defined at the word level: *skipping* (whether the word was fixated at least once or skipped); *first fixation duration* (the duration of the first fixation landing on the word); *gaze duration* (the summed duration of fixations on the word in the first pass, i.e., before the gaze leaves this word for the first time); *total fixation duration* (the summed duration of all fixations on the word); *number of fixations* on the word; *regression* (whether the gaze returned to the word after inspecting further text), etc. A detailed discussion of these and other measures can be found in (Boston et al. 2008, Clifton et al. 2007).

Eye tracking is widely used in psycholinguistics and other cognitive sciences to study different phenomena from low-level reading and viewing strategies to the most complex processes like decision-making. However, most studies are experimental: in psycholinguistics, they usually focus on the properties of preselected target words, presented in isolation or inside artificially constructed sentences and, sometimes, small texts. Recently, a number of eye-tracking corpora have been created for several languages, including English (Frank et al. 2013, Luke & Christianson 2018), German (Kliegl et al. 2006), Hindi (Husain et al. 2014), and Russian (Laurinavichyute et al. 2019). The Provo corpus (Luke & Christianson 2018) includes short text passages, while all other corpora mentioned above rely on individual sentences. There are also several corpora based on two languages, including the Dundee corpus (Pynte & Kennedy 2006) with texts in English and French and the GECO corpus (Cop et al. 2017) based on a novel by Agatha Christie (the English original and a Dutch translation). The motivation to create the MECO corpus (Kuperman et al. 2022a, b) used in the present study was to provide comparable data for a much larger set of languages and to do so using complete coherent short texts rather than sentences or text fragments. Among other things, this gives us a unique opportunity to study various text-level phenomena.

First of all, eye movement corpora have been instrumental in establishing the fundamental characteristics of eye movements within and across languages, the so called eye movement benchmarks. Three main parameters of words that influence eye movements were identified: frequency, length, and predictability. Other properties, like the age of acquisition of the word, the number of meanings or the morphological structure, may also play a role, but to a lesser extent (Clifton et al. 2007).

Studies of sentence-level phenomena mostly focused on such topics as syntactic ambiguity processing, which play a crucial role in developing syntactic parsing models. However, some basic generalizations have also been established: the first word of the sentence tends to have longer reading times, then the reader speeds up and slows down again at the last word (Just & Carpenter 1980). This final slowdown is associated with the so-called wrap up when the reader integrates all information presented in the sentence.

2.2. Discourse studies

It has been shown by many researchers (e.g. Hasan & Halliday 1976, van Dijk 2019, Givon 1993) that various phenomena, like anaphora or connectives, cannot be described within an isolated sentence. One can easily distinguish a random sequence of sentences from a coherent text. Thus, it is assumed that discourse is organized in a kind of structure. There are different approaches to representing this structure, cf. the connective-based approach to the relation between discourse segments (Prasad et al. 2018) and the semantic-based approach within the Rhetorical Structure Theory, or RST (Mann & Tompson 1987).

Our research is based on the RST. Its basic assumption is that discourse is a set of nested discourse units up to elementary ones. Each discourse unit has to be related to another one. A set of relation types varies through different research groups. The basic set of relations was worked out by Mann and Tompson (1987). It resembles a set of clause types within a complex sentence, though it is bigger. The relations can be symmetric ('Join', 'Sequence') or asymmetric ('Cause-Effect', 'Purpose', etc.).

Consequently, a coherent text can be split into elementary discourse units, or EDUs (Grosz & Sidner 1986, Mann & Thompson 1987, Polanyi 1988). There are various approaches to EDU splitting depending on whether spoken or written discourse is analyzed, or whether prosodic, cognitive, semantic or pragmatic criteria for discourse segmentation are taken into account. Some approaches combine different dimensions for segmentation, e.g. prosodic and syntactic dimensions (Degand & Simon 2005), or semantic and prosodic dimensions (Kibrik et al. 2009). In the majority of cases, a discourse unit corresponds to a clause, which can be finite or non-finite. Semantically it denotes an event or a state of affairs. In addition to that, there are units larger or smaller than a clause (see Kibrik et al. 2009).

As we are dealing with written texts, we consider clauses as elementary discourse units, and not prosodic units, as in (Hirschberg & Litman 1993, Chafe 1994, Kibrik & Podlesskaya 2003). Structures smaller than a finite clause, such as nominalized constructions or infinitival clauses, can also be treated as EDUs (Carlson & Marcu 2001, Schauer 2000). For example, a preposition introducing a noun phrase can signal causal relations between its dependent expressed via nominalization and the rest of the clause, as in the following case: ***iz-za*** *Petinogo pozdnego vozvrascheniya* 'due to Petya's late return'. At the same time, some EDUs can consist of two clauses. These are EDUs including sentential arguments and restrictive relative clauses. Appositive relatives are treated as separate EDUs.

2.3. Eye tracking studies of discourse-level phenomena

The majority of eye-tracking studies of linguistic complexity are limited to within-sentence phenomena. Significantly fewer studies deal with discourse phenomena, though discourse coherence can influence sentence comprehension and hence reading performance.

One of the research questions is whether there is a difference in the reading behavior inside a discourse segment (a sentence, a paragraph, an intonational unit or a clause) and at a segment boundary. A great number of works focus on the so-called wrap-up effect briefly mentioned in the previous section (cf. Balogh et al. 1998, Hirotani et al. 2006, Warren et al. 2009, among many others). The main claim of these studies is that clause or sentence final words are read slower than identical words within a clause.

Many works also investigate the start-up effect in the beginning of the clause and the general reading time dependence on the word position in a segment (e.g. Kuperman et al. 2010). In particular, it was found that sentence-initial words tend

to be processed slower (e.g. Gernsbacher 1990). Several experiments report readers' tendency to speed up as they proceed through a sentence (Aaronson & Ferres 1983, Aaronson & Scarborough 1976, Chang 1980, Ferreira & Henderson 1993). Another question is what types of segments (sentences or clauses) trigger these effects. It is also important whether the presence or absence of a comma can affect words reading parameters.

3. Data

3.1. The dataset of eye movements and the procedure used to collect the data

The dataset of eye movements used in the present study comes from the Multilingual Eye-movement Corpus, or MECO (Kuperman et al. 2022a, b). The goal of the MECO project was to collect comparable cross-linguistic eye-tracking data on reading. Native speakers of different languages who learnt English as their second language were recruited to read 12 short texts in their native language (L1) and 12 texts in English (L2). Participants whose native language was English read all 24 texts in their native language and served as the control group in some of the comparisons. After each text, there were two 4-alternative-forced-choice comprehension questions tapping into factual knowledge and inferencing.

The first release of the MECO corpus includes data from 12 languages that differ typologically and orthographically and belong to different linguistic families: English, German, Dutch, Norwegian, Italian, Spanish, Russian, Greek, Turkish, Finnish, Hebrew and Korean. As a result, reading in different L1 could be compared, as well as the influence of different L1 on reading in English as L2. In addition to that, all participants filled in an abridged version of the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian et al. 2007) and passed several tests in their L1 and in English, assessing vocabulary size, word and pseudoword naming, phonological/morphological awareness, and other component skills of reading. The goal was to evaluate how different skills influence reading in L1 and L2. Notably, only participants with an intermediate or advanced level of English as L2 were invited to take part in the study. In other words, the MECO project did not aim to assess reading in L2 at the stages when the learners could not read fluently.

Materials used in the MECO project were encyclopedia-style texts on a variety of topics including historical figures, events, and natural or social phenomena. Firstly, 24 English texts were created. They were loosely based on Wikipedia entries and contained 6–12 sentences (107–185 words). 12 texts were selected for the L2 part of the project, while the other 12 texts were used in the L1 part. Out of the latter 12 texts, five were translated into 11 languages. For the other 7, original texts on the same topic and in the same genre were created in 11 languages. In the present study, we use eye movement data for 12 Russian texts (101 sentences, 1831 words in total).

For each of the 12 languages, at least 45 participants were recruited. We analyze data from 35 native Russian speakers (25 female and 10 male, 20–31 years old). All participants signed an informed consent form before taking part in the study.

Eye-movements were recorded with an EyeLink 1000+ eye-tracker (SR Research, Kanata, Ontario, Canada) with a sampling rate of 1000 Hz. Each text appeared on a separate screen. Consolas font (16 points) was used. Participants were asked to read the texts silently and to press the space bar when they were ready to answer comprehension questions.

3.2. Discourse properties of the texts in the dataset: genre and discourse segmentation

The texts in the MECO corpus are loosely based on Wikipedia entries. They have the same genre and overall structure, but are shorter: every text consists of approximately 10 sentences. The majority of them start with the introduction of a new entity or notion, providing a definition or a basic description of it. Then some narration about these entities follows. All these texts have clear-cut topics repeated throughout the text. Besides, as encyclopedic texts, they include join and elaboration rhetoric relations, other kinds of enumerations, and comparison and contrast relations.

For the subsequent analysis, we split the texts into elementary discourse units (EDUs). As a result, 283 EDUs have been identified in our dataset. As we noted in section 2.2, there are different approaches to this procedure. We used the rules adopted from Ru-RSTreebank with some revisions. The instruction is worked out for written texts. According to this approach, there are EDUs smaller than a clause, while some EDUs (relative and argument clauses with their matrix clauses) are larger than a clause and can contain a comma.

Besides, we introduced several rule revisions. According to the current work on the Russian spoken discourse, coordinate noun phrase constructions (enumerations, like *красные фрукты* '*red fruits', некоторые овощи и даже ягоды 'some vegetables and even berries'*) often have intonational phrase boundaries in speech (pauses, specific intonation patterns) after each member of the list. The texts in our dataset are not read out loud, and we have no opportunity to judge whether to identify a separate EDU in each particular case relying on their prosodic features. Therefore, we decided to split all NP lists into separate discourse units in our set: for example, (1) *красные фрукты* '*red fruits',* (2) *некоторые овощи 'some vegetables'* (3) *и даже ягоды 'and even berries'.*

To sum up, there are different types of enumerations in our dataset. Firstly, there are coordinate clauses (e.g *[Он спросил,] [и она ответила] [He asked,] [and she answered]*) and coordinate clauses with the coordinate subject deletion (e.g. *[Он пришел] [и сказал] '[He came] [and said]'*). Secondly, there are NP lists. In the former case, the EDUs are in the multinuclear 'Join' or 'Sequence' relations. In the latter case, they are in the 'Enumeration' or 'Specification' relations.

4. Experiments looking for correlations between discourse unit characteristics and eye tracking parameters

4.1. EDU boundaries

The first question that we tested was whether the eye-tracking data for reading coherent texts provide evidence for the relevance of units that are smaller than sentences. In other words, we aimed to check whether elementary discourse units singled out according to semantic and structural criteria differ in terms of the reading patterns. As was mentioned in the section 2.3, there is a tendency to read the first word in a sentence slower than the following words. Therefore, we tested the hypothesis that the first words in EDUs are read slower than others.

According to some eye-tracking studies (e.g. Hirotani et al. 2006), commas influence eye tracking parameters in a significant way. Thus, the effect of EDU boundaries, if we find it, may be a result of a strong correlation between the end of the EDU and the comma following it. Indeed, many intra-sentential EDUs are separated by commas in Russian. Therefore, we also checked that the effect we found is due to EDU boundaries and not to punctuation marks.

4.2. The first-word effect

4.2.1. Data and procedure

As we demonstrated in section 2.1, eye-tracking research provides multiple measures that may be associated with different stages of text processing. In our pilot study, we confine ourselves to two parameters that are often selected as reflecting very early and advanced processing stages: the *first fixation duration* (the duration of the first fixation landing on the word) and the *total fixation duration*, or *total time* (the summed duration of all fixations on the word, including possible multiple fixations during the first pass and refixations following regressions if there were any).

Eye-tracking research usually focuses on the properties of separate words rather than larger units as a whole. To study the latter, we transformed selected measures to take into account the crucial factor of word length. Namely, we analyzed the relative first fixation duration (RFFD) and the relative total fixation duration (RTFD): the average first fixation duration per symbol and the average total fixation duration per symbol calculated for every word (token) in our texts. The longer the duration the slower the reading speed. These measures were used in different analyses that we conducted.

To test for start-up effects, we compared the first words in EDUs to the third words (because the start-up effect may cover not only the very first word, but also the second word in the unit). For this analysis, we filtered out EDUs that are shorter than four words or have no fixations on the first or on the third words. We also did not include sentences consisting of a single EDU. Then we calculated an average RFFD and RTFD for both positions for every participant. Using these average

values, we performed a paired two-sided t-test of the null hypothesis of independence for average RFFDs and RTFDs.

4.2.2. Results and discussion

The t-test revealed a statistically significant difference in RFFD and RTFD between the first and the third word in the EDU ($t = 8.59$, $df = 32$, $p < 0.001$ for RFFD; $t = 4.21$, $df = 32$, $p < 0.001$ for RTFD). Average RFFDs for the words in different positions are also presented in Figure 1. The thin gray lines are for separate EDUs, the black dots represent an average RFFD for a position. The blue lines are model predictions for EDUs. We can see that there is a tendency for decreasing the relative first fixation duration while moving further away from the first word in the EDU. In addition to that, the t-test comparing the first word RFFD characteristics in an EDU vs. in a sentence revealed no significant differences between sentences and EDUs ($t = 1.16$, $df = 51$, $p = 0.27$).

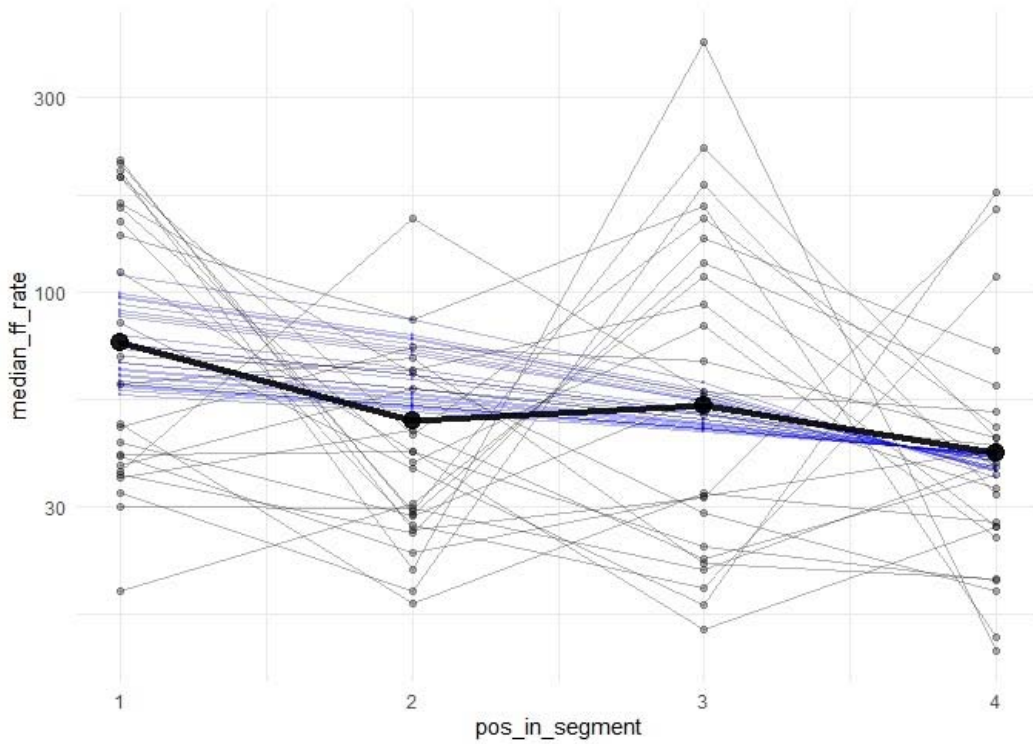

Fig. 1. The average RFFD plot for different word positions in the EDU

To conclude, our analysis provides additional evidence for the effect of the first word in a discourse segment on Russian real-text data. Though the particular patterns of fixation duration may vary greatly for different speakers and different EDUs (cf. gray lines in Fig. 1), the initial fixation on the first word in a segment is longer than on the following words. Moreover, this difference is significant

irrespective of the discourse segment level: a sentence vs. an EDU. We can conclude that for this effect, the EDU boundaries matter. Thus, these results confirm the role of EDU boundaries for reading a coherent text.

4.3. EDU boundaries vs. punctuation marks

An alternative hypothesis is that the first-word effect is due not to the EDU boundaries, but to the punctuation marks. To test this hypothesis, we compared RFFDs for the first unskipped word in a segment under different conditions: (1) the first word of an EDU after a comma; (2) the first word after a comma inside an EDU; (3) the first word of an EDU not following comma, (4) the first word after a dot.

4.3.1. Data and procedure

Firstly, we checked the hypothesis of the independence of first word RFFD from the comma position (inside an EDU vs. before an EDU). Secondly, we checked whether there is a difference between EDUs after a comma and after another EDUs without a comma. To do so, we used linear mixed effect models (LMEs) in the R package *lme4* (Bates et al. 2015). Participants and words were treated as random effects. Finally, we performed a pairwise comparison of the four conditions enumerated above using the Tukey test. For the first analysis we selected EDUs after a comma and EDUs with a comma inside and detected the first unskipped word after the comma.

4.3.2. Results

The results for the two LME models are presented in Figures 2 and 3. Figure 2 confirms that the RFFD on a word after a comma is significantly longer when it is the first word in an EDU than when it is in a middle position. Figure 3 shows that there are no significant differences for the first words in an EDU preceded or not preceded by a comma.

We can conclude that EDU boundaries play a more important role for the RFFD than punctuation marks. Finally, we performed a pairwise comparison of all the four conditions (after a comma in the middle of an EDU, after a comma in the beginning of an EDU, after a dot in the beginning of an EDU, in the beginning of an EDU without any punctuation marks) using the Tukey method. The results are provided in Figure 4.

As Figure 4 shows, there is a difference in RFFD depending on the position of the word inside an EDU (in the beginning vs. in the middle). The dot vs. comma is a significant factor, but there is no statistically significant difference between the word in the beginning of an EDU preceded or not preceded by a comma. To sum up, our data shows the impact of EDU boundaries on reading parameters.

		firstfix_rate	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	35.20	30.61 – 39.79	**<0.001**
segment startTRUE	5.51	2.89 – 8.13	**<0.001**
Random Effects			
σ^2	511.18		
$\tau_{00\ words.word}$	1002.53		
$\tau_{00\ words.subid}$	32.53		
ICC	0.67		
$N_{\ words.subid}$	33		
$N_{\ words.word}$	317		
Observations	5627		
Marginal R^2 / Conditional R^2	0.004 / 0.671		

Fig. 2. The LME model for a word after a comma in the beginning vs. in the middle of an EDU

		firstfix_rate	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	39.58	34.63 – 44.52	**<0.001**
start after commaTRUE	0.65	-2.82 – 4.13	0.712
Random Effects			
σ^2	847.71		
$\tau_{00\ words.word}$	1045.55		
$\tau_{00\ words.subid}$	44.64		
ICC	0.56		
$N_{\ words.subid}$	33		
$N_{\ words.word}$	336		
Observations	6104		
Marginal R^2 / Conditional R^2	0.000 / 0.563		

Fig. 3. The LME model for the first word in an EDU preceded or not preceded by a comma

Conditions	Estimate	SE	z	p
start_after_comma – comma_in_the_middle == 0	6.12	1.47	4.16	< 0.001***
start_after_dot – comma_in_the_middle == 0	12.87	2.12	6.07	< 0.001***
start_after_neither – comma_in_the_middle == 0	4.98	1.82	2.74	0.030*
start_after_dot – start_after_comma == 0	6.75	1.85	3.65	0.001**
start_after_neither – start_after_comma == 0	-1.15	1.45	-0.79	0.852
start_after_neither – start_after_dot == 0	-7.90	1.59	-4.98	< 0.001***

Fig. 4. Pairwise comparisons of the four conditions using Tukey method

4.4. Outliers as a window onto discourse effects

4.4.1. Finding outliers: sentences vs. EDUs

It is clear from previous studies that the variation in reading times is great both among participants and among sentences. As there could be a lot of low-level factors and discourse factors influencing the resulting average total fixation duration and other measures (different types of conjunctions, discourse topicality of a noun phrase, different types of rhetorical relations, the position of an EDU in the discourse tree etc.), we started with trying to find outliers in our dataset. Firstly, we performed an analysis to find the sentences that are read too slowly or too fast as compared to the average reading rate. Then we performed the same analysis to find outliers for EDUs.

To start with, we got the histograms and violin plots for sentences and EDUs. We calculated the average sentence duration rate, or ASDR, for all the sentences: the sum of total fixation durations of all unskipped words divided by the sentence length in symbols. Then, we used the standard deviation from the median ASDR per participant to calculate the deviation of the rate for every sentence.

We used the paired t-test with Bonferroni correction for multiple comparisons and identified 35 sentences for which the ASDR differs significantly from the estimated value ($p<0.05$). We also singled out 82 EDU outliers among 283 using the same method. Out of these EDUs, the majority (59) had *shorter* ASDR than expected. We look at this group below because it is large enough to observe some general tendencies. In order to formulate hypotheses for further research and to single out features that can subsequently be used for EDU classification, we started with a qualitative analysis of these outliers in the present paper.

4.4.2. Qualitative analysis of EDU outliers

In our qualitative analysis we aimed to identify some recurring features in the discourse units that tend to be read with shorter total fixation durations. We analyzed 59 outlier EDUs and found that 47 of them belonged to three groups characterized by common semantic and syntactic properties and specific discourse

functions. These three groups may be loosely called “starters”, “contrasts” and “listings”.

In general, what they all have in common is an explicit discourse structure. This structure sets certain relations between the units of expressed information within the EDU and embeds it in a given way relative to the general representation of textual information. In other words, it is easy for the reader to immediately answer the question of what is being said here: the EDUs that we have classified as starters assert that a certain object belongs to a class of objects; the EDUs belonging to the contrast group introduce the opposite poles of a certain scale; the listing group includes EDUs that report different manifestations of the same property or situation. Looking at these results somewhat more broadly, we can conclude that the decrease in fixation times is somehow related to the idea of predictability. Predictability, which manifests itself here at a higher level, has been proven to have a significant effect on fixation patterns (e.g. Clifton et al. 2007).

Below we will examine each group separately. In addition to analyzing ASDR outliers, we will also pay attention to EDUs that have been completely skipped by three or more readers and have a structure similar to one of the groups. We will use the following notation to denote EDUs and the differences in fixation patterns: [...] is used to define the boundaries of EDUs, bold font represents EDUs characterized by shorter ASDR than estimated, crossed-out EDUs are those skipped by more than 3 readers.

4.4.3. Starters

The texts of the MECO corpus have a similar composition, as they mostly describe natural, cultural or social phenomena. Loosely based on Wikipedia, they obey a common pattern, which is especially noticeable in the first sentences of these texts. 10 out of 12 first sentences have a similar syntactic structure: one or two joint EDUs consisting of a nominal predication with omitted copula, following the Russian grammar rules for the present tense. (1)–(3) show some examples of this type.

(1) ***[Янус – бог всех начинаний] [и переходов в древнеримской религии и мифологии.]***
[Yanus – bog vsekh nachinanij] [i perekhodov v drevnerimskoj religii i mifologii.]
‘Janus is the god of all beginnings and transitions in the ancient Roman religion and mythology.’

(2) ***[Дегустация – это сенсорный анализ и оценка вина.]***
[Degustaciya – eto sensornyj analiz i ocenka vina.]
‘Tasting is a sensory analysis and evaluation of wine.’

(3) *[Апельсиновый сок – это напиток, который получают из плодов апельсинового дерева.]*
[Apel'sinovyj sok – eto napitok, kotoryj poluchayut iz plodov apel'sinovogo dereva.]
‘Orange juice is a drink made from the fruit of the orange tree.’

One example (4) begins with an EDU clause with a light verb, and we have an example of a text (5) that begins with an EDU with verbal predication.

(4) *[В профессиональном спорте допингом называют применение спортсменами запрещенных стимулирующих веществ.]*
[V professional'nom sporte dopingom nazyvayut primenenie sportsmenami zapreshchennyh stimuliruyushchih veshchestv.]
'In professional sports, doping refers to the use of banned performance-enhancing substances by athletes.'

(5) ***[Всемирный день окружающей среды отмечают ежегодно пятого июня.]***
[Vsemirnyj den' okruzhayushchej sredy otmechayut ezhegodno pyatogo iyunya.]
'The World Environment Day is celebrated annually on June 5.'

At the semantic level, 11 out of 12 sentences establish taxonomic relations by introducing the main topic of the text as belonging to a particular category. But only 6 sentences out of 12 (including sentence (5), which is semantically and syntactically different) are read with a significantly shorter ASDR. The other six share a common syntactic feature – they have a restrictive relative clause that is not split up into a separate EDU according to the RST marking rules. We can speculate that it is the restrictive relative clause that determines the longer fixation times. The tendency for restrictive relatives to increase reading times was previously noted in (Hirotani et al. 2006) with the following explanation: restrictive relative clauses belong to the assertive part of the sentence and contribute to the truth conditions of the sentence, they are part of focused material that is known to require more attention. The only exception to this observation in our corpus is (6), in which the restrictive nature of the relative clause is controversial.

(6) ***[Жест «шака» – это знак дружеских намерений, который часто связывают с Гавайями и сёрф-культурой.]***
[Zhest «shaka» – eto znak druzheskih namerenij, kotoryj chasto svyazyvayut s Gavajyami i syorf- kul'turoj.]
'The "shaka" gesture is a sign of friendly intentions that is often associated with Hawaii and surf culture.'

Another type of EDUs that can be assigned to the starter category are EDU clauses that introduce macropropositions in the text in the sense of Van Dijk (2019). Here we observe two groups of cases. Firstly, there are clauses that introduce a new block of information the subject of which repeats the topical subject of the first sentence of the text. Secondly, they may be EDUs that follow the initial thematic sentence and serve as the first introductory piece of the narrative, elaborating on the points declared earlier. The beginning of the narrative may be marked by a tense change: from the present tense of the opening sentence to the narrative past tense.

4.4.4. Contrasts

This category includes pairs of EDUs that express some sort of opposition.

(7) *[Начиная с древних гонок на колесницах]* ***[и до недавних скандалов в бейсболе и велоспорте]***
[Nachinaya s drevnih gonok na kolesnicah] ***[i do nedavnih skandalov v bejsbole i velosporte]***
'From the ancient chariot races to the recent scandals in baseball and cycling…'

Shorter fixations are characteristic either for both EDUs or only for the second element of the comparison. In some cases, as in (8), no specific fixation effects are statistically significant, but we observe skipping of the entire EDU by a certain number of readers.

(8) *[что более дорогое вино будет обладать лучшими характеристиками,] [~~чем менее дорогое~~.]*
[chto bolee dorogoe vino budet obladat' luchshimi harakteristikami,] [~~chem menee dorogoe~~.]
'…that a more expensive wine will have better characteristics than a less expensive one.'

We can conjecture that the shorter fixation effect is related to the lexical parallelism in these EDUs, such as antonymy.

(9) ***[Ворота его храма были открыты во время войны] [и закрыты в мирное время.]***
[Vorota ego hrama byli otkryty vo vremya vojny] [i zakryty v mirnoe vremya.]
'The gates of his temple were open during the war and closed during peacetime.'

Moreover, in some cases the contrast is not expressed at the semantic level of the whole sentence, but only at the level of the individual lexemes, which form a kind of binary opposition.

(10) *[допинг – явление не новое,]* ***[а такое же древнее, как и сам спорт.]***
[doping – yavlenie ne novoe,] ***[a takoe zhe drevnee, kak i sam sport.]***
'…doping is not a new phenomenon, but is as old as sport itself.'

(11) ***[так как он смотрит и в будущее,]*** *[и в прошлое]*
[tak kak on smotrit i v budushchee,] *[i v proshloe.]*
'as he looks both to the future and to the past'

In addition, there are several cases in which we observe shorter fixations on a single EDU containing a lexical opposition represented by two opposite parameters or situations.

(12) ***[нет ни одного флага, у которого высота была бы больше ширины.]***

[net ni odnogo flaga, u kotorogo vysota byla by bol'she shiriny.]
'there is not a single flag that has a height greater than its width'

(13) ***[также связывали с входом и выходом из дома.]***
[takzhe svyazyvali s vhodom i vyhodom iz doma.]
'also associated with entering and exiting the house'

We can assume that the binary scale manifested in the lexical structure supports the processing of EDUs. In any case, we do not see specific effects of this kind within the EDUs in which the objects of comparison cannot be contrasted by a unique parameter, such as the presence or absence of a quality or situation, see (14) and (15).

(14) *[и включила в свои проекты не только сохранение природы,] [но и вопросы экологически безопасного развития.]*
[i vklyuchila v svoi proekty ne tol'ko sohranenie prirody,] [no i voprosy ekologicheski bezopasnogo razvitiya.]
'…and incorporated into its projects not only the preservation of nature, but also issues of environmentally safe development'

(15) *[В некоторых странах номера регистрируются в едином реестре,] [в других реестры ведутся в отдельных штатах и областях.]*
[V nekotoryh stranah nomera registriruyutsya v edinom reestre,] [v drugih reestry vedutsya v otdel'nyh shtatah ili oblastyah.]
'In some countries numbers are registered in a single registry, in others registries are maintained in individual states or provinces.'

Thus, we see in this group a compact discursive structure supported both at the syntactic level (by conjunctions) and at the lexical level. All these means contribute to building up the reader's expectations, which are expressed in the acceleration of information processing, especially in the second part of the contrast.

4.4.5. Listings

Finally, the last category has proven to be the most numerous, as it includes groups of three or more EDUs that together form an enumeration. As we mentioned earlier, enumeration elements are defined as separate EDUs in our analysis. So far, we have distinguished three components in the enumeration list and, accordingly, three EDU types: the beginning, the middle, and the end. We observe that the middle component (or one of the middle components) tends to require less fixation time or is even skipped, as examples (16) and (17) illustrate.

(16) *[Янус олицетворял золотую середину между варварством и цивилизацией,]* ***[деревней и городом,]*** *[юностью и зрелостью.]*
[Yanus olicetvoryal zolotuyu seredinu mezhdu varvarstvom i civilizaciej,] ***[derevnej i gorodom,]*** *[yunost'yu i zrelost'yu.]*
'Janus represented the golden mean between barbarism and civilization, village and city, youth and maturity.'

(17) *[например, его географическое происхождение,] [~~репутация~~] [~~и прочие характеристики~~.]*
[naprimer, ego geograficheskoe proiskhozhdenie,] [~~reputaciya~~] [~~i prochie harakteristiki~~.]
'such as its geographic origin, reputation, and other characteristics'

It can be assumed that in this case the discursive structure is graphically supported. It is interesting to note that, despite the fact that the comma generally slows down processing, in this case the placement of an EDU within two commas may be perceived by the reader as a way to save processing efforts by using the same cognitive schemas as in the preceding EDU.

We also found cases in which we have fixation acceleration on the first element of a list, but these lists are characterized by the fact that the preceding EDU contains a generalizing lexeme.

(18) *[и с тех пор празднование сопровождается широкомасштабными кампаниями, посвященными важнейшим экологическим проблемам:] [~~загрязнению мирового океана,~~] [перенаселенности планеты,] [глобальному потеплению.]*
[i s tekh por prazdnovanie soprovozhdaetsya shirokomasshtabnymi kampaniyami, posvyashchennymi vazhnejshim ekologicheskim problemam:] [~~zagryazneniyu mirovogo okeana,~~] [perenaselennosti planety,] [global'nomu potepleniyu.]
'and since then, the celebration has been accompanied by widespread campaigns on major environmental issues: ocean pollution, overpopulation of the planet, and global warming'

However, we do not observe this effect if there are only two EDUs in the list itself.

(19) *[Все страны требуют регистрационных знаков для таких наземных транспортных средств,] [как легковые и грузовые автомобили,] [а также мотоциклы.]*
[Vse strany trebuyut registracionnyh znakov dlya takih nazemnyh transportnyh sredstv,] [kak legkovye i gruzovye avtomobili,] [a takzhe motocikly.]
'All countries require registration plates for land vehicles, such as light and cargo vehicles, and also motorcycles.'

The last element of the list usually does not show a shorter fixation: probably the list is "assembled" as a whole at this moment. This hypothesis needs to be further tested using regression and saccade analyses. The exceptions are the examples with extremely short and very homogeneous lists.

(20) *[Наиболее популярные цвета, используемые для государственных флагов,] – [красный,] [~~белый,~~] [~~зеленый~~] [~~и синий~~.]*
*[Naibolee populyarnye cveta, ispol'zuemye dlya nacional'nyh flagov,] – [krasnyj,] [~~belyj~~,] [~~zelenyj~~] [~~**i sinij**~~.]*

'The most popular colors used for national flags are red, white, green, and blue.'

It is interesting that not only nominal enumerations, but also enumerations of situations expressed by verbal predicates with arguments behave as listings, as in example (21).

(21) *[Более того, некоторые производители выпаривают сок,]* ***[и затем снова добавляют в него воду]*** *[или разбавляют водой заранее изготовленный концентрат.]*
[Bolee togo, nekotorye proizvoditeli vyparivayut sok,] ***[i zatem snova dobavlyayut v nego vodu]*** *[ili razbavlyayut vodoj zaranee izgotovlennyj koncentrat.]*
'Moreover, some producers evaporate the juice and then add water to it again, or dilute a pre-made concentrate with water.'

We suppose that the cognitive mechanisms that determine how lists are read, processed, and remembered within a coherent text require, in principle, require a separate study because they may work differently from the processing of narrative fragments. Not coincidentally, longer lists usually require special formatting with a separate line for each element and bullet points to be better understood and remembered. Perhaps the fact that middle elements tend to be skipped or have shorter fixation times has a significant impact on the processing of the entire list and requires further investigation using other eye tracking measures, such as regression probabilities.

5. Conclusion

In this paper, we tried to identify different ways in which discourse structure affects texts complexity. To do so, we analyzed eye movement data of 35 readers for a collection of 12 Russian texts from the MECO project (Kuperman et al. 2022a, b). The analysis of eye movements is the most precise method that can be used to assess the complexity of the text for a reader. Moreover, it provides the researcher with many measures potentially reflecting different processing stages that can be used to study linguistic phenomena from the word level up to the whole text level.

However, this richness made our task more challenging. The influence of low-level factors, primarily the length, frequency and predictability of individual words, tends to obscure the effects of the higher-level factors, and their contribution cannot be measured directly. In the present paper, we came up with an approach that let us overcome this problem and identify several types of EDUs that are read significantly faster than expected. We hypothesized that this can be explained by their discourse properties, but could only prove this by qualitative analysis on the dataset we analyze in the present paper.

We view this as the first step that opens multiple avenues for further research. Firstly, we plan to test the hypotheses we formulated on other sets of eye-tracking data, both from Russian and from other languages. Some predictions may

eventually be tested experimentally. Secondly, we plan to explore other eye-tracking measures: at first, using the same qualitative approach that we adopted I the present study and then validating the emerging generalizations on other data sets.

Acknowledgements

We would like to thank Varvara Magonmedova for her help on data preprocessing and Ivan Pozdnyakov for his help in statistical processing of the data in R.
The work has been supported by the Ministry of Science and Higher Education of the Russian Federation within Agreement No 075-15-2020-793.

REFERENCES

Aaronson, Doris & Steven Ferres. 1983. Lexical categories and reading tasks. *Journal of Experimental Psychology: Human Perception and Performance* 9(5). 675.

Aaronson, Doris & Hollis S. Scarborough. 1976. Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance* 2(1). 56.

Balogh, Jennifer, Edgar Zurif, Penny Prather, David Swinney & Lisa Finkel. 1998. Gap-filling and end-of-sentence effects in real-time language processing: Implications for modeling sentence comprehension in aphasia. *Brain and Language* 61(2). 169–182.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01

Boston, Marisa Ferrara, John T. Hale, Reinhold Kliegl & Umesh Patil. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Researc* 2. 1–12.

Carlson, Lynn & Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545 54*. 56.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.

Chang, Frederick R. 1980. Active memory processes in visual sentence comprehension: Clause effects and pronominal reference. *Memory & Cognition* 8(1). 58–64.

Clifton, Charles Jr., Adrian Staub & Keith Rayner. 2007. Eye movements in reading words and sentences. In R. van Gompel (ed.), *Eye movements: A window on mind and brain*, 341–372. Amsterdam, Netherlands: Elsevier.

Cop, Uschi. 2017. Presenting GECO: An eye-tracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods* 49. 602–615.

Degand, Liesbeth & Anne-Catherine Simon. 2005. Minimal Discourse Units: Can we define them, and why should we. *Proceedings of SEM-05. Connectors, Discourse Framing and Discourse Structure: From Corpus-Based and Experimental Analyses to Discourse Theories* 477. 65–74.

Frank, Stefan L., Irene Fernandez Monsalve, Robin L. Thompson & Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods* 45. 1182–1190.

Ferreira, Fernanda & John M. Henderson. 1993. Reading processes during syntactic analysis and reanalysis. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale* 47(2). 247.

Gernsbacher, Morton A., Varner R. Kathleen & Mark E. Faust. 1990. Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16(3). 430.

Givón, Talmy. 1993. Coherence in text, coherence in mind. *Pragmatics & Cognition* 1(2). 171–227.

Grosz, Barbara J. & Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12. 175–204.

Halliday, M. A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Hirotani, Masako, Lyn Frazier & Keith Rayner. 2006 Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language* 54(3). 425–443.

Hirschberg, Julia & Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19(3). 501–530.

Husain, Samar, Shravan Vasishth & Narayanan Srinivasan. 2014. Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research* 8. 1–12.

Just, Marcel A. & Patricia. A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87. 329–354.

Kibrik, Anrej A. & Vera I. Podlesskaya. 2009. *Night Dream Stories: Corpus Study of Russian Discourse*. Moscow: Yazyki slavyanskih kultur. (In Russ.)

Kliegl, Reinhold, Antje Nuthmann & Ralf Engbert. 2006. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General* 135. 12.

Kuperman, Victor. 2010. The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology* 63(9). 1838–1857. https://doi.org/10.1080/17470211003602412

Kuperman, Victor. 2022a. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). To appear in *Behavior Research Methods.*

Kuperman, Victor. 2022b. Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus (MECO). To appear in *Studies in Second Language Acquisition.*

Laurinavichyute, Anna K., Irina A. Sekerina, Svetlana Alexeeva, Kristine Bagdasaryan & Reinhold Kliegl. 2019. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods* 51. 1161–1178.

Luke, Steven G. & Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods* 50. 826–833.

Mann, William C. & Sandra A. Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization.* Los Angeles: University of Southern California, Information Sciences Institute.

Marian, Viorica, Henrike K. Blumenfeld & Margarita Kaushanskaya. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research* 50. 940–967.

Podlesskaya, Vera I. & Andrej A. Kibrik. 2003. Methods of oral speech corpora research: Discourse transcription development experience. *Proc. of Cognitive Modeling in Linguistics*. Varna, Bulgaria.

Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics* 12. 601–638.

Prasad, Rashmi, Bonnie Webber & Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. 87–97.

Pynte, Joël & Alan Kennedy. 2006. An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French. *Vision Research* 46. 3786–3801.

Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124. 372.

Rayner, Keith. 2012. *Psychology of Reading*. New York, NY: Psychology Press.

Schauer, Holger. 2000. From elementary discourse units to complex ones. *1st SIGdial Workshop on Discourse and Dialogue*. 46–55.

Van Dijk, Teun A. 2019. *Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition*. Routledge.

Warren, Tessa, Sarah J. White & Erik D. Reichle. 2009. Investigating the causes of wrap-up effects: Evidence from eye movements and E–Z Reader. *Cognition* 111(1). 132–137.

Dictionaies and internet resources / Интернет-ресурсы

MECO. https://meco-read.com (accessed 26.05.2022).

Ru-RSTreebank. https://rstreebank.ru/ (accessed 26.05.2022).

Article history:
Received: 20 October 2021
Accepted: 01 February 2022

Bionotes:

Svetlana Yu. TOLDOVA holds a Ph.D. in Philology and is Associate Professor of the School of Linguistics at the Faculty of Humanities and Head of Formal Linguistics Lab at the National Research University "Higher School of Economics". Her research interests include NLP, corpus linguistics, discourse analysis and linguistic typology.

Contact information:

e-mail: toldova@yandex.ru

room A-114, Building 1, 21/4 Staraya Basmannaya Ulitsa, Moscow, 117218, Russia

ORCID: 0000-0002-5777-9161

Natalia A. SLIOUSSAR is Doctor Habil., Associate Professor of the School of Linguistics at the Faculty of Humanities, Associate Professor of the Department of the Problems of Convergence in Natural Sciences and Humanities at St. Peterburg State University. Her research interests embrace psycholinguistics, syntax and morphology.

Contact information:

e-mail: slioussar@gmail.com

room 518, Building 1, 21/4 Staraya Basmannaya Ulitsa, Moscow, 117218, Russia

ORCID: 0000-0003-1706-6439

Anastasia A. BONCH-OSMOLOVSKAYA is an Associate Professor of the Faculty of Humanities at the Higher School of Economics, Moscow, Russia. She holds a PhD in Language Theory. Her research interests include digital humanities, computational linguistics, corpus linguistics, theoretical linguistics and quantitative methods in linguistics.

Contact information:
room 521, Building 1, 21/4 Staraya Basmannaya Ulitsa, Moscow, 117218, Russia
e-mail: abonch@gmail.com
ORCID: 0000-0001-5826-8286

Сведения об авторах:
Светлана Юрьевна ТОЛДОВА – кандидат филологических наук, доцент Школы лингвистики факультета гуманитарных наук НИУ ВШЭ, заведующая научно-учебной лабораторией по формальным моделям в лингвистике. Ее научные интересы включают компьютерную лингвистику, корпусную лингвистику, анализ дискурса, лингвистическую типологию.
Контактная информация:
e-mail: toldova@yandex.ru
Россия, 117218, Москва, Старая Басманная ул., д. 21/4, стр. 1, каб. А-114
ORCID: 0000-0002-5777-9161

Наталия Анатольевна СЛЮСАРЬ – доктор филологических наук, доцент Школы лингвистики факультета гуманитарных наук НИУ ВШЭ, доцент кафедры проблем конвергенции естественных и гуманитарных наук, старший научный сотрудник института когнитивных исследований СПбГУ. В сферу ее научных интересов входят психолингвистика, морфология и синтаксис.
Контактная информация:
e-mail: slioussar@gmail.com
Россия, 117218, Москва, Старая Басманная ул., д. 21/4, стр. 1, каб. А-114
ORCID: 0000-0003-1706-6439

Анастасия Александровна БОНЧ-ОСМОЛОВСКАЯ – доцент факультета гуманитарных наук университета «Высшая школа экономики» (Москва), кандидат филологических наук по специальности «Теория языка». Сфера ее научных интересов включает цифровую гуманитаристику, компьютерную лингвистику, корпусную лингвистику, теоретическую лингвистику, квантитативные методы в лингвистике.
Контактная информация:
Россия, 117218, Москва, Старая Басманная ул., д. 21/4, стр. 1, каб. 521
e-mail: abonch@gmail.com
ORCID: 0000-0001-5826-8286

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 471–492
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-31187

Research article

Word-formation complexity: a learner corpus-based study

Olga LYASHEVSKAYA[1,2] ✉, Julia PYZHAK[1] and Olga VINOGRADOVA[1]

[1]National Research University Higher School of Economics, Moscow, Russia
[2]Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia
✉olesar@yandex.ru

Abstract
This article explores the word-formation dimension of learner text complexity which indicates how skilful the non-native speakers are in using more and less complex – and varied – derivational constructions. In order to analyse the association between complexity and writing accuracy in word formation as well as interactive effects of task type, text register, and native language background, we examine the materials of the REALEC corpus of English essays written by university students with Russian L1. We present an approach to measure derivational complexity based on the classification of suffixes offered in Bauer and Nation (1993) and then compare the complexity results and the number of word formation errors annotated in the texts. Starting with the hypothesis that with increasing complexity the number of errors will decrease, we apply statistical analysis to examine the association between complexity and accuracy. We found, first, that the use of more advanced word-formation suffixes affects the number of errors in texts. Second, different levels of suffixes in the hierarchy affect derivation accuracy in different ways. In particular, the use of irregular derivational models is positively associated with the number of errors. Third, the type of examination task and expected format and register of writing should be taken into consideration. The hypothesis holds true for regular but infrequent advanced suffixal models used in more formal descriptive essays associated with an academic register. However, for less formal texts with lower academic register requirements, the hypothesis needs to be amended.
Keywords: *linguistic complexity, morphological complexity, writing accuracy, word formation, English, learner corpora*

For citation:
Lyashevskaya, Olga, Julia Pyzhak & Olga Vinogradova. 2022. Word-formation complexity: a learner corpus-based study. *Russian Journal of Linguistics* 26 (2). 471–492. https://doi.org/10.22363/2687-0088-31187

Научная статья

Словообразовательная сложность и ошибки учащихся в экзаменационных эссе

О.Н. ЛЯШЕВСКАЯ[1,2] ✉, Ю.В. ПЫЖАК[1], О.И. ВИНОГРАДОВА[1]

[1]Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
[2]Институт русского языка им. В. В. Виноградова РАН, Москва, Россия
✉olesar@yandex.ru

Аннотация
В статье рассматривается словообразовательная сложность учебных текстов, которая трактуется как система измерений, показывающих разнообразие приемов словообразования разного уровня, от простых до продвинутых, используемых учащимся. Анализируется взаимосвязь между сложностью и ошибками, которые учащиеся допускают в словообразовании. Исследование основано на материалах REALEC – корпуса английских экзаменационных эссе, написанных студентами университета с родным русским языком. Предлагается подход к измерению словообразовательной сложности, основанный на классификации суффиксов Бауэра и Нейшена (Bauer & Nation 1993), и анализируется соответствие между показателями индексов сложности и количеством ошибок словообразования, размеченных в текстах корпуса, с учетом типа экзаменационного задания. Постулируется гипотеза о том, что с увеличением сложности количество ошибок должно уменьшаться, и проводится статистический анализ параметров сложности и безошибочности. В работе показано, во-первых, что использование словообразовательных суффиксов более высокой сложности связано с количеством ошибок в текстах. Во-вторых, разные уровни иерархии сложности оказывают разнонаправленное влияние на точность: в частности, использование нерегулярных словообразовательных моделей положительно связано с количеством ошибок. В-третьих, следует учитывать тип экзаменационного задания, в том числе ожидаемые формально-регистровые особенности текста. Гипотеза была подтверждена для регулярных, но нечастотных суффиксальных моделей при их использовании в описаниях рисунков и графиков – текстах, следующих определенному формату и включающих элементы академического письма. Однако в случае аргументативных эссе выдвинутая гипотеза требует уточнения.
Ключевые слова: *лингвистическая сложность, морфологическая сложность, безошибочность письма, словообразование, английский язык как иностранный, учебные корпуса*

Для цитирования:
Lyashevskaya O.I., Pyzhak J.V., Vinogradova O.N. Word-formation complexity: a learner corpus-based study. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 471–492. https://doi.org/10.22363/2687-0088-31187

1. Introduction

Text complexity is one indicator that can be used to assess authors' written and spoken skills and how varied and complex are the means of linguistic expression

they apply. Starting with the work of Norris and Ortega (2009) and two publications by Bulté and Housen (2012, 2014), researchers agree that complexity is a multifaceted phenomenon, consisting of several sub-constructs, dimensions, levels, and components, and many of these constructs and sub-areas have been independently evaluated. Consequently, the term ‘‘complexity’’ has been widely applied to manifestation of objective properties of linguistic production. Bulté and Housen (2012) also draw the distinction between propositional complexity, discourse-interactional complexity, and linguistic complexity. Of these three, linguistic complexity of written production, will be the target of the present article.

In addition to the general consent on including lexical, syntactic, discursive, and morphological parameters of texts when looking at text complexity, it soon became clear that the majority of researchers focus on the first three areas, while morphological and phonological complexity or complexity phenomena at the interfaces between the traditional levels of linguistic analysis were largely ignored, which resulted in some appeals to the linguistic community to expand the construct of, and research on, complexity beyond the syntactic and lexical levels. In 2019, Centre for English Corpus Linguistics at UCLouvain hosted the colloquium, Broadening the Scope of L2 Complexity Research, and in a way the research in this paper is our contribution to bridging this gap. More specifically, we set aside the topic of inflectional complexity, which has already gained considerable attention in learner data analysis and has acquired its own methodology (Brezina & Pallotti 2016, Yoon 2017, Tywoniw & Crossley 2020), and focus exclusively on the diversity of word-formation models.

Roots and affixes are the building blocks of morphological competence. Acquiring the morphology of a second language can be seen as not only learning strategies for composing and decomposing forms from and to the building blocks and enhancing vocabulary. It is also learning relations between the stored word forms that facilitates processing at the morphological level (Baerman et al. 2015). Word-formation skills give L2 learners flexibility when modifying and adapting word meaning to the context, coercing a word’s syntactic class to fit into an available grammatical construction, choosing the right vocabulary for a given type of discourse, or in constructing new words. Inappropriate use of derivational models may result in communication fallacy and other losses in social interactions, just as any kind of erroneous linguistic behaviour may do.

Recent discussion on the status of word(-formation) families has brought to light many derivation-related issues important for L2 research and education (Brown et al. 2020, Laufer et al. 2021, Nation 2021). To put it in a nutshell, the researchers emphasise that the derived forms cover a significant portion of texts and hence learners’ word-formation competence impacts text comprehension. Affix knowledge develops with general proficiency and facilitates vocabulary learning. However, many advanced learners have limited or patchy knowledge of affixes and find it challenging to identify derivational forms of known headwords given in context, even in structures with top-frequent affixes. Brown et al. (2020) present some evidence that the learners are concentrated on the recognition of the affix

meaning rather than its grammatical function. Although the evidence provided largely concerns the receptive aspects of acquisition, the discussion has important implications for the theories of L2 production.

In the task of examination writing, learners are expected to achieve a balance in the interaction of the performance areas such as complexity, accuracy, and fluency in choosing preferable derivational strategies. Although the lack of high-level and complex skills is not the only source of word-formation errors and that low writing quality may also be related to, for example, time and stress management, genre of the task and familiarity of the topic, and learners' individual effects, learner corpora and annotated research datasets provide the basis for empirical studies focused on the relationship between complexity and accuracy (Skehan 2009, Bardovi-Harlig & Bofman 1989, Plakans et al. 2019, Lahuerta 2018, among others).

Our study is based on the learner corpus REALEC (Vinogradova et al. 2017), which includes examination writings of students with Russian L1. We examine the use of word-formation constructions with the focus on suffixal ones, since their number in the examination texts significantly exceeds the number of prefixes and other word-formation units. We will test the following hypothesis: the higher the parameter of derivational morphological complexity, the less often errors occur in word formation, since the scale for measuring morphological complexity is based on the order in which students studying English as a foreign language learn derivational affixes. Accordingly, the more the student uses advanced suffixes, the higher their language proficiency.

2. Word-formation complexity

There are several ways to define complexity, among which relative complexity (difficulty) and absolute (structural) complexity are most discussed (de la Torre García et al. 2021). Relative complexity refers to the cognitive difficulty of the task, the amount of effort and resources that a speaker has to employ in order to process and make use of a linguistic structure. In contrast to this, absolute complexity is defined as the numerical characteristics of a text based on the quantity of encoded and encoding linguistic units and the number of connections between these components. Morphological complexity belongs to the area of formal parameters of absolute complexity, according to the classification of parameters critical for measuring the acquisition of a target language (Bulté & Housen 2012), see Fig. 1. These are features which can be measured objectively at the word level in a text.

Figure 1 shows that the criteria of morphological complexity are divided into two groups: inflectional and derivational. The first type refers to the use of grammatical forms, for example, the frequency of tense forms, the frequency of modals, the number of different verb forms, the variety of past tense forms, and MCI (Morphological Complexity Index) (Brezina & Pallotti 2019). The second group of criteria deals with the use of derivational affixes, composites (multi-root words), and conversives, and refers to the variability of word-formation models and the size of word families.

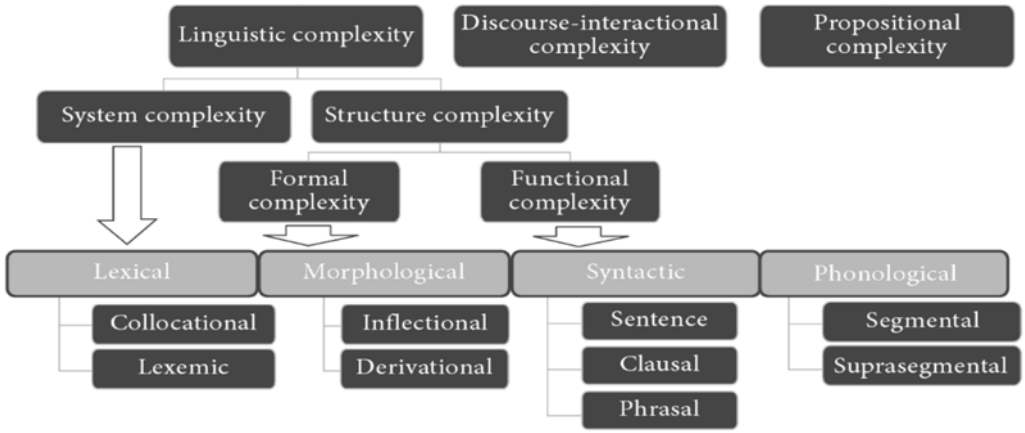

Fig. 1. A taxonomy of absolute complexity criteria (Bulté & Housen 2012: 23)

Word-formation complexity criteria have gained much less attention compared to inflectional complexity criteria. A possible explanation for this is that there is low agreement among theoretical approaches to this level of linguistic description, the coverage of derivational models, affixes, and that their taxonomy in lexical resources is sketchy and inconsistent (cf. Table 1), and it is problematic to use (a very few) tools for automatic morpheme segmentation of non-standard texts. It is indicative that Biber used only one derivation-related feature Nominalizations (words ending in *-tion*, *-ment*, *-ness*, *-ity*) among the other 67 linguistic features in his multidimensional analysis of English register (Biber 1988). A slightly longer list of regular affixes (*-able*, *-er*, *-ish*, *-less*, *-ly*, *-ness*, *-th*, *-y*, *non-*, and *un-*) was used in the calculation of a types-per-family ratio in a study of vocabulary structure (Horst & Collins 2006), yet no distinction was drawn between inflection and derivation in the construct of word families.

Recently, Tywoniw and Crossley (2020) introduced the TAMMI method, in which the density of derived and non-derivational words is measured in the text, among other metrics of (inflectional) morphological complexity. To calculate the Derived Word Tokens per Word index, they divide the number of word tokens with any derivational affixes by the number of words in a text. The Derived Word Types per Type index is the number of distinct word types with any derivational affixes divided by the number of types in a text. Similar calculations apply to the measurement of Non-Derivational Word Tokens per Word and Non-Derivational Word Types per Type indices.

Tywoniw and Crossley's approach is a generalisation over statistical measures developed for tokens (N) and types (V) of a given morphological class in corpora-based research (Plag et al. 1999). Other derivational complexity indicators recommended in the literature include:

– the length of the affix;

– the complexity-based rank, constructed on the affix ordering principle: suffixes with a rank greater than the rank of a given affix may follow that suffix in a word, while suffixes with rank lower than this will never follow it;

– the derivative's junctural phonotactics: the probability of the sequence of sounds spanning the juncture between the morphemes (e.g. "nh", which is highly unlikely to occur within morphemes and more likely to create morphological boundaries, as in *inhumane*) (see Baayen 2009 for an overview).

Table 1. **Numbers of distinct affixes attested in English L1 corpora, dictionaries, and other resources according to (Laws & Ryder 2014: 6, 10)**

	Word-initial affixes	Word-final affixes	Totals
Marchand (1969)	65	104	169
Hay and Baayen (2002)	26	54	80
Stein (2007)	547	296	843
Incl. prefixes/suffixes	171	164	
combining forms	125	107	
both	251	25	
Affixes in BNC	268	222	490
Incl. prefixes/suffixes	96	141	
combining forms	41	61	
both	131	20	
Affixes in MorphoQuantics, not in BNC	286	59	345
Incl. prefixes/suffixes	81	22	
combining forms	84	32	
both	121	5	

Bauer and Nation (1993) proposed a scale for categorising English affixes on the following criteria:

1) frequency – the number of different words an affix occurs in;
2) productivity – the likelihood that the affix will be used to form new words;
3) the predictability of the meaning of the affix;
4) the regularity of the orthographic form of the base – the predictability of change in the written form when the affix is added;
5) the regularity of the spoken form of the base – the amount of phonetic change when the affix is added;
6) the regularity of the spelling of the affix – the number of allomorphs attested in different words;
7) the regularity of the pronunciation of the affix;
8) the regularity of function – the degree to which the affix attaches to a base of a known form-class and produces a word of a known form-class.

Their classification implies that the affixes are acquired at different stages of language mastery. For example, the first level means that a speaker perceives a different form as a different word. Being on the second level, they understand that regularly inflected words are part of the same family. At subsequent levels, different derivational affixes are acquired, see Table 2 for higher level suffixes. One can notice that the same orthographic manifestations of suffixes can be attested at different levels (marked with asterisk in Table 2), cf. the suffix *-able* that appears at both the third and 6th levels. The third level includes cases of attaching this suffix

to transitive verbs, which is a very frequent, productive, and regular word-formation construction. At the 6th level, the base has to be truncated (e.g., the suffix *-ate* removed) when the suffix *-able* is attached, as in *attenuable*, *permeable*, thus leading to more complexity. At the last, seventh, level, the student demonstrates near-native knowledge of every existing morpheme including rare combinations of affixes and roots of Latin and Greek origin, cf. *differentiate*.

Table 2. **Levels of English suffixes, according to (Bauer & Nation 1993)**

Level	Suffixes
3: the most frequent and regular affixes	*-able** (*eatable*), *-er* (*writer*), *-ish* (*selfish*), *-less* (*endless*), *-ly** (*fortunately*), *-ness* (*kindness*), *-th** (*fourth*), *-y** (*smelly*)
4: frequent, orthographically regular affixes	*-al** (*normal*), *-ation* (*preparation*), *-ess* (*heiress*), *-ful* (*useful*), *-ism* (*socialism*), *-ist** (*socialist*), *-ity* (*sensitivity*), *-ize* (*legalize*), *-ment* (*government*), *-ous* (*ambitious*)
5: regular but infrequent affixes	*-age* (*percentage*), *-al** (*approval*), *-ally* (*idiotically*), *-an* (*American*), *-ance* (*clearance*),*-ant* (*consultant*), *-ary* (*revolutionary*), *-atory* (*confirmatory*), *-dom* (*kingdom*), *-en* (*wooden*), *-en* (*widen*), *-ence* (*emergence*), -ly* (*leisurely*) ...
6: frequent but irregular affixes	*-able** (*permeable*), *-ee* (*nominee*), *-ic* (*geographic*), *-ify* (*quantify*), *-ion* (*description*), *-ist** (*tobacconist*), *-ition* (*addition*), *-ive* (*representative*), *-th** (*length*), *-y** (*diplomacy*)

The consistency of Bauer and Nation's hierarchy was confirmed in Leontjev (2016): the study tested how successfully the learners of English recognised affixes belonging to different levels. With the exception of no difference between levels 5 and 6, the lower the affix level, the easier it is for a student to recognise it. However, other studies suggest that the Bauer and Nation levels "only partly agree with learner knowledge data" (Nation 2021: 969). There are also known challenges in mapping the affix levels to the CEFR Lexical Profile levels and word meanings (Capel 2010).

In Lyashevskaya et al. (2021), Bauer & Nation's classification of suffixes was operationalised through four quantitative indices for levels 3 to 6. To calculate the index of each level, they divide the number of tokens with suffixes of level N by the number of tokens that have any inflectional or derivational suffix. Although a wider variety of indices can be designed that take into account all types of word-formation devices (e.g. prefixes and morphemes within composite words), the relation between the suffix-based derivational indices and inflectional complexity index is straightforward for English morphology since both calculation methods rely on the number of suffixes that have to be correctly supplied in writing.

3. Related research

While many studies have addressed the relationship between morphological complexity and the level of proficiency in a foreign language, unsurprisingly, most research is based on inflectional complexity only under the notion of morphological complexity. Over the years, a wealth of methods, such as MCI, and tools, such as

LancsBox (Brezina et al. 2020), have been developed to facilitate the empirical study of inflectional data. One of the illustrative examples is Brezina and Palloti (2019), whose aim was to reveal the relationship between the realised inflectional complexity in texts of Italian learners and their level of language proficiency. They show significant correlation between measures of inflectional complexity and other indicators of complexity such as a standardised type-token ratio and sentence length. However, the effects are not observed in groups of advanced learners in the study based on written argumentative essays in English produced by Italian university students, taken from the International corpus of learner English (ICLE). In the case of English, which is less complex than Italian in terms of verbal inflection, the authors at CEFR levels B1 to C1 demonstrate a native-like ability, thus reaching a threshold "after which inflectional diversity remains constant" (Brezina & Palloti 2019: 99).

Ehret and Szmrecsanyi (2019) studied the relationship between the number of years of L2 instruction in English and morphological and syntactic complexity as well as holistic, information-theory-based complexity of text (Kolmogorov complexity). The authors found that writers with higher levels of language mastery wrote more complex texts, although the relationship between complexity and proficiency was not always linear. While holistic text complexity and morphological complexity of the text increased, the syntactic complexity might decrease as more advanced students were less likely to adhere to the correct word order.

The effect of the derivational complexity of the native language in L2 acquisition was investigated in van der Slik et al. (2019). In particular, the authors found a link between the derivational complexity of a student's native language and their success in learning Dutch. Students whose native language was morphologically less difficult than Dutch found it more difficult to acquire the morphological system of Dutch.

In Kimppa et al. (2019), the acquisition of word formation in adult students of Finnish was studied using psycholinguistic methods. It was found that more advanced students showed performance comparable to that of native Finnish speakers when processing derivational morphology. This suggests that with an increase in the level of proficiency in a foreign language, the processing of derivational morphology also develops, and some learners can reach the level of a native speaker.

Our study aims at narrowing the gap in the studies of word-formation complexity by focusing attention on its relationship to accuracy. It addresses the following research questions:

1. Is there an association between word-formation complexity and error-free use of the word-formation models in L2 English production?

2. Is there a task effect? Does the word-formation complexity affect the frequencies of the word-formation errors in Task 1 texts the same way as in Task 2 texts? Which specific levels of word-formation complexity contribute to the accuracy of word formation the most?

4. Data and method

This study is based on REALEC, Russian Error-Annotated English Learner Corpus. This corpus currently includes about 6,000 texts (roughly 1.5 million words) of the examination essays written by Russian-speaking learners of English during their second-year examination at university. The writing tasks are similar to those used in the IELTS examination. The first task tests the ability to describe graphic material in the task (graph description essay), and in the second one the participant is expected to express their opinion about a certain problem given in a short written text prompt (argumentative essay). The required size of the first type of essay is at least 150 words; the second one, at least 250. A substantial part of the corpus was manually processed by EFL experts: they annotated and corrected errors at different linguistic levels from orthographic and morphological to discursive.

We have selected from the corpus 1307 examination texts containing errors in suffixal word formation. In our sample, there are no texts without derivational errors due to the well-known problem of recall in spotting such errors by experts. The errors that we selected were labelled with the following tags: "Formational suffix", "Word formation" and "Confusion of categories". The "Formational suffix" tag marks inappropriate use of a suffix or its absence where a suffix is needed – see examples (1) and (2). The "Word formation" tag combines three types of errors: incorrect use of both a suffix and a prefix, or the absence of both where they are needed, or the combination of the first two types – see (3) and (4). The corrections suggested by experts for the errors in focus in this research are given in square brackets, while corrections suggested for all other errors are not presented, so the authors' spelling, grammar and vocabulary are intact in the examples.

(1) *Business books empahsize, that society and all its members must feel confident in every step taking by **politics [politicians]** and what is more, to have an essence of non-restricted protection.*
(2) *First of all, if we speak of equality of men and women we should make a **notice [note]** that this also mean that women could not do some work which is not suit them (take heavy things).*
(3) *One of researches showed, that the **borned [inborn]** characteristics more important for our personality and development.*
(4) *Today a lot of international organization move their businesses to **undeveloped [developing]** countries.*

The "Confusion of categories" label is used to mark cases of incorrect choice of a part of speech from a word family, see (5). This tag, unlike the previous two, describes the nature of the error rather than the formal type of what needs to be corrected in the error span. However, among the errors labelled with this tag, there are also cases of incorrect use of word-formation suffixes.

(5) *What is about global warming, recent studies say that it is a result of **climatic [climate]** changes which are essential for the earth.*

Spelling mistakes in affixes and on morpheme boundaries such as *useage* (cf. *usage*), *privelage* (cf. *priviledge*), *begining* (cf. *beginning*) are tagged as

"Spelling" and not as errors in word formation, according to the REALEC annotation scheme. Such cases are excluded from consideration. We also removed texts in which only the incorrect uses of prefixes and composites, and not suffixes, were annotated.

The resulting dataset consists of 595 graph description essays and 712 argumentative essays. Table 3 reports the number of errors labelled in the texts of each examination task type. Most of the texts contain only one derivational error. Texts in which no more than four derivational errors were reported account for 96% of documents in each type.

Table 3. **Distribution of the number of errors in two types of examination writing**

# errors	graph description	argumentative essay	total
1	376	391	767
2	118	170	288
3	44	95	139
4	33	25	58
5	12	17	29
6	8	5	13
7	2	5	7
8	1	1	2
9		2	2
10		1	1
11	1		1
total	595	712	1307

There is no available information regarding the proficiency level of the authors, however, it can be estimated within the range from B1 to C1 on the CEFR scale.

To measure derivational complexity, we used the application Inspector (Lyashevskaya et al. 2021). In each text, it calculates (token-wise) the number of word-formation affixes on each of the levels 3 (most frequent and regular), 4 (frequent and orthographically regular), 5 (infrequent and regular), and 6 (frequent and irregular) of Bauer & Nation's classification of word affixes (see Section 2). The simple base forms are identified using the nltk package PorterStemmer, after which the total number of suffixed words in the text is calculated, thus taking into account both inflectional and derivational morphemes at the end of the word. The relative metrics of each level are calculated according to the formula:

$$\frac{number\ of\ suffixes\ on\ n^{ı}s\ level}{number\ of\ suffixed\ words}.$$

The basic statistics of the mean, standard deviation (SD), median values of the suffix level indices, and accuracy index (number of errors) are summarised in Table 4. The average length of texts is: 182.8 words (*SD*=37.4) for the graph description essays and 275.2 words (*SD*=62.3) for argumentative essays.

Table 4. **Descriptive statistics of word-formation complexity measures and word-formation errors**

task	stats	level 3	level 4	level 5	level 6	# errors
graph description	Mean	0.033	0.083	0.046	0.042	1.709
	SD	0.032	0.071	0.041	0.042	1.234
	Median	0.026	0.069	0.038	0.029	1.000
argumentative	Mean	0.049	0.080	0.067	0.052	1.829
essay	SD	0.034	0.042	0.039	0.034	1.256
	Median	0.045	0.075	0.065	0.045	1.000

Several types of statistical techniques were used to investigate the research questions. The Pearson's analysis, which presupposes the continuous distribution of variables, was conducted to detect possible correlation among complexity indices, while the non-parametric rank-based Kendall correlation analysis was applied to measure pairwise the ordinal association between continuous measures of complexity and a discrete (paucal integer) measure of accuracy.

Analysis of variance of the complexity and accuracy measures was conducted to compare the groups of texts such as graph description and opinion essays, or essays with one vs. more than one word-formation error. Since the measures are distributed non-normally we performed a non-parametric Kruskal-Wallis rank-sum test (*kruskal.test* function for two groups in the R package *stats*, R Core Team 2019, Hollander & Wolfe 1973: 115–120, 185–194).

In addition, we applied two regression algorithms. We assume that the word-formation errors follow a Poisson distribution, since it describes the likelihood of events that occur over a fixed period of time, and the events are independent of each other. When a student writes an essay, an error may or may not occur at any given moment. We used a Poisson regression (*vglm* function in the R package VGAM, Yee 2015) to model the number of errors (count dependent variable that ranges from 1 to 11) by the indices of derivational complexity (four independent non-normally distributed variables). The zero-truncated model, based on a positive Poisson distribution, is better suited for data in which no zeros in the response variable is attested, as in our case.

In order to determine whether we need to apply a one-inflated Poisson regression model due to the excess of ones in our response variable (# errors=1, see Table 3), we fitted a binary logistic regression model (*glm* function in the R package *stats*, R Core Team 2019, Dobson 1990) which estimated the probability of the response falling into one of two groups: texts having one error *vs.* texts having two and more errors. The same four complexity indices were used in this model as independent variables.

5. Results

We ran Pearson's correlation test to evaluate possible correlation among four levels of the complexity scores, considering values ($0.5 \leq r < 1$), with $p \leq 0.05$ to be

a strong correlation. Only a weak correlation was observed between levels 3 and 6 (r=0.23), levels 3 and 4 (r=-0.09), and levels 4 and 5 (r=0.08) in graph description essays. As for argumentative essays, there was a medium positive correlation between levels 4 and 5 (r=0.38, p<0.05) and a weak correlation between some other levels (r=0.24 for levels 3 and 6, -0.19 for levels 3 and 4, -0.09 for levels 4 and 6, -0.08 for levels 5 and 6).

In what follows, we assess the effect of the examination task using a non-parametric analysis of variance, and argue for the need for a separate analysis of data in two task types. After that, we show the results of a Poisson regression analysis that estimates the effect of complexity on the number of errors in essays of each examination type. In each group, we further split the data into two subgroups by the number of word-formation errors and present the results of non-parametric analysis of variance and regression analysis performed on these subgroups.

5.1. Effect of examination task

The results of comparisons based on Kruskal-Wallis rank sums are given in Table 5. The analysis reveals that at all levels of derivational complexity and with respect to the number of word-formation errors, there is a significant difference (p<0.05) between the texts of graph description and argumentative essays. This is in line with the conclusion of (Lyashevskaya et al. forthc.) that the texts of the two examination tasks invoke different patterns of complexity and accuracy. In Sections 5.2, 5.3, and 5.4 the analysis is conducted separately for the two task types.

Table 5. ***Non-parametric analysis of variance in the groups of graph description and opinion essays***

	H chi-squared	*p*-value
der_level3	96.732 .	2.2e-16 ***
der_level4	3.9733 .	0.04623 ** .
der_level5	107.96 .	2.2e-16 ***
der_level6	45.258 .	1.727e-11 ***
errors	7.5932 .	0.005859** .

5.2. Derivational complexity and accuracy in graph description

Table 6 shows the estimated coefficients and statistical significance of Poisson's zero-truncated regression model fitted for graph description essays. The number of errors in the model is conditioned by four word-formation complexity metrics. The complexity of suffixes at level 5 (orthographically regular but infrequent affixes) and level 6 (frequent but orthographically irregular models) was found to be significant predictors ($p < 0.05$). The model suggests that the number of errors decreases by 30% with each additional 0.1-point increase in the level 5 complexity, and increases by 29.5% for 0.1-increase in the level 6 complexity.

Table 6. **Summary of Poisson's zero-truncated regression model for graph description data**

	Estimate	***p*-value**
const	0.2884 .	0.00728 ***
der_level3	-1.1467 .	0.43706 . .
der_level4	-0.5626 .	0.40022 . .
der_level5	-3.4913 .	0.00495 ***
der_level6	2.5884 .	0.00981 ***

5.3. Derivational complexity and accuracy in argumentative essays

The analysis was repeated for the texts of argumentative writing. These data also show a very weak rank correlation between each of the suffix level indices and the number of errors (Kendall's τ=0.006 in the case of level 3, level 4 – 0.02, level 5 – 0.013, and level 6 – 0.029, all $p < 0.001$).

Table 7 reports the output of Poisson's zero-truncated regression model that predicts the count of word-formation errors conditioned by four suffix level complexity measures. Only the suffix complexity at level 6 (frequent but orthographically irregular models) was found to be a significant predictor ($p < 0.05$). The model suggests that with each additional 0.1-point increase in the level 6 complexity, the average number of word-formation errors increases by 30.8% while holding all other variables in the model constant.

Table 7. **Summary of Poisson's zero-truncated regression model for opinion essays data**

	Estimate	***p*-value**
const	0.1818 .	0.1500 . .
der_level3	-1.3744 .	0.2635 . .
der_level4	0.7608 .	0.4453 . .
der_level5	-0.1944 .	0.8548 . .
der_level6	2.6819 .	0.0145 *.

It should be mentioned that no interaction was observed between the complexity measures in the models for both task types, suggesting that these measures are independent and combine additively such that the outcome is better predicted by a simple weighted sum of the indices.

5.4. Texts with one error vs. more than one error

56% of essays have only one word-formation error in our sample. So it is possible that the regression models we presented above underestimate the probability of ones in the response – an effect known as one-inflation (Hassanzadeh & Kazemi 2017).

We ran a rank-sum one-way analysis of variance, dividing the graph description essays into two groups: texts with one error (376 documents) and texts with two and more errors (219 documents). The results suggest that there is no difference between these two groups in regard to their complexity indices, except for level 6, see Table 8. Furthermore, the effect of the suffix complexity was not found in the binary logistic regression models for these groups conditioned by complexity (p-value>0.05 for all four coefficients in various combinations, also with backward elimination of predictors from a full model).

We repeated the same experiments with the argumentative essays (391 documents with one error, 321 documents with more than one error). No effect of complexity was found in both non-parametric analysis of variance and logistic regression (all p-values > 0.05). Therefore we conclude that there is not enough evidence to support the need for selective modelling one-inflation in our datasets.

Table 8. **Non-parametric analysis of variance in the subgroups with one error vs. more than one error**

graph description		argumentative	
	***p*-value**		***p*-value**
der_level3	0.6891	der_level3	0.9654
der_level4	0.3823	der_level4	0.2786
der_level5	0.1477	der_level5	0.7961
der_level6	0.01621**	der_level6	0.3318

6. Analysis

According to our analysis, the use of level 5 and level 6 word-formation suffixes affects the number of derivational errors in graph description, and the use of level 6 suffixes affects the number of errors in argumentative writing. With the increase in the frequency of level 5 suffixes, the number of errors decreases, and with the increase in the frequency of level 6 suffixes, the number of errors increases.

We have to bear in mind that the expected CEFR level of learner proficiency in the examination is stated as B2, in other words, its range is from low intermediate to high intermediate level. At the intermediate level of English, the learners are expected to have acquired frequent and regular suffixal models such as *-er* in *writer* (level 3) and *-ity* in *sensitivity* (level 4). Regular but infrequent suffixes, such as *-ence* in *emergence* (level 5), had most likely been encountered by students during training and had most likely been practised sufficiently. If so, their performance in using such suffixal constructions might be at the top-right end of the U-shaped curve (Abrahamsson 2013). Much the same can be said of the level 6 suffixes (e. g. *-th* in *length*), with the refinement that irregular word-formation models are likely to be more prone to errors. Admittedly, the lexical unit encoded at level 6 is not only an idiosyncratic, non-prototypical form-function pairing (due to the complexity and irreproducibility of the form), but also belongs to a word family having a non-prototypical and non-transparent structure.

When analysing the complexity and accuracy of the university examination writing in English, several further considerations are to be taken into account. First, equivalents of English words with level 5 and 6 suffixes should have been acquired by undergraduate students in their native language. More often than not, in the case of Russian L2, such words are loanwords and/or the product of word-formation, and one can argue for the existence of near-equivalent suffixes and near-equivalent word-formation models in two languages. If a word has not been acquired and/or sufficiently trained in L2, the learner can still be successful resting on the mechanisms of generalisation, or overgeneralise and thus come to failure.

Second, a word-formation construction can be acquired in terms of morphology but not in terms of syntactic behaviour and co-occurrence. This is the case of partial lexical equivalence – when an existing L2 word-formation construction is inappropriate in a given context, for example, when a learner uses a correctly formed gerund wrongly applying a pattern of the noun to it (the decreasing in the number instead of either decreasing the number or the decrease in the number).

Third, examination writing can be considered as a product of a trade-off between complexity and accuracy according to Skehan's (1998, 2009) Trade-off Hypothesis. In principle, the learner should be interested in maximising the text complexity, including word-formation, but not at the expense of accuracy, and can therefore adopt various strategies to increase the success rate.

Fourth, task complexity and available cognitive resources can either facilitate or inhibit interactions between complexity and the quality of the output according to the later version of the Trade-off Hypothesis and the Cognition Hypothesis by Robinson (2001, 2011), see also overview in (Vasylets et al. 2017). The examination tasks in question differ in many ways and essentially diverge in that, in the case of graph description, the author apparently adheres to specific academic-like stylistically stringent register and can rely on the task prompt as a source of lexical material, whereas in the the case of argumentative writing, the text is expected to be longer, can be less formal and objective, and has to involve argumentation. When sharing his/her opinion, the student has to demonstrate advanced vocabulary knowledge and a rich supply of diverse constructions (Vinogradova et al. 2017). It is usually agreed that the learner might experience greater cognitive load in the latter task, for example, because of the need for adjusting the discourse strategy, perspective taking, choosing appropriate time and space reference, and more complex task planning in general. This can hypothetically result in a beneficial effect of increasing complexity on attention and control processes. At the same time, less advanced students may strive to avoid underdeveloped derivational patterns, but downgrading the complexity does not necessarily interfere with the number of errors.

Level 3 and 4 suffixes are not significant predictors of accuracy in either task type group, which means that lower complexity indices do not account for variance in accuracy in the essays of intermediate learners on their way towards advanced proficiency, even of the learners with a mature vocabulary in their L1. This confirms the intuition that "if we are examining text coverage for high-proficiency learners, Level 6 of Bauer and Nation is likely to be suitable" (Nation 2021).

Qualitative analysis of errors reveals a few noticeable patterns. Expectedly, non-existent derivational constructions are observed in place of the models with irregular suffixes, cf. a wrong choice of the suffix *-ion* in the word **tendention*:

(6) *There we can see an upward* ***tendention [tendency]*** *througout the years*.

Such an error can be explained as L1 interference, cf. Russian *tendencija*.

Nevertheless, the incorrect use of existing words is more common. In most cases, it is accompanied by the incorrect choice of the part of speech:

(7) *The **long [length]** of railways a more than 100 kilometres*.
(8) *Moreover, it has to be noticed, that population of aged people has tendency to **growth [grow]***.

Note that in (8), the error presumably appeared due to interference, since there is an expression *tendentsija k rostu*, lit. 'tendency to growth' in Russian.

Only the most frequent level 3 model triggers the errors in word usage, as illustrated by examples (6) and (7).

Examples (9) and (10) show the error in use of the very frequent level 3 model.

(9) *But in projection for 2050 in Yemen of population the number of **workable [working]** people will increase and will be 57,3%.*
(10) *Overall, Instagram is more **usable [used more]** by people of the age of 18–29 (approximately 50%).*

Even though the forms *workable* and *usable* are attested in L1 English, they are inappropriate in a given context. Such errors show that learning the syntactic properties of derived forms and understanding the relationship between the functions of the base and derived forms should be part of word-formation instruction.

7. Conclusion

Our study of the association between derivational complexity and the number of errors in English examination writing was motivated by the hypothesis that with increasing complexity, the number of errors will decrease: the more complex suffixes a student uses, the higher his/her level of language proficiency is. To support this hypothesis, we used the classification of derivational suffixes by Bauer and Nation (1993) and a number of statistical methods, such as non-parametric analysis of variance and regression models. Our analysis shows that the two examination tasks applied in the end-of-course examination exhibit partly different patterns.

For shorter and more formal texts, which contain descriptions of the graphical materials, only the use of advanced word-formation structures have a significant effect on the number of errors in word formation. Moreover, the effect is twofold: with an increase in the frequency of level 5 suffixes, the number of errors in word formation decreases, and with the increase in the frequency of level 6 suffixes, it increases. This may indicate that the acquisition of morphology is not linear, but wave-like: the level 5 suffixes are learned well and used confidently, whereas the level 6 suffixes may be familiar to the student (that is, he/she most probably has come across them before), yet they can still be used incorrectly. But one could make an alternative argument: it is the irregularity of word-formation models attested at level 6 that accounts for the decrease in derivational accuracy. The latter approach

indirectly supports both parallel and (semi-)sequential acquisition of the level 5 and 6 suffixes.

As for the texts written in answer to the second examination task, argumentative essays, the word-formation complexity effect narrows down to the suffixes at level 6. We can stipulate that the very format of opinion essays (rather, absence of one specific format) allows authors to choose words more at will and, accordingly, adjust the level of morphological complexity to their level of L2 acquisition in word formation in order to avoid inappropriate usage of words with certain infrequent suffixes.

Once university students have mastered how to deal with the range of word-formation means, their accuracy at this level seems to be getting more dependent on other factors, such as syntactic and discursive complexity of their writing, the range of their lexis, and individual psychological and neurophysiological reaction to the complex cognitive task.

It is important to note that our empirical findings and analysis call for a future research agenda in the area of EFL word formation, such that will examine more direct interactions at particular levels of acquisition, involves analysis of other morphological, syntactic, and discourse structure parameters as well as empirical measurements of extralinguistic factors and, as a result, will develop the methods of exploratory data analysis and computational modelling to reveal distinct learner group profiles and register-sensitive text clusters (Crossley 2020).

Acknowledgements

The research was carried out within the HSE University project ADWISER – Automated detection of writing inaccuracies for students of English in Russia, 2021. The work of Olga Lyashevskaya was partly supported by the Korean National Research Foundation (2021 General Joint Research Support Project TROIKA, 2021–2023).

The authors are grateful to Irina Panteleeva and Anna Scherbakova, who developed Inspector, a tool that has been used to calculate the complexity parameters. We also thank Lilia Rodionova for her assistance in the design and analysis of the study.

REFERENCES

Abrahamsson, Niclas. 2013. U-shaped learning and overgeneralization. In Peter Robinson (ed.), *The routledge encyclopedia of second language acquisition*, 663–664. London: Routledge. https://doi.org/10.4324/9780203135945

Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 899–919. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.899

Baerman, Matthew, Dunstan Brown & Greville G. Corbett (eds.). 2015. *Understanding and Measuring Morphological Complexity*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198723769.001.0001

Bardovi-Harlig, Kathleen & Theodora Bofman. 1989. Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition* 11(1). 17–34.

Bauer, Laurie & Paul Nation 1993. Word families. *International Journal of Lexicography* 6(4). 253–279.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Brezina, Vaclav & Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second Language Research* 35(1). 99–119. https://doi.org/10.1177/0267658316643125

Brezina, Vaclav, Pierre Weill-Tessier & Antony McEnery. 2020. #LancsBox v. 5.x. URL: http://corpora.lancs.ac.uk/lancsbox (accessed 25.05.2021).

Brown, Dale, Tim Stoeckel, Stuart Mclean & Jeff Stewart. 2020. The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, amaa061. https://doi.org/10.1093/applin/amaa061

Bulté, Bram & Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken & Ineke Vedder (eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 21–46. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.32.02bul

Bulté, Bram & Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26. 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Capel, Annette. 2010. A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal* 1(1). 2–7. https://doi.org/10.1017/S2041536210000048

Crossley, Scott. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research* 11(3). 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

de la Torre García, Nuria, María Cecilia Ainciburu & Kris Buyse. 2021. Morphological complexity and rated writing proficiency: The case of verbal inflectional diversity in L2 Spanish. *ITL – International Journal of Applied Linguistics* 172(2). 290–318. https://doi.org/10.1075/itl.20009.del

Dobson, Annette J. 1990. *An Introduction to Generalized Linear Models*. London: Chapman and Hall.

Ehret, Katharina & Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research* 35(1). 23–45. https://doi.org/10.1177/0267658316669559

Hassanzadeh, Fatemeh & Iraj Kazemi. 2017. Regression modeling of one-inflated positive count data. *Statistical Papers* 58(3). 791–809. https://doi.org/10.1007/s00362-015-0726-7

Hay, Jennifer & R. Harald Baayen. 2002. Parsing and productivity. In Geert E. Booij & Jaap Van Marle (eds.), *Yearbook of morphology 2001*, 203–235. Dordrecht: Kluwer Academic. https://doi.org/10.1007/978-94-017-3726-5_8.

Hollander, Myles & Douglas A. Wolfe. 1973. *Nonparametric Statistical Methods*. New York: John Wiley & Sons.

Horst, Marlise & Laura Collins. 2006. From faible to strong: How does their vocabulary grow? *Canadian Modern Language Review* 63(1). 83–106. https://doi.org/10.1353/cml.2006.0046

Kimppa, Lilli , Yury Shtyrov, Suzanne C.A. Hut, Laura Hedlund, Miika Leminen & Alina Leminen. 2019. Acquisition of L2 morphology by adult language learners. *Cortex* 116. 74–90. https://doi.org/10.1016/j.cortex.2019.01.012

Lahuerta, Ana Cristina. 2018. Study of accuracy and grammatical complexity in EFL writing. *International Journal of English Studies* 18(1). 71–89. https://doi.org/10.6018/ijes/2018/1/258971

Laufer, Batia, Stuart Webb, Su Kyung Kim & Beverley Yohanan. 2021. How well do learners know derived words in a second language? The effect of proficiency, word frequency and type of affix. *ITL – International Journal of Applied Linguistics* 172(2). 229–258. https://doi.org/10.1075/itl.20020.lau

Laws, Jacqueline & Chris Ryder. 2014. Getting the measure of derivational morphology in adult speech a corpus analysis using MorphoQuantics. *University of Reading Language Studies Working Papers* 6. 3–17. http://morphoquantics.co.uk/Resources/Laws%20&%20Ryder%20(2014).pdf (accessed 25.06.2021)

Leontjev, Dmitri. 2016. L2 English derivational knowledge: Which affixes are learners more likely to recognise? *Studies in Second Language Learning and Teaching* 6(2). 225–248. https://doi.org/10.14746/ssllt.2016.6.2.3

Lyashevskaya, Olga, Irina Panteleev & Olga Vinogradova. 2021. Automated assessment of learner text complexity. *Assessing Writing* 49, article 100529. https://doi.org/10.1016/j.asw.2021.100529

Lyashevskaya, Olga, Olga Vinogradova & Anna Scherbakova. (forthc.) Accuracy, syntactic complexity, and task type at play in examination writing: A corpus-based study. In Agnieszka Leńko-Szymańska & Sandra Götz (eds.), *Complexity, accuracy, and fluency in learner corpus research*.

Marchand, Hans. 1969. *The Categories and Types of Present-Day English Word-Formation*. 2nd ed. Munich: C. H. Beck.

Nation, Paul. 2021. Thoughts on word families. *Studies in Second Language Acquisition* 43(5). 969–972. https://doi.org/10.1017/S027226312100067X

Norris, John & Lourdes Ortega. 2009. Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. *Applied Linguistics* 30(4). 555–578. https://doi.org/10.1093/applin/amp044

Plakans, Lia, Atta Gebril & Zeynep Bilki. 2019. Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing* 36(2). 161–179. https://doi.org/10.1177/0265532216669537

Plag, Ingo, Christiane Dalton-Puffer & Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language & Linguistics* 3(2). 209–228.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org (accessed 25.06.2021).

Robinson, Peter. 2001. Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics* 22(1). 27–57. https://doi-org.proxylibrary.hse.ru/10.1093/applin/22.1.27

Robinson, Peter. 2011. Second language task complexity, the Cognition Hypothesis, language learning, and performance. In Peter Robinson (ed.), *Second language task complexity*: *Researching the Cognition Hypothesis of language learning and performance*, 3–39. Amsterdam: John Benjamins. https://doi.org/10.1075/tblt.2.05ch1

Skehan, Peter. 1998. *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Skehan, Peter. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30(4). 510–532. https://doi.org/10.1093/applin/amp047

Stein, Gabriele. 2007. *A Dictionary of English Affixes: Their Function and Meaning*. Munich: Lincom Europa.

Tywoniw, Rurik & Scott Crossley. 2020. Morphological complexity of L2 discourse. In Eric Friginal & Jack A. Hardy (eds.), *The Routledge handbook of corpus approaches to discourse analysis*, 269–297. London: Routledge. https://doi.org/10.4324/9780429259982-17

van der Slik, Frans, Roeland van Hout & Job Schepens. 2019. The role of morphological complexity in predicting the learnability of an additional language: The case of La (additional language) Dutch. *Second Language Research* 35(1). 47–70. https://doi.org/10.1177/0267658317691322

Vasylets, Olena, Roger Gilabert & Rosa M. Manchón. 2017. The effects of mode and task complexity on second language production. *Language Learning* 67(2). 394–430 https://doi.org/10.1111/lang.1222

Vinogradova, Olga, Olga Lyashevskaya & Irina Panteleeva. 2017. Multi-level student essay feedback in a learner corpus. In *Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue 2017*. 373–387. Moscow.

Yee, Thomas W. 2015. *Vector Generalized Linear and Additive Models*. Springer. https://doi.org/10.1007/978-1-4939-2818-7

Yoon, Hyung-Jo. 2017. Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System* 66. 130–141. https://doi.org/10.1016/j.system.2017.03.007

Supplementary materials

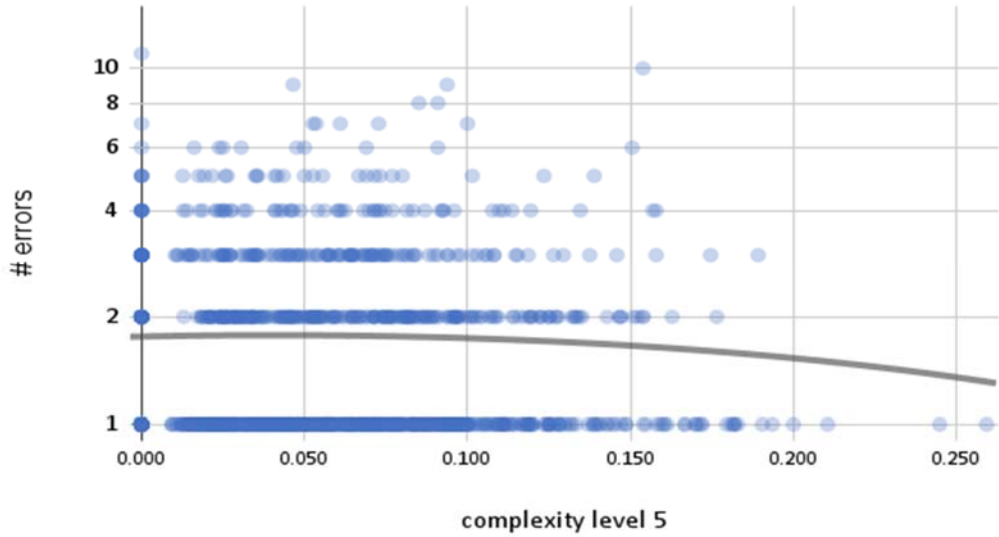

Figure B. Complexity at level 6 and number of errors (both task types)

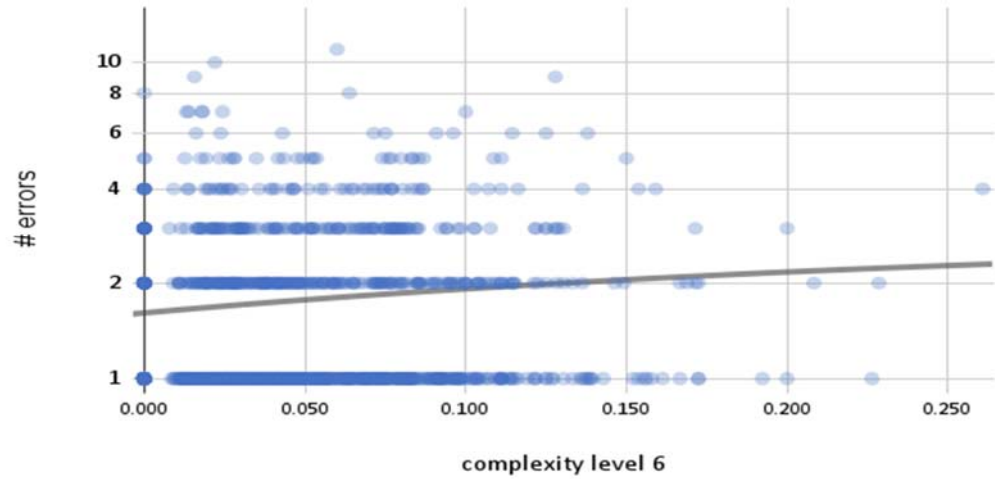

Figure A. Complexity at level 5 and number of errors (both task types)

Article history:
Received: 20 October 2021
Accepted: 08 February 2022

Bionotes:
Olga N. LYASHEVSKAYA is Professor at the School of Linguistics, National Research University "Higher School of Economics", Moscow, and a Senior Research Fellow at the Vinogradov Russian Language Institute, RAS, Moscow. Her research interests embrace semantics of grammar, lexical semantics, construction grammar, paleo-Slavic grammar, cognitive linguistics, corpus linguistics, lexicography, quantitative data analysis and computational linguistics. She coordinates the natural language processing and database projects of the Russian National Corpus team and PI in a number of lexical resource projects including Russian Framebank and Russian Constructicon.
Contact information:
National Research University "Higher School of Economics",
room 519, building A, 21/4, Staraya Basmannaya ul., Moscow, Russia
e-mail: olesar@yandex.ru
ORCID: 0000-0001-8374-423X

Julia V. PYZHAK is a student at the Department of Humanities, National Research University "Higher School of Economics", Moscow. Her research interests include corpus linguistics, machine learning and data analysis, as well as English as a foreign language.
Contact information:
National Research University "Higher School of Economics",
room 519, building A, 21/4, Staraya Basmannaya ul., Moscow, 117218, Russia
e-mail: jeneavas41@yandex.ru
ORCID: 0000-0003-3439-9788

Olga I. VINOGRADOVA is Associate Professor at the School of Linguistics, National Research University "Higher School of Economics", Moscow, and a Research Fellow at the Laboratory of Learner Corpora at the Department of Humanities, National Research University "Higher School of Economics", Moscow. Her areas of research are learner corpus linguistics, computer aided language learning, and corpus-based assessing writing, L2 acquisition and lexical typology. She leads the development of the learner corpus REALEC and a number of tools for automated assessment of learner text complexity and accuracy.
Contact information:
National Research University "Higher School of Economics".
room 519, building A, 21/4, Staraya Basmannaya ul., Moscow, 117218, Russia
e-mail: olgavinogr@gmail.com
ORCID: 0000-0001-5928-1482

Сведения об авторах:
Ольга Николаевна ЛЯШЕВСКАЯ – профессор Школы лингвистики Национального исследовательского университета «Высшая школа экономики», старший научный сотрудник Института русского языка имени В. В. Виноградова РАН (Москва). Ее исследовательские интересы охватывают широкий круг проблем, включая

грамматическую и лексическую семантику, грамматику конструкций, палеославянскую грамматику, когнитивную лингвистику, корпусную лингвистику, лексикографию, количественный анализ данных и компьютерную лингвистику. Она координирует проекты по разработке Национального корпуса русского языка, русского ФреймБанка, русского Конструктикона и т. д., а также является автором публикаций по когнитивной и функциональной лингвистике, анализу лингвистических данных и языковым технологиям.

***Контактная информация**:*
Национальный исследовательский университет «Высшая школа экономики»
Россия, 117218, Москва, Старая Басманная ул., 21/4, корпус А, комн. 519
e-mail: olesar@yandex.ru
ORCID: 0000-0001-8374-423X

Юлия Вячеславовна ПЫЖАК – студентка факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики», стажер-исследователь научно-учебной лаборатории учебных корпусов факультета гуманитарных наук. Ее исследовательские интересы включают корпусную лингвистику, машинное обучение и анализ данных, английский язык как иностранный.

***Контактная информация**:*
Национальный исследовательский университет «Высшая школа экономики»
Россия, 117218, Москва, Старая Басманная ул., 21/4, корпус А, комн. 519
e-mail: jeneavas41@yandex.ru
ORCID: 0000-0003-3439-9788

Ольга Ильинична ВИНОГРАДОВА – доцент Школы лингвистики Национального исследовательского университета «Высшая школа экономики», научный сотрудник научно-учебной лаборатории учебных корпусов факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики» (Москва). Область ее научных интересов включает корпусную лингвистику, компьютерные технологии в обучении языку, оценку письменных навыков корпусными методами, преподавание и освоение иностранных языков и лексическую типологию. Она руководит разработкой учебного корпуса REALEC и ряда компьютерных инструментов оценивания сложности и структуры ошибок в текстах, написанных изучающими язык.

***Контактная информация**:*
Национальный исследовательский университет “Высшая школа экономики”
Россия, 117218, Москва, Старая Басманная ул., 21/4, корпус А, комн. 519
e-mail: olgavinogr@gmail.com
ORCID: 0000-0001-5928-1482

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 493–514
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30084

Research article

Word frequency and text complexity: an eye-tracking study of young Russian readers

Antonina N. LAPOSHINA ✉, Maria Yu. LEBEDEVA and Alexandra A. BERLIN KHENIS

Pushkin State Russian Language Institute, Moscow, Russia
✉ANLaposhina@pushkin.institute

Abstract
Although word frequency is often associated with the cognitive load on the reader and is widely used for automated text complexity assessment, to date, no eye-tracking data have been obtained on the effectiveness of this parameter for text complexity prediction for the Russian primary school readers. Besides, the optimal ways for taking into account the frequency of individual words to assess an entire text complexity have not yet been precisely determined. This article aims to fill these gaps. The study was conducted on a sample of 53 children of primary school age. As a stimulus material, we used 6 texts that differ in the classical Flesch readability formula and data on the frequency of words in texts. As sources of the frequency data, we used the common frequency dictionary based on the material of the Russian National Corpus and DetCorpus – the corpus of literature addressed to children. The speed of reading the text aloud in words per minute averaged over the grades was employed as a measure of the text complexity. The best predictive results of the relative reading time were obtained using the lemma frequency data from the DetCorpus. At the text level, the highest correlation with the reading speed was shown by the text coverage with a list of 5,000 most frequent words, while both sources of the lists – Russian National Corpus and DetCorpus – showed almost the same correlation values. For a more detailed analysis, we also calculated the correlation of the frequency parameters of specific word forms and lemmas with three parameters of oculomotor activity: the dwell time, fixations count, and the average duration of fixations. At the word-by-word level, the lemma frequency by DetCorpus demonstrated the highest correlation with the relative reading time. The results we obtained confirm the feasibility of using frequency data in the text complexity assessment task for primary school children and demonstrate the optimal ways to calculate frequency data.
Keywords: *text complexity, text readability, word frequency, eye tracking*

For citation:
Laposhina, Antonina N., Maria Yu. Lebedeva & Alexandra A. Berlin Khenis. 2022. Word frequency and text complexity: An eye-tracking study of young Russian readers. *Russian Journal of Linguistics* 26 (2). 493–514. (In Russian). https://doi.org/10.22363/2687-0088-30084

Научная статья

Влияние частотности слов текста на его сложность: экспериментальное исследование читателей младшего школьного возраста методом айтрекинга

А.Н. ЛАПОШИНА ✉, М.Ю. ЛЕБЕДЕВА, А.А. БЕРЛИН ХЕНИС

Государственный институт русского языка имени А.С. Пушкина, Москва, Россия
✉ANLaposhina@pushkin.institute

Аннотация
Параметр частотности слова во многих исследовательских трудах связывается с когнитивной нагрузкой на читателя и широко используется в автоматических системах анализа сложности текста. Однако к настоящему моменту для русскоязычного материала не представлено достаточное количество экспериментальных данных о влиянии параметра частотности слов на сложность текста, собранных с помощью метода айтрекинга. Кроме того, не определены оптимальные способы учета частотности отдельных слов для характеристики целого текста. Целью данной статьи является заполнение этих лакун. Исследование проводилось на выборке 53 детей младшего школьного возраста. Материалом для эксперимента выступили 6 текстов, отличающихся по параметрам классической формулы читабельности Флеша и данным о частотности слов в текстах. В качестве источников данных о частотности слов использованы как стандартный частотный словарь на материале Национального корпуса русского языка, так и корпус литературы, адресованной детям, ДетКорпус. В качестве меры сложности текста использовался параметр скорости чтения текста вслух в словах в минуту, усредненный по классам. Для более детального анализа были произведены подсчеты корреляции параметров частотности конкретных словоформ и их лемм с тремя параметрами глазодвигательной активности: средней относительной скорости чтения слова, средней длительности фиксаций и средним количеством фиксаций. На пословном уровне анализа наивысший коэффициент корреляции с относительным временем чтения продемонстрировали данные частотности леммы по корпусу детской литературы. На уровне анализа текстов наиболее высокую корреляцию со средним временем чтения фрагмента показал параметр процента покрытия текста списком 5 000 самых частотных слов, при этом данные по разным источникам показали близкие значения. Приведенные результаты айтрекингового эксперимента подтверждают связь сложности текста и частотности входящих в него слов на материале для младших школьников, а также обозначают оптимальную методику и источники подсчета частотности для данной задачи.

Ключевые слова: *сложность текста, читабельность текста, частотность слова, айтрекинг*

Для цитирования:
Лапошина А.Н., Лебедева М.Ю., Берлин Хенис А.А. Влияние частотности слов текста на его сложность: экспериментальное исследование читателей младшего школьного возраста методом айтрекинга. *Russian Journal of Linguistics.* 2022. Т. 26. № 2. С. 493–514. https://doi.org/10.22363/2687-0088-30084

1. Введение

Тенденция к постоянному росту объема информации, характерная для современного этапа развития человечества, в полной мере касается и учебной информации. Новые учебники и пособия, образовательный контент, размещаемый на различных цифровых платформах, требуют тщательной и одновременно быстрой оценки с точки зрения их доступности читателям определенного возраста. Проблема предиктивной оценки сложности учебных материалов стоит особенно остро для категории детей младшего школьного возраста: в это время ребенок не только осваивает технику чтения, но и формирует читательскую грамотность – способность к осмыслению письменных текстов и эффективному взаимодействию с ними. Критически важно, чтобы в этот период человек не сталкивался с нецелесообразными сложностями при чтении.

За более чем столетнюю историю разработки формальных способов оценки сложности текста исследовались самые разные группы признаков текста – лексические, морфологические, синтаксические, семантические, способные указывать на его уровень сложности (DuBay 2007). В качестве базовых и наиболее распространенных рассчитываемых показателей сложности текста выступают значения средней длины слова и средней длины предложения. Логика подсчета таких признаков может варьироваться: кроме длины в знаках, есть формулы, учитывающие длину слова в слогах или процент слов длиннее заданного количества слогов. К настоящему моменту предложено огромное количество подобных формул для разных групп читателей, а также для русскоязычных учебных материалов (Солнышкина, Кисельников 2015, Криони и др. 2008, Шпаковский 2008) и сервисов по автоматизированной оценке сложности текста[1]. При этом такие формальные показатели имеют существенные ограничения.

В качестве аргумента против использования подобных формул часто указывается их низкая интерпретируемость (Мизернов, Гращенко 2015) и игнорирование лексики, составляющей текст (Graesser et al. 2011), – очевидно, что семантическая сложность слова не всегда коррелирует с его длиной (ср., например, слова *здравствуйте* и *штифт*). Чтобы снять это ограничение оценки сложности текста базовыми подсчетами, предлагается учитывать лексическую и семантическую сложность слов, входящих в текст. Одним из наиболее распространенных способов расчета сложности лексики являются подсчеты процента лексики из специально составленных списков «простых»

[1] Аналитик чтения. http://read-analytic.ru (дата обращения: 20.12.2021); Оценка читабельности текста. http://ru.readability.io (дата обращения: 20.12.2021)

слов (например, Chall & Dale 1995). В отечественных классических трудах 1970-х гг., посвященных сложности русских учебных текстов, Я.А. Микк предлагает учитывать «знакомость» слова: количество знакомых слов в тексте, которое определяется испытуемыми эмпирически по пятибалльной шкале (5 – очень хорошо знакомое слово, 0 – незнакомое слово) (Микк 1970).

С развитием корпусной лингвистики и появлением больших корпусов текстов для получения статистики стало возможным объективизировать интуитивное понятие «знакомости» слова с помощью данных о его частотности по релевантным большим коллекциям текстов. Основная идея такого подхода заключается в том, что частотные слова воспринимаются и производятся быстрее, чем редкие (Raney & Rayner 1995): возвращаясь к приведенному выше примеру, все формы слова *здравствовать* встречаются в основном корпусе НКРЯ 17 965 раз, а слова *штифт* – 129 раз. Соответственно, текст с высокой долей частотных слов должен восприниматься лучше (Chen & Meurers 2018). Частотность слова в качестве одного из признаков, оказывающих влияние на сложность, широко используется как в исследованиях для англоязычных материалов (Lexile 2007, Graesser et al. 2014), так и для текстов на русском языке (Solovyev et al. 2018, Glazkova et al. 2021, Иомдин, Морозов 2021). С другой стороны, исследователи указывают на отсутствие значимой корреляционной связи частотности и сложности текста, оцениваемой с помощью классической формулы Флеша-Кинкейда (Мартынова и др. 2020). Следовательно, экспериментальное установление связи (или ее отсутствия) частотности слова со сложностью на материале текстов на русском языке представляется актуальной исследовательской задачей.

Целью данного исследования является экспериментальное уточнение связи параметров движений глаз и частотности слова на материале русскоязычных текстов для младших школьников и определение оптимальной методики учета информации о частотности.

В ходе исследования мы отвечаем на следующие исследовательские вопросы:

1) Оказывает ли параметр частотности слова значимое влияние на сложность текста на русском языке для учеников младших классов?

2) Какой источник информации о частотности слова наиболее релевантен для этой задачи?

3) Какая из предлагаемых методик подсчета данных о частотности – среднее от логарифма нормализованной частотности всех слов текста, процент покрытия текста частотными списками объемом 5 000 слов по двум указанным источникам, процент слов в тексте со значением ipm ниже 5 – оптимальна для поставленной цели?

2. Частотность слова как параметр оценки сложности текста: теоретическое обоснование

Помимо установления связи параметра частотности слов и сложности составленного из них текста отдельную исследовательскую проблему

представляет выбор источника данных о частотности лексики, релевантного выбранной возрастной категории, так как данные о частотности слова сильно зависят от типа и наполнения корпуса, по которому ведутся подсчёты (Ляшевская, Шаров 2009, предисловие к словарю). Ряд исследователей используют для этих целей данные больших национальных корпусов текстов (Dorofeeva et al. 2019, Glazkova et al. 2021, Иомдин, Морозов 2021). Аргументами в пользу этого выбора могут служить большой размер таких корпусов, а также представленность в их составе различных жанров «официального» кодифицированного языка, с которым учащимся предстоит столкнуться в жизни: художественная литература, новости, публицистика – всё это составляет основу данных.

Однако, с другой стороны, остается открытым вопрос о правомерности сравнения материалов для детей с языком «взрослых». Например, ряд слов, высокочастотных по данным Национального корпуса русского языка (НКРЯ), полностью отсутствует в текстах учебников младших классов (*цена, проблема, государство*). Напротив, относительная встречаемость другой группы слов (*лес, птица, мороз)* значительно выше в текстах учебников, чем в коллекции текстов НКРЯ (Лапошина et al. 2019). Поэтому альтернативным решением здесь может стать подсчет частотности по специальным коллекциям текстов, предназначенным для детей (Lexile 2007).

Наконец, существуют разные способы учета частотной информации для текстовых фрагментов. Классический способ, представленный еще в ранних формулах читабельности, предлагает расчет процента слов текста, входящих в релевантный список слов, одной из разновидностей которого может стать частотный список. Этот метод расчета и сейчас используется в ряде исследований сложности текста (Glazkova et al. 2021, Sato 2014). Ещё один популярный способ учета частотности слов текста – это расчет среднего или медианного значения из частотности каждого слова текста (Francois & Fairon 2012, Reynolds 2016). Система анализа сложности англоязычных текстов Lexile предлагает средний логарифм нормализованной частотности всех слов текста как меру, демонстрирующую наивысшую корреляцию со сложностью, однако для расчетов используют не стандартную меру количества вхождений на миллион, ipm (instances per million), а на 5 миллионов слов (Lexile 2007). В работе, посвященной сравнению метрик частотности для текстов на английском языке с помощью построения предсказательных моделей, лучший результат показала модель, основанная на двух показателях: среднего логарифма с основанием 10 от нормализованной частотности слова на миллиард слов и стандартного отклонения (Chen & Meurers 2016). В работе на материале русского языка, посвященной диагностической валидности Стандартизированной методики исследования навыков чтения на русском языке, предлагается в качестве метрики средняя нормализованная частотность (ipm) только полнозначных слов текста (Dorofeeva et al. 2019). Часть исследований вообще использует не числовые значения, а систему деления текстов на группы по

частотности входящей в них лексики: «тексты с частотными словами/тексты с нечастотными словами» (Rello et al. 2013). Таким образом, представляется перспективным сравнение способов выражения данных о частотности на материале целых текстов и выбор оптимальной методики для данной задачи.

Метод бесконтактной регистрации движений глаз (айтрекинг) является одним из наиболее точных и трудозатратных способов проверки механизмов восприятия текста, который активно используется уже более 25 лет (Rayner 1998). Он позволяет отслеживать непроизвольное направление взгляда с высокой точностью (вплоть до доли градуса) и временным разрешением (вплоть до сотых долей секунды). Данная технология применима как для экспериментальной проверки гипотез о влиянии лингвистических и паралингвистических параметров текста на его сложность, так и для обнаружения различных затруднений при чтении. Чаще всего применительно к исследованиям чтения используются такие глазодвигательные параметры, как средняя продолжительность фиксаций, относительное время чтения слова и среднее количество фиксаций. Средняя продолжительность фиксаций (fixation duration) отражает время остановки взора на конкретном слове, и, чаще всего, характеризует скорость лексической активации и восприятия прочитанного (Henderson et al. 1989, Rayner 1998). Относительное время чтения слова (dwell time, %) является показателем процентного отношения суммы всех фиксаций на слове к другим слова в тексте (Griffin & Spieler 2006). Среднее количество фиксаций (fixation count) отражает количество фиксаций на слове, дополняя понимание о стратегии чтения (Clifton et al. 2007).

В современной науке накоплена большая база данных и установлены связи между определенными параметрами движений глаз и сложностью текста (Jian & Ko 2017). Сложность текста может быть представлена в различных лексических параметрах слов. Широко изучено влияние таких факторов, как длина слова (Rayner 2011), регулярность и согласованность слова (Farris-Trimble et al. 2018), орфографические и фонологические характеристики слова (Tiffin-Richards & Schroeder 2015), семантическое разнообразие значений слова (Luke et al. 2015) и прочее.

В контексте нашего исследования особый интерес представляют работы, оценивающие изменения в глазодвигательном поведении в зависимости от частотности слова (Rello et al. 2013, White et al. 2018). Так, в ряде исследований реципиенты демонстрировали значительно большую продолжительность взгляда на низкочастотных словах, чем на высокочастотных (Rau et al. 2014). Есть также свидетельства о большем влиянии фактора длины слова на параметры движения глаз у детей при чтении слов с низкой частотностью в сравнении с чтением слов с высокой частотностью (Rau et al. 2015). В работе на русскоязычном материале в записью движений глаз, техника чтения оценивалась по методикам «Чтение регулярных и нерегулярных слов» Т.В. Ахутиной и «Стандартизованная методика исследования навыка чтения» (СМИНЧ) А.Н. Корнева и О.А. Ишимовой (Корнеев и др. 2019). В описании обеих групп

этих методических материалов частотность указывается в качестве одной из характеристик, однако не поясняется источник данных о частотности слова и методика подсчета этой метрики для целого текста.

3.Материалы и методы

Участники эксперимента. Для установления зависимости частотности слов текста на русском языке с его сложностью был проведен эксперимент, в котором приняли участие 53 ученика 1–3 классов средних школ города Москвы: 26 учеников 1 класса (10 мальчиков, 16 девочек), 15 учеников 2 класса (4 мальчика, 11 девочек), 12 учеников 3 класса (2 мальчика, 10 девочек). Исследования проводились в апреле и мае, в конце учебного года, когда предполагается освоение навыков чтения, соответствующих классу обучения.

В качестве ***материала для эксперимента*** были использованы 6 текстов из современных учебников русского языка для 2–3 класса (табл. 1). В некоторых случаях текст незначительно модифицировался: слова заменялись на синонимы для получения более контрастирующих значений длины и частотности.

Методика подсчета данных о частотности лексики русского языка. Для определения оптимального источника информации о частотности слова, релевантного задаче оценки сложности текста для младшей школы, мы использовали два источника информации: Частотный словарь современного русского языка (по материалам Национального корпуса русского языка)[2] (далее – ЧС НКРЯ) и корпус детской литературы[3] (далее – ДетКорпус).

Частотный словарь основан на выборке текстов Национального корпуса русского языка объемом 100 млн словоупотреблений и включает в себя 20 тысяч наиболее употребительных слов современного русского языка (2-я половина XX – начало XXI вв.). Для получения частотных данных использовалась мера нормализованной частоты (ipm) общего частотного списка лемм. В данном словаре снята омонимия, поэтому частоты для разных значений омонима приводятся отдельно.

ДетКорпус – это аннотированный корпус русской детской литературы, включающий более 2097 прозаических произведений, написанных на русском языке в период с 1920-х по 2010-е гг. и адресованных детям и подросткам. Корпус содержит как художественные тексты различных жанров (реализм, приключения, детектив, ужастик), так и текст нон-фикшн. В данной коллекции текстов омонимия не снята. Поэтому в дальнейших пословных подсчетах мы не учитывали многозначные слова. Примеры анализируемых слов в табл. 2 дают представление о возможных различиях в частотных данных в зависимости от выбранного корпуса. Так, частотность тематически и

[2] *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

[3] ДетКорпус. http://detcorpus.ru (дата обращения: 20.12.2021).

стилистически нейтральной леммы *себя* показывает весьма близкие между собой значения по указанным коллекциям текстов, лемма *мальчик* значительно чаще встречается в корпусе детской литературы, а лемма *современный,* более характерная для документальной прозы и публицистики, намного частотнее в ЧС НКРЯ. Частотные списки 5 000 самых частотных слов по этим двум источникам, которыми мы будем оперировать далее, пересекаются на 81%.

Для учета частотных данных для целых текстов были рассчитаны следующие метрики:

1. Процент покрытия текста списком 5000 самых частотных слов;
2. Процент слов в тексте со значением ipm ниже 5;
3. Среднее от логарифма нормализованной частотности всех слов текста;
4. Средняя относительная частотность (ipm) полнозначных слов текста.

Анализ результатов показал, что метрика № 4 на текстах небольшого объема показывает слишком сильный разброс: например, значение для текста «В траве» равно 42, а для текста «Собака» – 1545. Вероятнее всего, это связано с тем, что в текстах малого объема появление хотя бы нескольких слов из 100 самых частотных для русского языка (*она, человек, такой* и т.п.*)* и отсутствие логарифмирования приводят к увеличению среднего значения частотности в разы. Поскольку такие данные трудно интерпретировать в дальнейшем, мы отказались от этой метрики.

В качестве ***традиционной метрики сложности текста***, основанной на средних значениях длины и предложения, был использован индекс читабельности Флеша со скорректированными для русского языка коэффициентами (Оборнева 2006).

$$\text{FRE(Oborneva)} = 206{,}835 - 1{,}52 \times \text{ASL} - 65{,}14 \times \text{ASW},$$

где FRE(Oborneva) – оценка читабельности текста; AWL – среднее число слогов в слове; ASL – среднее количество слов в предложении. Результатом этой формулы является число от 0 до 100, где 100 – очень легкий текст, 65 – легкий текст, 30 – немного трудно читать, а 0 – очень сложный для чтения текст.

В табл. 1 для каждого отобранного текста представлены данные о лингвистических параметрах, влияние которых на сложность исследовалось в ходе эксперимента. Первые две строки табл. 1 содержат параметры, по которым контрастируют отобранные текста. Тексты № 1 и № 2 уравновешены по проценту частотных слов, но отличаются по индексу FRE(Oborneva), тексты № 3 и № 4, наоборот, контрастируют по частотным данным, текст № 5 является примером текста, определенного как сложный и по индексу FRE(Oborneva) и по количеству частотных слов, текст № 6, напротив, пример простого текста и по индексу FRE(Oborneva), и по количеству частотных слов.

Таблица 1. **Основные лингвистические параметры используемых в эксперименте текстов**

Параметр текста	Текст 1. Трактор	Текст 2. Умка	Текст 3. В траве	Текст 4. Мышка	Текст 5. Цветы	Текст 6. Собака
FRE (Oborneva)	49	78	75	66	15	80
Покрытие текста частотным списком 5 000 (ЧС НКРЯ)	84%	85%	35%	89%	62%	92%
Средняя длина слова	6.4	4.8	5.6	5.6	6.8	4.1
Покрытие текста частотным списком 5 000 (ДетКорпус)	81%	83%	61%	93%	69%	96%
Процент слов с ipm < 5 (ЧС НКРЯ)	8%	14%	43%	4%	21%	2%
Процент слов с ipm < 5 (ДетКорпус)	11%	12%	22%	4%	24%	0%
Среднее от логарифма ipm всех слов текста (ЧС НКРЯ)	4.5	4.2	3.6	4.7	3.8	4.9
Среднее от логарифма ipm всех слов текста (ДетКорпус)	4.2	4.6	3.8	4.8	4	4.9

Table 1. **Main linguistic parameters of the texts used in the experiment (FD RNC is a frequency dictionary based on Russian National Corpus, DetCorpus is a corpus of literature addressed to children)**

Параметр текста	Text 1. Tractor	Text 2. Umka	Text 3. In the grass	Text 4. Mouse	Text 5. Flowers	Text 6. Dog
FRE (Oborneva)	49	78	75	66	15	80
Text coverage by the list 5000 (FD RNC)	84%	85%	35%	89%	62%	92%
Average word length	6.4	4.8	5.6	5.6	6.8	4.1
Text coverage by the list 5000 (DetCorpus)	81%	83%	61%	93%	69%	96%
Percent of words with ipm < 5 (FD RNC)	8%	14%	43%	4%	21%	2%
Percent of words with ipm < 5 (DetCorpus)	11%	12%	22%	4%	24%	0%
Average log word frequency (FD RNC)	4.5	4.2	3.6	4.7	3.8	4.9
Average log word frequency (DetCorpus)	4.2	4.6	3.8	4.8	4	4.9

Пословный анализ. Для более детального анализа и выбора оптимального источника данных о частотности и методике подсчета, были также произведены подсчеты корреляции лингвистических параметров конкретных слов с данными глазодвигательной активности. Каждая словоформа длиной более 3 символов была размечена по длине в символах и в слогах, частотности конкретной словоформы и частотности леммы (табл. 2). В анализе не участвовали многозначные слова. Поскольку слово в тексте может встретиться в непривычной, редкой форме, в ходе эксперимента отдельно проверялось влияние частотности конкретной словоформы на параметры глазодвигательной активности респондентов. Так, частотность всех форм существительного

волна составляет 31 662 вхождения, тогда как конкретная словоформа, представленная в тексте эксперимента, *волною*, встретилась в корпусе лишь 279 раз, поэтому все подсчеты были выполнены отдельно по совокупности частот всех словоформ этой лексемы (леммы *волна*) и отдельно по словоформе (*волною).*

Таблица 2. **Пословные параметры длины, частотности и характеристик движений глаз**

Словоформа	мальчики	гладиолусов	себе	современный
Лемма	мальчик	гладиолус	себя	современный
Длина словоформы в знаках	8	11	4	11
Длина словоформы в слогах	2	4	2	4
Частотность леммы в ЧС НКРЯ, ipm	188	0	2272	236
Частотность леммы в ДетКорпусе, ipm	597	1.1	2243	14
Частотность словоформы в ЧС НКРЯ, ipm	19	0	90	33
Частотность словоформы в ДетКорпусе, ipm	91	0.4	86	4
Относительное время чтения слова	0.026	0.089	0.019	0.032
Средняя длительность фиксаций, мс	257	288	255	250
Среднее количество фиксаций	3.22	9.15	2.46	4.43

Table 2. **Word-by-word values of word length, frequency and eye movement parameters (FD RNC is a frequency dictionary based on Russian National Corpus, DetCorpus is a corpus of literature addressed to children)**

Word form	мальчики	гладиолусов	себе	современный
Lemma	мальчик	гладиолус	себя	современный
Length of word form in characters	8	11	4	11
Length of word form in syllables	2	4	2	4
Lemma frequency by FD RNC, ipm	188	0	2272	236
Lemma frequency by DetCorpus, ipm	597	1.1	2243	14
Word form frequency by FD RNC, ipm	19	0	90	33
Word form frequency by DetCorpus, ipm	91	0.4	86	4
Dwell time, %	0.026	0.089	0.019	0.032
Fixation duration, ms	257	288	255	250
Fixation count	3.22	9.15	2.46	4.43

В ***ходе эксперимента*** испытуемых просили вслух и с максимальной скоростью прочитать предъявляемые тексты и предупреждали, что после прочтения будут заданы вопросы на понимание. Параллельно велась аудиорегистрация чтения и ответов на вопросы. Перед чтением каждого текста участник отвечал на вопрос, знаком ли ему предложенный текст. Учеников случайным образом разделили на 2 равные группы. Для уменьшения утомления каждая из групп читала только 3 из 6 отобранных текстов в случайном порядке, а также первый «тренировочный» текст, данные которого впоследствии не учитывались. Рис. 1 иллюстрирует пример результата прочтения текста одним из

испытуемых: точками обозначены фиксации, линии демонстрируют траекторию перемещения взгляда при чтении.

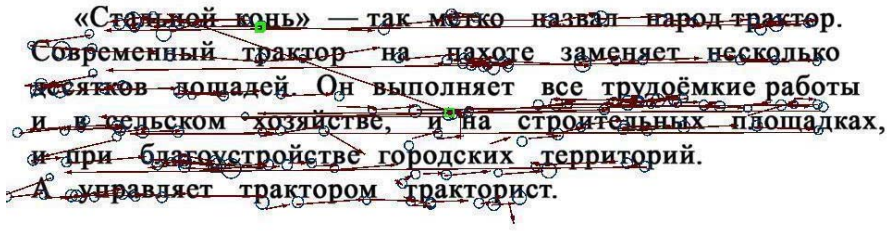

Рис. 1. Пример анализируемых данных глазодвигательной активности / Pic. 1. An example of the analyzed data of oculomotor activity

Оборудование. Исследование проводилось с применением айтрекера SR Research Eyelink 1000+, с частотой регистрации 500 Гц и 13-точечной калибровкой перед началом эксперимента. Испытуемые садились перед экраном компьютера диагональю 23 дюйма, с разрешением 1920 на 1080 точек (расстояние между глазами и экраном 940 мм), голова фиксировалась с помощью лобной опоры. В центре экрана предъявлялись отобранные тексты в той же верстке, которая использовалась в исходных учебниках, в виде изображения шириной 1400 пикселей и соответствующей тексту высотой. Это обеспечивало соответствие угловых размеров текста таковым при чтении учебника в привычном положении.

Методика подсчетов. В качестве меры сложности целого текста использовался параметр скорости чтения текста вслух в словах в минуту, усредненный по классам. При пословном анализе в качестве показателей сложности использовались значения средней относительной скорости чтения слова (dwell time %) – эта величина показывает, какую часть от времени прочтения всего текста конкретным испытуемым занимает чтение данного слова; средней длительности фиксаций (fixation duration); средним количеством фиксаций (fixation count).

4. Результаты

4.1. Анализ на уровне текстов

Для всех текстов ожидаемо наблюдалось увеличение скорости чтения от первого к третьему классу, хотя для разных текстов средние скорости были различны (Рис. 2). Следует отметить, что часть учеников 1 класса, участвовавших в исследовании, уже были знакомы с текстом «Собака», что дополнительно могло повысить скорость чтения этого фрагмента. Все остальные тексты были отмечены учениками как незнакомые.

Для оценки влияния класса обучения и конкретного текста на скорость чтения был проведен двухфакторный дисперсионный анализ. Результаты анализа показали, что оба этих фактора играют статистически значимую роль и

не зависят друг от друга, то есть «простой» или «сложный» текст оставался таковым независимо от класса обучения (см. рис. 1) (ANOVA, фактор «класс» ($F(2,135) = 28{,}55$, $p < 0{,}0001$) и «текст» ($F(5,135) = 8{,}40$, $p < 0{,}0001$).

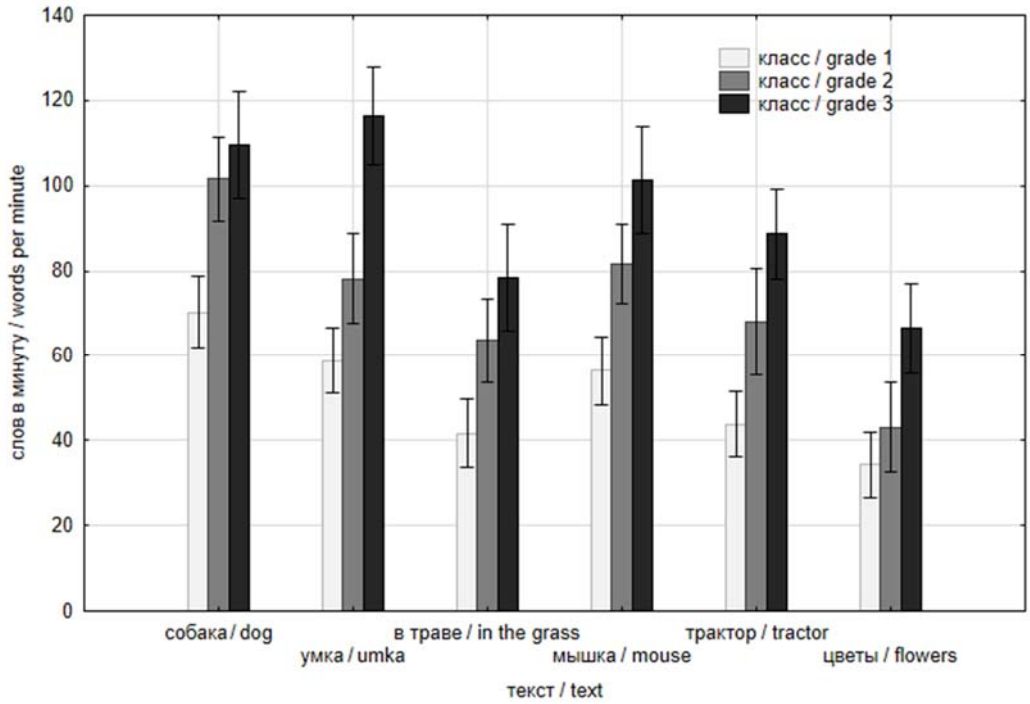

Рис. 2. Средние скорости чтения исследуемых текстов учениками 1–3 классов / Fig. 2. Average reading speed of the texts by students of grades 1–3

Анализ связи скорости чтения с характеристиками текстов иллюстрирует, что значения классической формулы Флеша не всегда способны адекватно предсказывать скорость чтения (рис. 2). Так, первая контрастирующая пара текстов с близкими показателями частотной лексики, но отличающаяся значениями формулы FRE (Oborneva), – «Умка» («простой» текст) и «Трактор» («сложный» текст), демонстрирует стабильную связь скорости чтения и показателей формулы: средняя скорость «простого» текста заметно выше по всем трем классам. Вторая пара текстов, имеющая близкие значения формулы FRE (Oborneva), соответствующие «простым» текстам, и контрастирующая по проценту покрытия текста частотным списком ЧС НКРЯ, иллюстрирует интересный случай расхождения показателей сложности по формуле и частотным данным. Так, на рис. 2 тексты расположены в порядке возрастания сложности по формуле FRE (Oborneva). Если допустить, что такой расчет точно измеряет сложность текста, мы ожидаем увидеть на графике плавное падение скорости чтения. Однако обратим внимание, что скорость чтения текста «В траве» (пример 1) заметно ниже ожидаемой.

(1) В траве трещат кузнечики, скрипит жук. Воркуют дикие голуби. Стучат по деревьям дятлы, пищат рябчики. Жужжит золотая пчёлка. Поют певчие дрозды, трещит сойка.

Этот текст действительно состоит из очень коротких слов и предложений, что и определило его значение формулы FRE (Oborneva), соответствующее самому простому тексту из всей коллекции. Однако при этом текст имеет наименьший процент покрытия списками частотной лексики. Можно предположить, что в данном случае речь идет о повышении сложности текста исключительно за счет трудности лексического состава текста. И оценка его сложности с помощью традиционных показателей (формуле FRE, основанной на длине слов и предложений) не соответствует его наблюдаемой сложности, выраженной средней относительной скоростью чтения.

Таким образом, были экспериментально зафиксированы две глобальные возможные причины повышения сложности текста, выраженные в снижении средней скорости чтения: структурная сложность, связанная с длиной слов и предложений («Трактор»), и лексическая сложность («В траве»). Быстрее всего испытуемые справлялись с текстами, отнесенными по обеим группам параметров к «простым»: «Собака», «Умка» и «Мышка». Эти тексты отличаются короткими словами и частотной лексикой. А наименьшая скорость чтения была показана на тексте «Цветы», содержащем и длинные слова, и крайне нечастотную лексику (*хризантема – 1.2 ipm; клубень – 3 ipm; гладиолус – 1.1 ipm; георгин – 1.4 ipm*)[4].

Результаты корреляционного анализа параметров частотности слов текста, полученных различными методиками, с параметром средней по всем классам скорости чтения исследуемых текстов (табл. 3), демонстрируют наивысшую корреляцию средней относительной скорости чтения с метриками текстов, представляющими процент покрытия текста частотными списками. Тип источника данных о частотности – ЧС НКРЯ или ДетКорпус – не всегда играет значимую роль в исследуемых текстах: при подсчетах среднего логарифма частотности слов текста и процента слов с ipm ниже 5 коэффициент корреляции оказывается выше у подсчетов по данным ДетКорпуса; процент же покрытия текстов частотными списками объемом 5000 слов по двум исследуемым источникам показывает одинаковый коэффициент детерминации. Интересно также, что индекс FRE (Oborneva) не показал статистически значимой корреляции со временем чтения в изучаемых текстах.

Малый объем текстов, на материале которых трудно делать общие выводы, стоит отнести к ограничениям эксперимента, связанным со спецификой процесса записи данных глазодвигательной активности. В айтрекинговых экспериментах текст выводится на экране, расположенном относительно далеко от глаз участника исследования. Для получения достаточно точных для анализа чтения результатов шрифт не может быть мелким, а текст не должен занимать значительную часть экрана, прокрутка также невозможна. Кроме

[4] Приведены расчеты частотности по леммам по данным ДетКорпуса.

того, возраст участников данного исследования также налагал ограничения: для детей 7–10 лет неподвижное сидение с фиксированным положением головы в сочетании с чтением на скорость утомительно. Поэтому весь эксперимент не мог длиться более 15 минут, включая калибровку, чтение тренировочного текста и ответы на вопросы.

Таблица 3. **Корреляционный анализ параметров глазодвигательной активности с параметрами частотности (Корреляция Спирмена, выделенные жирным значения с p-value < 0,05)**

Параметр	Средняя скорость чтения текста
Средняя длина слова	**-0.83**
FRE(Oborneva)	0.66
Покрытие текста частотным списком 5 000 (ЧС НКРЯ)	**0.89**
Покрытие текста частотным списком 5 000 (ДетКорпус)	**0.89**
Процент слов с ipm < 5 (ЧС НКРЯ)	-0.77
Процент слов с ipm < 5 (ДетКорпус)	**-0.83**
Среднее от логарифма ipm всех слов текста (ЧС НКРЯ)	**0.78**
Среднее от логарифма ipm всех слов текста (ДетКорпус)	**0.85**

Table 3. **Correlation analysis of oculomotor activity parameters with word frequency parameters (Spearman correlation, bold values have p-value <0.05)**

Parameter	Average reading speed
Average word length	**-0.83**
FRE(Oborneva)	0.66
Text coverage by the list 5000 (FD RNC)	**0.89**
Text coverage by the list 5000 (DetCorpus)	**0.89**
Percent of words with ipm < 5 (FD RNC)	-0.77
Percent of words with ipm < 5 (DetCorpus)	**-0.83**
Average log word frequency (FD RNC)	**0.78**
Average log word frequency (DetCorpus)	**0.85**

4.2. Результаты пословного анализа

Результаты корреляционного анализа (табл. 4) иллюстрируют степень связи параметров глазодвигательной активности с лингвистическими параметрами отдельных словоформ. Затемнённым фоном отмечены наиболее высокие значения корреляции в столбце. Максимально тесную связь с относительным временем чтения демонстрируют данные частотности леммы в ДетКорпусе, хотя все остальные варианты подсчета частотности слова по лемме и словоформе показывают очень близкие по значению результаты. Это говорит о том, что гипотеза о необходимости подсчета именно словоформ, а не лемм, на данных текстах не подтверждается. Возвращаясь к выбору оптимального источника данных о частотности, можно отметить, что, несмотря на то, что лучший результат показали данные по ДетКорпусу, разница не является существенной.

Таблица 4. **Корреляционный анализ параметров глазодвигательной активности с лингвистическими параметрами словоформ (Корреляция Спирмена, выделенные жирным значения имеют значение p-value < 0.05)**

Параметр	Среднее относительное время чтения	Средняя длительность фиксации	Среднее количество фиксаций
Длина словоформы в знаках	**0.53**	-0.02	**0.73**
Длина словоформы в слогах	**0.36**	-0.09	**0.55**
Частотность леммы в ЧС НКРЯ, ipm	**0.55**	**0.49**	**0.46**
Частотность леммы в ДетКорпусе, ipm	**0.59**	**0.42**	**0.54**
Частотность словоформы в ЧС НКРЯ, ipm	**0.58**	**0.47**	**0.53**
Частотность словоформы в ДетКорпусе, ipm	**0.58**	**0.42**	**0.53**

Table 4. **Correlation analysis of oculomotor activity parameters and linguistic parameters of word forms (Spearman correlation, bold values have a p-value <0.05)**

Parametr	Dwell time	Fixation duration	Fixation count
Length of word form in characters	**0.53**	-0.02	**0.73**
Length of word form in syllables	**0.36**	-0.09	**0.55**
Lemma frequency by FD RNC, ipm	**0.55**	**0.49**	**0.46**
Lemma frequency by DetCorpus, ipm	**0.59**	**0.42**	**0.54**
Word form frequency by FD RNC, ipm	**0.58**	**0.47**	**0.53**
Word form frequency by DetCorpus, ipm	**0.58**	**0.42**	**0.53**

Показательно, что длина слова в знаках сильнее всего коррелирует со средним количеством фиксаций, но не демонстрирует значимой связи со средней длительностью фиксаций, в отличие от частотных данных. Иными словами, от длины слова зависит, на какое количество «отрезков» глаз делит слово при чтении, тогда как то, сколько времени он задерживается на каждом таком отрезке, зависит именно от частотности слова. Можно предположить, что длительность фиксаций свидетельствует о когнитивных усилиях, требуемых для распознавания и обработки данного слова. Например, в выборке слов одинаковой длины в 5 знаков самые высокие значения длительности фиксаций занимают низкочастотные слова *сойка* (2.4 ipm / 307 мс.), *юркий* (4 ipm / 302 мс.), а самые низкие – частотные *разве* (240 ipm / 230 мс.) и *белые* (425 ipm / 234 мс.)[5].

Количество слогов в данном эксперименте показывает статистически значимую корреляцию со временем чтения и количеством фиксаций, но заметно меньшую, чем длина слов в знаках.

Также важно отметить, что на материале пословного анализа в совокупности отобранных для исследования текстов взаимная корреляция между параметрами частотности словоформы по НКРЯ и ее длины в знаках крайне слабая (корреляция Спирмена, $r = –0,21$, $p < 0,05$).

[5] Приведены расчеты частотности по леммам по данным ДетКорпуса.

4.3. Практическое применение результатов эксперимента

Полученные в ходе эксперимента данные, подтверждающие предиктивный потенциал информации о частотности слова, были использованы при создании пилотной версии системы автоматизированной оценки уровня сложности текста. Разработанный сервис Текстометр[6] при переключении в режим «русский язык как родной» предлагает оценку сложности текста по двум векторам: структурному и лексическому. Структурная сложность текста основывается на традиционном индексе FRE(Oborneva), приведенной для удобства интерпретации[7] к шкале возрастающей сложности от 0 до 10 по формуле:

$$\text{TEXTOMETR\_Struc} = \frac{100 - FRE(Oborneva)}{10}$$

Коэффициент лексической сложности текста основывается на результатах проведенного эксперимента и подсчитывается с помощью формулы:

$$\text{TEXTOMETR\_Lex} = 10 - \frac{(Freq - 50)}{5}$$

где Freq – процент покрытия всех слов текста списком 5 000 наиболее частотных лемм ДетКорпуса, показавший наивысшую корреляцию с относительным временем чтения текста, а остальные параметры являются константами. Результатом вычислений становится коэффициент лексической сложности текста, число от 0 до 10[8], означающее степень вероятной знакомости слов текста читателями детской аудитории. Предположительный возраст, для которого данный текст является оптимальным по сложности, рассчитывается на основании усредненной оценки двух описанных параметров. В качестве иллюстрации приведем значения параметров структурной и лексической сложности изученных текстов.

Как видно из табл. 5, почти все тексты отмечены программой как соответствующие возрасту учеников российской начальной школы. Исключение составляет лишь текст «Цветы» из-за входящих в его состав длинных и крайне нечастотных слов, которые привели к высоким показателям структурной и лексической сложности текста. Дополнительной причиной таких высоких показателей сложности может являться и небольшой объем текста, при котором появление даже нескольких длинных или низкочастотных слов может значительно повлиять на финальные оценки теста. Стоит упомянуть, однако, что данный текст действительно показал наименьшие значения усредненной скорости чтения по всем трем классам, что также свидетельствует о возникших трудностях при его чтении. Описанный выше нетипичный текст

6 Текстометр. https://textometr.ru (обращение 27.12.2021)

7 Оригинальный вариант коэффициента представляет собой число от от 0 до 100, причем коэффициент измеряет уровень простоты текста, т.е. меньшие значения означают большую сложность текста, что приводит к неудобству интерпретации.

8 Формально максимальным значением коэффициента является 20, однако подавляющее большинство русскоязычных текстов, включая новостные, научные, художественные тексты и определения сложных терминов, укладывается в шкалу от 0 до 10.

«В траве» оценивается по шкале структурной сложности в 2 балла из 10, тогда как по шкале лексической сложности – в 7 баллов из 10. Общая усредненная оценка остается при этом в границах возрастной группы младшей школы, однако такой способ оценки позволяет получить более полную и интерпретируемую информацию о сложности текста.

Таблица 5. **Пример работы сервиса Текстометр (русский язык как родной) на материале текстов из эксперимента**

Текст	Структурная сложность	Лексическая сложность	Предположительный возраст
Текст 1. Трактор	4	3	9–10 лет
Текст 2. Умка	3	3	9–10 лет
Текст 3. В траве	2	7	9–10 лет
Текст 4. Мышка	3	1	7–8 лет
Текст 5. Цветы	9	6	13–15 лет
Текст 6. Собака	2	1	7–8 лет

Table 5. **An example of the output of the Textometr tool (Russian as a native language section) for the texts from the experiment**

Text	Structural complexity	Lexical complexity	Estimated age
Text 1. Tractor	4	3	9–10 years
Text 2. Umka	3	3	9–10 years
Text 3. In the grass	2	7	9–10 years
Text 4. Mouse	3	1	7–8 years
Text 5. Flowers	9	6	13–15 years
Text 6. Dog	2	1	7–8 years

К ограничениям эксперимента, связанным со спецификой процесса записи данных глазодвигательной активности, стоит отнести малый объем текстов. В айтрекинговых экспериментах текст выводится на экране, расположенном относительно далеко от глаз участника исследования. Для получения достаточно точных для анализа чтения результатов шрифт не может быть мелким, а текст не должен занимать значительную часть экрана, прокрутка также невозможна. Кроме того, возраст участников данного исследования также налагал ограничения: для детей 7–10 лет неподвижное сидение с фиксированным положением головы в сочетании с чтением на скорость утомительно. Поэтому весь эксперимент не мог длиться более 15 минут, включая калибровку, чтение тренировочного текста и ответы на вопросы.

5. Выводы

Приведенные результаты айтрекингового эксперимента, осуществленного на материале текстов учебников русского языка для младшей школы, подтверждают связь сложности текста и частотности слов, в него входящих, и демонстрируют потенциал учета частотных данных в системах оценки сложности текста на русском языке.

На уровне текстов самую высокую корреляцию со средним временем чтения фрагмента показал параметр процента покрытия текста списком 5000

самых частотных слов, при этом данные по разным источникам – ЧС НКРЯ и ДетКорпусу – показали одинаковые значения. Также в результате эксперимента удалось выявить текст, где классическая формула Флеша, базирующаяся на средней длине слова и предложения, дает ошибку прогноза, тогда как данные о частотности слов текста верно диагностируют вероятную лексическую сложность текста. Целесообразно включение такого текста в коллекцию материалов для валидизации качества систем, оценивающих сложность текста на русском языке.

На пословном уровне самую тесную связь с относительным временем чтения продемонстрировали данные частотности леммы по корпусу детской литературы. Также анализ показал, что хотя длина слова в знаках сильнее всего коррелирует со средним количеством фиксаций, она не демонстрирует значимой связи со средней длительностью фиксаций. Напротив, частотность слова показала самую высокую корреляцию с этим параметром. Такие данные позволяют сделать предположение, что длина слова больше влияет на механическую часть процесса чтения, а именно на какое количество «отрезков» глаз делит слово для удобства прочтения, тогда как частотность слова оказывает влияние на когнитивный аспект чтения – на то, сколько времени потребуется на каждом таком отрезке для распознавания облика слова и его восприятия. При этом более трудозатратная методика подсчета частотности словоформ не дала значимого прироста качества в исследуемом материале, что позволяет сделать вывод о достаточности данных о частотности леммы в задаче оценки сложности лексики текста.

Среди направлений дальнейшей работы отметим, во-первых, проверку найденных закономерностей на большем количестве текстов, во-вторых, экспериментальную проверку влияния других формальных лингвистических показателей текста на русском языке на его сложность.

Благодарности и финансирование

Работа выполнена с использованием средств государственного бюджета по госзаданию на 2020–2024 годы (проект FZNM-2020-0005).

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Иомдин Б.Л., Морозов Д.А. Кто поймет «Незнайку»? Автоматическое определение сложности текстов для детей // *Русская речь*. 2021. № 5. С. 55–68. [Iomdin, Boris L. & Dmitry A. Morozov. 2021. Who can understand “Dunno”? Automatic assessment of text complexity in children’s literature. *Russian Speech* 5. 55–68 (In Russ.)]. https://doi.org/10.31857/S013161170017239-1

Корнеев А.А., Ахутина Т.В., Матвеева Е.Ю. Особенности чтения третьеклассников с разным уровнем развития навыка: анализ движений глаз // *Вестник Московского университета. Серия 14. Психология*. 2019. № 2. С. 64–87. [Korneev, Aleksei A., Tatiana V. Akhutina & Ekaterina Yu. Matveeva. 2019. Reading in third graders with different state of the skill: An eye-tracking study. *Vestnik Moskovskogo Universiteta. Seriya 14. Psikhologiya* 2. 64–87. (In Russ.)]. https://doi.org/10.11621/vsp.2019.02.64

Криони Н.К., Никин А.Д., Филиппова А.В. Автоматизированная система анализа сложности учебных текстов // *Вестник Уфимского государственного авиационного технического университета.* 2008. № 11 (1). С. 101–107. [Krioni, Nikolai K., Aleksei D. Nikin & Anastasia V. Filippova. 2008. Automated system for analyzing the complexity of educational texts. *Bulletin of the Ufa State Aviation Technical University* 11(1). 101–107. (In Russ.)].

Лапошина А.Н., Веселовская Т.С., Лебедева М.Ю., Купрещенко О.Ф. Лексический состав текстов учебников русского языка для младшей школы: корпусное исследование // *Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2019».* 2019. Т. 18 (25). С. 351–363. [Laposhina, Antonina N., Tatiana S. Veselovskaya, Maria U. Lebedeva & Olga F. Kupreshchenko. 2019. Lexical analysis of the Russian language textbooks for primary school: Corpus study. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*18. 351–363. (In Russ.)].

Мартынова Е.В., Солнышкина М.И., Мерзлякова А.Ф., Гизатулина Д.Ю. Лексические параметры учебного текста (на материале текстов учебного корпуса русского языка) // *Филология и культура.* 2020. № 3 (61). С. 72–80. [Martynova, Ekaterina V., Marina I. Solnyshkina, Amina F. Merzlyakova & Diana Yu. Gizatulina. 2020. Lexical parameters of the academic text (based on the texts of the academic corpus of the Russian language). *Philology and Culture* 3. 72–80. (In Russ.)]. https://doi.org/10.26907/2074-0239-2020-61-3-72-80

Мизернов И.Ю., Гращенко Л.А. Анализ методов оценки сложности текста. // *Новые информационные технологии в автоматизированных системах.* 2015. № 18. С. 572–581. [Mizernov, I. Yu. & L. A. Grashchenko. 2015. Analysis of methods for assessing text complexity. *New Information Technologies in Automated Systems* 18. 572–581. (In Russ.)].

Микк Я.А. О факторах понятности учебного текста: автореф. дис. … канд. пед. наук. Тарту, 1970. 22 с. [Mikk, Ya.A. 1970. Factors of educational text clarity. *Abstract of Pedagogy Cand. Diss. Tartu.* (In Russ.)].

Оборнева И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис... канд. пед. наук: 13.00.02. М., 2006. 165 с. [Oborneva, Irina V. 2006. Automated estimation of complexity of educational texts on the basis of statistical parameters. *Pedagogy Cand. Diss*. Moscow. (In Russ.)].

Солнышкина М.И., Кисельников А.С. Сложность текста: этапы изучения в отечественном прикладном языкознании. // *Вестник Томского государственного университета. Филология.* 2015. № 6 (38). С. 86–99. [Solnyshkina, Marina I. & Alexander S. Kiselnikov. 2015. Text complexity: Study phases in Russian linguistics. *Tomsk State University Journal of Philology* 6. 86–99. (In Russ.)]. https://doi.org/10.17223/19986645/38/7

Шпаковский Ю.Ф. Разработка количественной методики оценки трудности восприятия учебных текстов для высшей школы // *Научно-технический вестник информационных технологий, механики и оптики*. 2008. № 1 (83). С. 110–117. [Shpakovsky, Yury F. 2008. Development of a quantitative methodology for assessing the difficulty of perceiving educational texts for higher education. *Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics* 1(83). 110–117. (In Russ.)].

Chall, Jeanne S. & Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula.* Cambridge, MA: Brookline Books.

Chen, Xiaobin & Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In Joel Tetreault, Jill Burstein, Claudia Leacock & Helen Yannakoudakis (eds.),

Proceedings of the 11th workshop on innovative use of nlp for building educational applications, 84–94. San Diego: Association for Computational Linguistics.

Clifton, Jr. Charles, Adrian Staub & Keith Rayner. 2007. Eye movements in reading words and sentences. In Roger P. G. van Gompel, Martin H. Fischer, Wayne S. Murray & Robin L. Hill (eds.), *Eye movements: A window on mind and brain,* 341–371. Elsevier. https://doi.org/10.1016/B978-008044980-7/50017-3

Dorofeeva, Svetlana V., Victoria Reshetnikova, Margarita Serebryakova, Daria Goranskaya, Tatiana V. Akhutina & Olga Dragoy. 2019. Assessing the validity of the standardized assessment of reading skills in Russian and verifying the relevance of available normative data. *The Russian Journal of Cognitive Science* 6(1). 4–24.

DuBay, William H. 2007. *Smart Language: Readers, Readability, and the Grading of Text.* Costa Mesa, California: Impact Information.

Farris-Trimble, Ashley & Bob McMurray. 2018. Morpho-phonological regularities influence the dynamics of real-time word recognition: Evidence from artificial language learning. *Laboratory Phonology* 9(1). 1–34. https://doi.org/10.5334/labphon.41

Francois, Tomas & Cedrick Fairon. 2012. An 'AI readability' formula for French as a foreign language. *Proceedings of the EMNLP and CoNLL 2012, Jeju Island, Korea, 12–14 July 2012*. 466–477.

Glazkova, Anna, Yury Egorov & Maxim Glazkov. 2021. A comparative study of feature types for age-based text classification. In *Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science* 12602. 120–134.

Graesser, Arthur C., Danielle S. McNamara, Zhiqang Cai, Mark Conley, Haiying Li & James Pennebaker. 2014. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal* 115. 210–229.

Griffin, Zenzi M. & Daniel H. Spieler. 2006. Observing the what and when of language production for different age groups by monitoring speakers' eye movements. *Brain and Language* 99(3). 272–288.

Henderson, John M., Aleksander Pollatsek & Keith Rayner. 1989. Covert visual attention and extrafoveal information use during object identification. *Perception & Psychophysics* 45. 196–208. https://doi.org/10.3758/BF03210697

Jian, Yu-Cin & Hwawei Ko. 2017. Influences of text difficulty and reading ability on learning illustrated science texts for children: An eye movement study. *Computers & Education* 113. 263–279.

Lexile. 2007. *The Lexile Framework for Reading: Theoretical Framework and Development. Technical Report.* MetaMetrics, Inc., Durham, NC

Luke, Steven G., John M. Henderson & Fernanda Ferreira. 2015. Children's eye-movements during reading reflect the quality of lexical representations: An individual differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41(6). 1675–1683. https://doi.org/10.1037/xlm0000133

Raney, Gary E. & Keith Rayner. 1995. Word frequency effects and eye movements during two readings of a text. *Canadian Journal of Experimental Psychology* 49. 151–172.

Rau, Anne K., Kristina Moll & Karin Landerl. The transition from sublexical to lexical processing in a consistent orthography: An eye-tracking study. *Scientific Studies of Reading* 18. 224–233. https://doi.org/10.1080/10888438.2013.857673

Rau, Anne K., Kristina Moll, Margaret J. Snowling & Karin Landerl. 2015. Effects of orthographic consistency on eye movement behavior: German and English children and adults process the same words differently. *Journal of Experimental Child Psychology* 130. 92–105. https://doi.org/10.1016/j.jecp.2014.09.012.

Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124. 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, Keith, Timothy J. Slattery, Denis Drieghe & Simon P. Liversedge. 2011. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance* 37(2). 514–528.

Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco & Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In Paula Kotzé & Gary Marsden (eds.), *Human-Computer interaction – INTERACT 2013. Lecture notes in computer science vol 8120,* 203–219. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-40498-6_15

Reynolds, Robert. 2016. Insights from Russian second language readability classification: Complexity-dependent training requirements, and feature evaluation of multiple categories. *Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, San Diego, CA 2016*. 289–300.

Sato, Satoshi. 2014. Text Readability and Word Distribution in Japanese. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) 2014.* 2811–2815.

Schwarm, Sarah E. & Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05), USA, 2005.* 523–530.

Solovyev, Valery, Vladimir Ivanov & Marina Solnyshkina. 2018. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of Intelligent & Fuzzy Systems* 34. 3049–3058.

Tiffin-Richards, Simon P. & Sasha Schroeder. 2015. Children's and adults' parafoveal processes in German: Phonological and orthographic effects. *Journal of Cognitive Psychology* 27. 531–548. https://doi.org/10.1080/20445911.2014.999076

White, Sarah J., Denis Drieghe, Simon P Liversedge & Adrian Staub. 2018. The word frequency effect during sentence reading: A linear or nonlinear effect of log frequency? *Quarterly Journal of Experimental Psychology* 71(1). 46–55. https://doi.org/10.1080/17470218.2016.1240813

Словари/Dictionaries

Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник. 2009. [Lyashevskaya, Olga N. & Sergey A. Sharoff. 2009. Modern Russian Frequency Dictionary (based on the data from the Russian National Corpus). Moscow: Azbukovnik. (In Russ.)]

Article history:
Received: 19 November 2021
Accepted: 21 January 2022

Bionotes:

Antonina N. LAPOSHINA is a leading expert of the Laboratory of Cognitive and Linguistic Studies at Pushkin State Russian Language Institute, member of the research group “Teaching Russian in the Digital Age”. Her research interests include computer-assisted language learning and corpus-based language learning.

Contact information:
6 Akademika Volgina street, Moscow, 117485, Russia
e-mail: ANLaposhina@pushkin.institute
ORCID: 0000-0003-0693-7657

Maria Yu. LEBEDEVA holds a PhD in Philology and is a leading researcher of the Laboratory of Cognitive and Linguistic Studies, Associate Professor of the Department of Methods of Teaching Russian as a Foreign Language at Pushkin State Russian Language Institute. Her research interests are focused on corpus-based language learning, methods of online teaching Russian and reading strategies in the digital age.

Contact information:

6 Akademika Volgina street, Moscow, 117485, Russia
e-mail: MULebedeva@pushkin.institute
ORCID: 0000-0002-9893-9846

Alexandra A. BERLIN KHENIS is a specialist of the Laboratory of Cognitive and Linguistic Studies at Pushkin State Russian Language Institute. Her research interests include cognitive psychology, as well as psychophysiological aspects of reading and learning.

Contact information:

6 Akademika Volgina street, Moscow, 117485, Russia
e-mail: alexa.munxen@gmail.com
ORCID: 0000-0003-2034-1526

Сведения об авторах:

Антонина Николаевна ЛАПОШИНА – ведущий эксперт лаборатории когнитивных и лингвистических исследований Государственного института русского языка имени А.С. Пушкина. Сфера научных интересов: компьютерная лингвистика, цифровая лингводидактика.

Контактная информация:

Российская Федерация, 117485, Москва, ул. Академика Волгина, д. 6
e-mail: ANLaposhina@pushkin.institute
ORCID: 0000-0003-0693-7657

Мария Юрьевна ЛЕБЕДЕВА – кандидат филологических наук, ведущий научный сотрудник лаборатории когнитивных и лингвистических исследований, доцент кафедры методики преподавания РКИ Государственного института русского языка имени А.С. Пушкина. Сфера научных интересов: корпусная лингводидактика, особенности чтения в цифровую эпоху.

Контактная информация:

Российская Федерация, 117485, Москва, ул. Академика Волгина, д. 6
e-mail: MULebedeva@pushkin.institute
ORCID: 0000-0002-9893-9846

Александра Алехандровна БЕРЛИН ХЕНИС – специалист лаборатории когнитивных и лингвистических исследований Государственного института русского языка имени А.С. Пушкина. Сфера научных интересов: когнитивная психология, психофизиологические аспекты чтения и обучения.

Контактная информация:

Российская Федерация, 117485, Москва, ул. Академика Волгина, д. 6
e-mail: alexa.munxen@gmail.com
ORCID: 0000-0003-2034-1526

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 515–549
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-29475

Research article

Russian dictionary with concreteness/abstractness indices

Valery D. SOLOVYEV[1] ✉, Yulia A. VOLSKAYA[1], Mariia I. ANDREEVA[1,2] and Artem A. ZAIKIN[1]

[1]Kazan (Volga region) Federal University, Kazan, Russia
[2]Kazan State Medical University, Kazan, Russia
✉maki.solovyev@mail.ru

Abstract
The demand for a Russian dictionary with indices of abstractness/concreteness of words has been expressed in a number of areas including linguistics, psychology, neurophysiology and cognitive studies focused on imaging concepts in human cognitive systems. Although dictionaries of abstractness/concreteness were compiled for a number of languages, Russian has been recently viewed as an under-resourced language for the lack of one. The Laboratory of Quantitative Linguistics of Kazan Federal University has implemented two methods of compiling dictionaries of abstract/concrete words, i.e. respondents survey and extrapolation of human estimates with the help of an original computer program. In this article, we provide a detailed description of the methodology used for assessing abstractness/concreteness of words by native Russian respondents, as well as control algorithms validating the survey quality. The implementation of the methodology has enabled us to create a Russian dictionary (1500 words) with indices of concreteness/abstractness of words, including those missing in the Russian Semantic Dictionary by N.Yu. Shvedova (1998). We have also created three versions of a machine dictionary of abstractness/concreteness based on the extrapolation of the respondents' ratings. The third, most accurate version contains 22,000 words and has been compiled with the use of a modern deep learning technology of neural networks. The paper provides statistical characteristics (histograms of the distribution of ratings, dispersion, etc.) of both the machine dictionary and the dictionary obtained by interviewing informants. The quality of the machine dictionary was validated on a test set of words by means of contrasting machine and human evaluations with the latter viewed as more credible. The purpose of the paper is to give a detailed description of the methodology employed to create a concrete/abstract dictionary, as well as to demonstrate the methodology of its application in theoretical and applied research on concrete examples. The paper shows the practical use of this vocabulary in six case studies: predicting the complexity of school textbooks as a function of the share of abstract words; comparing abstractness indices of Russian-English equivalents; assessing concreteness/abstractness of polysemantic words; contrasting ratings of different age groups of respondents; contrasting ratings of respondents with different levels of education; analyzing concepts of "concreteness" and "specificity".
Keywords: *concreteness, abstractness, digital dictionary, Russian, academic texts*

For citation:
Solovyev, Valery D., Yulia A. Volskaya, Mariia I. Andreeva & Artem A. Zaikin. 2022. Russian dictionary with concreteness/abstractness indices. *Russian Journal of Linguistics* 26 (2). 515–549. https://doi.org/10.22363/2687-0088-29475

Научная статья

Словарь русского языка с индексами конкретности/абстрактности

В.Д. СОЛОВЬЕВ[1] ✉, Ю.А. ВОЛЬСКАЯ[1], М.И. АНДРЕЕВА[1,2], А.А. ЗАИКИН[1]

[1]Казанский (Приволжский) федеральный университет, Казань, Россия
[2]Казанский государственный медицинский университет, Казань, Россия
✉maki.solovyev@mail.ru

Аннотация.
Для целого ряда исследований в лингвистике, психологии, нейрофизиологии, посвященных репрезентации концептов в когнитивной системе человека, требуется словарь с численными оценками степени конкретности/абстрактности слов. Такие словари созданы для нескольких языков, но до последнего времени не было словаря для русского языка. В лаборатории квантитативной лингвистики Казанского федерального университета подготовлено несколько вариантов такого рода словаря для русского языка. При их создании использованы две методологии: опрос респондентов и разработка компьютерных программ для экстраполяции человеческих оценок. В статье подробно описана методология оценки абстрактности/конкретности слов респондентами-носителями русского языка, а также способы контроля качества их ответов. Применение данной методологии позволило создать словарь русского языка (1500 слов) с указанием индексов конкретности/абстрактности слов, в том числе отсутствующих в Русском семантическом словаре Н.Ю. Шведовой (1998). В нашей лаборатории созданы также три версии машинного словаря абстрактности/конкретности, полученные экстраполяцией оценок респондентов. Последняя версия словаря (22 тыс. слов), составлена с применением современной технологии глубокого обучения нейронных сетей и является наиболее точной. Приведены статистические характеристики (гистограммы распределения оценок, дисперсия и др.) и машинного словаря, и словаря, полученного опросом информантов. Оценка качества машинного словаря осуществлена на тестовом множестве слов путем сопоставлением машинных оценок с человеческими. Цель данной статьи – дать подробное описание методологии создания словаря конкретности/абстрактности, а также на конкретных примерах продемонстрировать методику его применения в теоретических и прикладных исследованиях. В статье показано практическое использование данного словаря в шести конкретных исследованиях: определение сложности текстов по доле абстрактных слов (на примере школьных учебников), сравнение оценок слов и их переводных эквивалентов в английском языке, оценки конкретности/абстрактности многозначных слов, сравнение оценок разных возрастных групп респондентов, сравнение оценок респондентов с разным уровнем образования, сравнение концепций «конкретность» и «специфичность».

Ключевые слова: *конкретность, абстрактность, электронный словарь, русский язык, учебные тексты*

Для цитирования:
Соловьев В.Д., Вольская Ю.А., Андреева М.И., Заикин А.А. Словарь русского языка с индексами конкретности/абстрактности. *Russian Journal of Linguistics.* 2022. Т. 26. № 2. С. 515–549. https://doi.org/10.22363/2687-0088-29475

1. Введение

Категория абстрактности/конкретности уже десятилетия находится в центре внимания когнитивных исследований. Проблема представления конкретных и абстрактных объектов в мозгу человека представляет собой серьезный вызов всей когнитивной науке (Borghi et al. 2017). Современный подход к ее изучению начинается с фундаментальной работы (Paivio 1965). Основной подход к определению этих понятий представлен в работе (Spreen & Shulz 1966). Конкретные понятия – те, которые воспринимаются органами чувств. Примеры конкретных слов – *кошка, стул, гора*. Абстрактные понятия не воспринимаются органами чувств. Например, *ответственность, взаимоотношения, непонимание*. Схожие трактовки встречаются во многих исследованиях. Так, в работе (Schmid 2000) приводится такое определение: «abstract nouns are those nouns whose denotata are not part of the concrete physical world and cannot be seen or touched» («денотаты абстрактных существительных не принадлежат физическому миру, т.е. их нельзя увидеть или дотронуться до них»[1]). Однако данные определения сильно упрощают ситуацию, давая характеристики прототипов конкретности и абстрактности. В действительности экспериментальные исследования показали, что изучаемая категория – континуум, но не дихотомия (Mkrtychian et al. 2019). В связи с этим весьма сложно дать совершенно строгое чисто лингвистическое определение этих понятий, которое позволило бы любое слово однозначно квалифицировать как конкретное или абстрактное.

Для поддержки вышеуказанных когнитивных исследований требуются словари с индексами, характеризующими степень конкретности/абстрактности слов. Обычно словарь создается методом опроса носителей языка, которым предлагается выставить рейтинг конкретности/абстрактности заданных слов, кроме этого применяются методы машинного обучения для расширения словарей путем экстраполяции уже имеющихся рейтингов на другие слова.

Статья подводит итоги первого этапа исследований в этом направлении лаборатории квантитативной лингвистики КФУ и обобщает опыт построения первого для русского языка словаря рейтингов конкретности/абстрактности, а также результаты первых исследований на его основе. Словарь свободно доступен по адресу (ENA, April 17, 2022)[2].

2. Обзор литературы

Исследования категории конкретности/абстрактности ведутся широким фронтом от психологии и психолингвистики до нейрофизиологии и

[1] Здесь и далее перевод выполнен авторами статьи.
[2] https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html.

медицины. Опубликованы тысячи статей, свежие обзоры можно найти в (Mkrtychian et al. 2019, Solovyev 2021). В нейрофизиологии изучался вопрос локализации понятий абстрактности/конкретности. Во многих экспериментах с помощью техники нейровизуализации было показано, что конкретные и абстрактные слова репрезентируются в разных нейроанатомических структурах мозга.

В психологических исследованиях установлен так называемый «эффект конкретности», демонстрирующий большую легкость обработки конкретных слов в человеческом сознании. Конкретные слова лучше запоминаются (Schwanenflugel et al. 1992), лучше распознаются (Fliessbach et al. 2006), быстрее читаются (Schwanenflugel & Shoben 1983), быстрее усваиваются (Mestres-Missé et al. 2014). Для таких слов легче написать словарные толкования, и они будут более детальными (Sadoski 1997). Респонденты легче продуцируют ассоциации в ответ на конкретные слова-стимулы (de Groot 1989).

В современной науке предложены две основные теории репрезентаций конкретной/абстрактной лексики: двойного кодирования (dual coding theory) (Paivio 1990) и доступного контекста (context availability theory) (Schwanenflugel & Shoben 1983). Теория двойного кодирования постулирует существование 2-х систем памяти: образной и словесной, причем образная система, в отличие от словесной, обеспечивает кодирование только конкретной информации. Согласно теории доступного контекста конкретные и абстрактные слова различаются количеством и силой ассоциативных связей. В русле теории доступного контекста было показано, что конкретные слова активируют более широкий вербальный контекст и поэтому обрабатываются быстрее, но не получают доступа к системе обработки изображений. В целом, исследователи соглашаются в отношении способов репрезентации конкретных, но не абстрактных понятий. Когнитивная лингвистика предлагает свой подход (Kousta et al. 2011) к репрезентации абстрактных понятий: согласно гипотезе воплощенности (embodied abstract semantics), эмоциональный опыт имеет критическое значение для репрезентации и обработки абстрактных слов.

Понятия конкретности и абстрактности сами по себе являются предметами изучения в лингвистике достаточно давно, однако в последнее время с появлением больших корпусов текстов и обширных лексических онтологий появились принципиально новые идеи исследований и результаты. К числу наиболее интересных можно отнести следующие. В работе (Snefjella et al. 2019) показано, что с течением времени степень конкретности слов возрастает. В статье (Reilly & Desai 2017) описано, что плотность множества семантически близких слов выше для конкретных слов, нежели для абстрактных. В работе (Naumann et al. 2018) замечено, что в корпусах текстов абстрактные слова чаще встречаются вместе с абстрактными, а конкретные – с конкретными. В работе (Ivanov & Solovyev 2021) проведено сопоставление категорий конкретности и специфичности, показано их существенное различие.

Исследования конкретности/абстрактности имеют различные прикладные аспекты. В медицине абстрактные слова играют важную роль в ходе терапии больных с афазией (Dallin et al. 2020). Доля абстрактных слов является одним из значимых показателей сложности текстов (Sadoski et al. 2001, McNamara et al. 2014). Этот параметр включен в доступные онлайн-пакеты расчета сложности текстов (Coh-Metrix). Автоматически оцененная доля абстрактных слов вместе с другими параметрами может быть использована в педагогике для оценки сложности текстов с целью адекватного выбора образовательных материалов.

Для проведения психологических, нейрофизиологических экспериментов нужны списки слов с оценками степени их конкретности/абстрактности. Далее в статье как синоним слова *оценка* будет использоваться и слово *рейтинг*. Оценки получаются методом опроса носителей языка, в результате которого составляется словарь с рейтингами абстрактности/конкретности слов. Для английского языка первый крупный словарь такого рода создан в 1981 г. (Coltheart 1981). Он содержит почти 4 тыс. слов и свободно доступен в составе психолингвистической базы данных MRC (ENA, April 17, 2022)[3]. Позднее был создан словарь, включающий почти 40 тыс. слов (Brysbaert et al. 2014a). Каждое слово получает не менее 25 оценок респондентов по 5-балльной шкале, которые усредняются. Кроме английского языка сравнимый по объему словарь создан лишь для нидерландского (Brysbaert et al. 2014b). Очевидной проблемой является большая трудоемкость составления подобных словарей. Для немецкого языка словарь (Maximilian & Walde 2016) содержит лишь 4 тыс. слов. Недавно опубликована база данных с рейтингами конкретности/абстрактности для хорватского языка на 6 тыс. слов (Peti-Stantić et al. 2021). Аналогичные словари созданы для итальянского (Vergallito et al. 2020), китайского (Yao et al. 2017) и ряда других языков.

В связи с большой трудоемкостью построения словаря путем проведения опросов актуальной является задача создания компьютерных словарей методом автоматической экстраполяции человеческих оценок, полученных на небольшом множестве слов, на большее множество. Основная идея экстраполяции человеческих оценок на ранее неоцененные слова состоит в использовании векторной семантики слов (Mikolov et al. 2013), построенной на базе большого корпуса текстов, и получении новых оценок на основе семантической близости слов в построенном семантическом пространстве. Таким образом, необходимым условием создания в определенном языке компьютерного словаря является существование большого корпуса текстов, на основе которого можно строить векторную семантику.

Принципиально важной является оценка качества машинных словарей. Они оцениваются путем сравнения со словарями, созданными на основе опросов, с вычислением коэффициента корреляции двух словарей, чаще всего по Спирмену. К настоящему времени лучший достигнутый результат –

[3] https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

машинный словарь для английского языка работы (Charbonnier & Wartena 2019), он имеет коэффициент корреляции со словарем на основе опросов 0,900. Словарь создан с использованием технологии fastText (Joulin et al. 2016) для построения семантического пространства и SVM (Cristianini & Shawe-Taylor 2000) в качестве классификатора. Экстраполяция человеческих оценок в данном словаре осуществляется путем кросс-валидации на 40-тысячном словаре (Brysbaert et al. 2014a). В (Brysbaert, Warriner & Kuperman 2014a) было проведено сопоставление двух словарей на основе опроса респондентов, и оказалось, что коэффициент корреляции между ними равен 0,919. Т.е. результат 0,900 – почти предельно возможный.

3. Словарь с индексами конкретности/абстрактности для русского языка

3.1. Построение словаря

Словарь с индексами конкретности/абстрактности для русского языка, включающий 1,5 тыс. слов, создан методом опроса респондентов в Казанском федеральном университете. Для оценки взяты наиболее частотные существительные из известного частотного словаря О.Н. Ляшевской, С.А. Шарова (Ляшевская & Шаров 2009). Слова предлагались респондентам в виде анкет (гугл-форм) по 50 слов в каждой. Мы считаем, что наш опрос проведен более тщательно по сравнению с опросом для английского языка. Дело в том, что анкеты для английского содержали по 300 слов (Brysbaert et al. 2014a, 2014b). Естественно ожидать, что к концу столь длинного списка слов концентрация внимания респондентов падает, и доля ошибочных оценок должна возрастать.

В проведенном исследовании не было установлено ограничение времени на заполнение анкет респондентами. Для оценки конкретности/абстрактности слов авторы руководствовались исследованием (Laming 2004) и использовали 5-балльную шкалу Ликерта, где 1 маркирует высокую степень конкретности, а 5 – высокую степень абстрактности. Исследования, проведенные на материале английского языка (Colheart 1981), основывались на 7-балльной шкале от 1 (абстрактные) до 7 (конкретные). Полученные оценки слов масштабировались с коэффициентом 100. Таким образом, была получена шкала от 100 до 700. Для возможности последующего сравнения данных обоих языков произведена перенормировка наших данных по формуле y = 100 * (1.5 * (5-x)) +1), где x – значение оценки конкретности для русского языка (Solovyev et al. 2019b). Таким образом рассчитаны другие значения рейтинга, удобные для сопоставления с английским рейтингом: наивысшее значение конкретности маркировалось 700 единицами, наивысшее значение абстрактности – 100. В дальнейшем в разных исследованиях мы использовали различные шкалы – и от 1 до 5, и от 100 до 700.

Опрос был разделен на две части. В первой (Solovyev et al. 2019a) в качестве респондентов участвовали около 400 студентов (от 17 до 25 лет) очной

формы обучения Казанского федерального университета и около 300 студентов Белорусского государственного педагогического университета, носителей русского языка (Zhuravkina et al. 2020). В этой части получены оценки для 1000 слов, для каждого слова – не менее 40 оценок.

Во второй части опроса (500 слов) (Вольская 2020) использовалась система Яндекс.Толóка, в экспериментах могли принять участие все желающие. Во второй части для каждого слова опрашивалось 60 человек. В первой части какие-либо дополнительные инструкции участникам не давались. Однако в аналогичном построении словарей для английского языка респонденты подробно инструктировались (Brysbaert et al. 2014a, 2014b). Поэтому во второй части перед прохождением опроса мы также давали респондентам детальные инструкции, причем максимально близкие к приведенным в зарубежных работах. В них были даны подробные описательные определения абстрактных и конкретных слов с примерами. Определения в духе тех, что приведены во введении, опирались на возможность восприятия слов органами чувств. Далее подчеркивалось, что некоторые слова могут сочетать в себе как признаки конкретности, так и абстрактности; приводилось описание принципа оценки слов (указывалось соответствие числовых значений степени проявления абстрактности). Приведем фрагмент пояснений: «слово “любовь” более абстрактно, так как означает некое отвлеченное понятие, которое лишено физической очерченности, а вот слово “стол” – более конкретно, это реальный предмет, который можно потрогать, увидеть и т.д».

При использовании системы Яндекс.Толóка респондентам было необходимо указать возраст, пол, уровень образования (среднее, средне специальное, неоконченное высшее, высшее, высшее филологическое). После этого пользователям открывался доступ к оценке лексических единиц. Система предоставляет возможность отбирать респондентов по уровню образования, возрасту, родному языку, их квалификации, судя по предыдущей работе в Яндекс.Толóка. Мы выделили две возрастные группы: от 18 до 30 и от 31 до 55 лет. Допускались лишь те респонденты, для которых русский язык является родным. Также данный опрос могли проходить только лучшие исполнители сервера, число которых составляет 20% от общего количества зарегистрированных на Яндекс.Толóка участников.

В рамках первой части опроса для перепроверки оценок для 100 слов получены дополнительные оценки других участников эксперимента. Коэффициент корреляции для этих двух независимых оценок оказался равен 0,879. Аналогичное сравнение двух вышеупомянутых экспериментов для английского дало коэффициент корреляции 0,919. Несколько более низкий результат у нас можно объяснить тем, что в этой части эксперимента мы, в отличие от опросов для английского языка, не давали респондентам четких определений конкретности/абстрактности. В итоге впервые создан словарь слов русского языка с численными оценками конкретности/абстрактности, полученными опросом респондентов.

3.2. Коррекция исходных данных. Статистика

В ходе визуальной проверки оценок при проведении первого опроса был выявлен ряд недобросовестных респондентов, например таких, которые оценили все слова одним и тем же баллом. В связи с этим возникла проблема очистки собранных данных от мусора. На основе работы (Chandola et al. 2009) реализованы 5 способов очистки данных.

1. Расчет автокорреляции оценок респондента. Слишком высокий уровень автокорреляции первого порядка указывает на несерьезное или по меньшей мере недостаточно вдумчивое отношение к эксперименту. Удаляются ответы респондентов, у которых оценки выходят за рамки стандартного распределения.

2. Расстояние от вектора оценок респондента до вектора средних оценок. Расстояние измерялось по манхэттенской метрике (Black 2019). Респонденты, оценки которых слишком отличались от средних, исключались из исследования.

3. Совпадение результатов двух и более респондентов. Если у двух респондентов оценки полностью совпадали, то результаты одного из них отбрасываются.

4. Алгоритм иерархической кластеризации с одиночным связыванием применялся к множеству векторов оценок респондентов. Далее выделялись кластеры, слишком далеко отстоящие от остальных, и также удалялись.

5. В каждом опросе (50 слов, оценки не менее 40 респондентов) отбиралось одно слово с наименьшей средней оценкой и одно слово с наибольшей оценкой. Удаляются результаты тех респондентов, которые оценили слово обратным образом – 5 баллами или 1 баллом соответственно (что могло быть связано с простой ошибкой в полюсах семантического дифференциала).

При создании словаря реализован жесткий подход, при котором удалялись не только явные выбросы, но и все сомнительные случаи. В итоге удалено около четверти всех результатов. Таким образом мы получили около 30 оценок для каждого слова. Отметим, что при создании словаря для английского языка для каждого слова исходно предполагалось получить не менее 30 оценок, однако после аналогичного отбрасывания оценок недобросовестных респондентов для ряда слов количество оценок уменьшилось до 25 (Brysbaert et al. 2014a, 2014b). Гистограмма данных, оставшихся после удаления ошибочных, приведена на рис. 1. Большинство оценок приходится на интервал от 1,4 до 3,6 с некоторым преобладанием оценивания слов как конкретных. Среднее значение равно 2,5. Выделяются три пика: скорее конкретных слов, скорее абстрактных и промежуточных.

Разности оценок до и после очистки распределены по нормальному закону (рис. 2). Большинство разностей по абсолютной величине не превышает 0,2. p-значение критерия Шапиро–Уилка равно 0,576. Коэффициент корреляции Пирсона между средними оценками исходными и очищенными составил 0,978. Таким образом, очистка практически не повлияла на конечный результат.

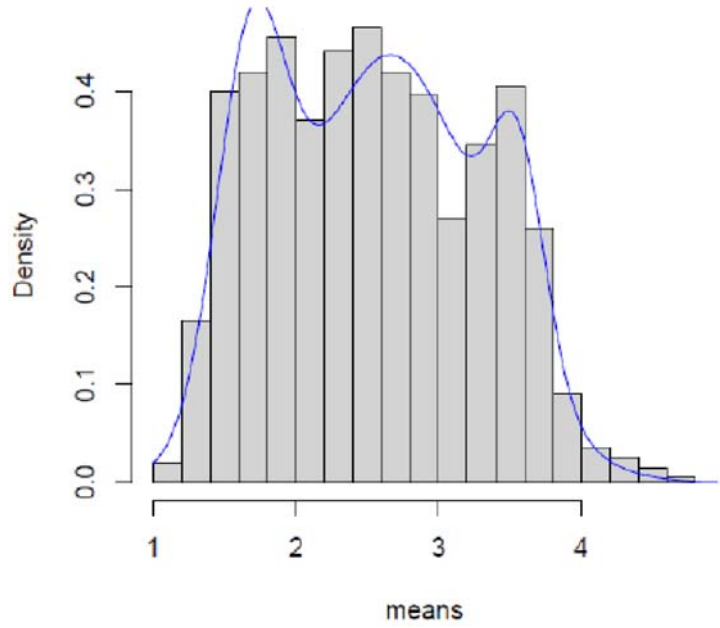

Рис. 1. Гистограмма распределения оценок / Fig. 1. Histogram of ratings distribution

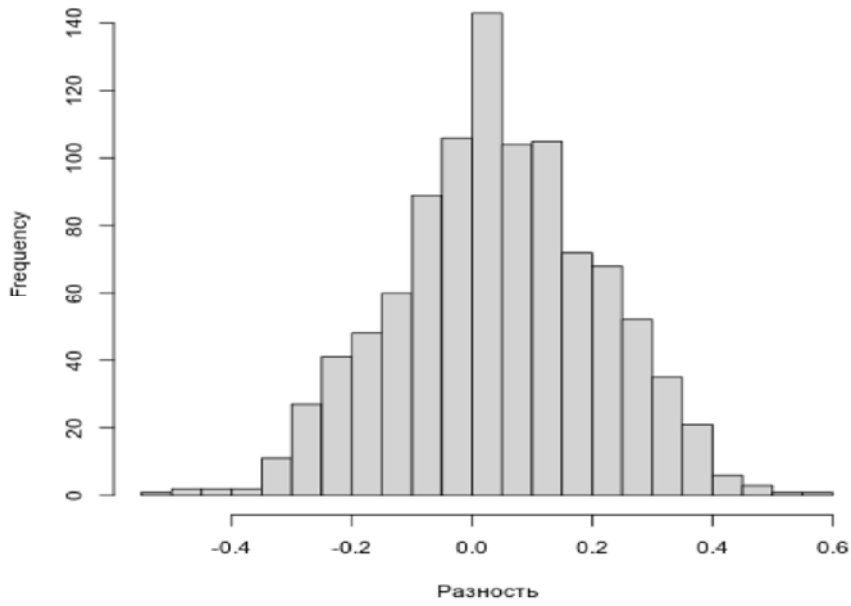

Рис. 2. Распределение разностей оценок до и после очистки /
Fig. 2. Distribution of ratings difference prior to and after filtration

На рис. 3 приведена гистограмма распределения дисперсии оценок. Для большинства слов дисперсия находится в пределах от 0,8 до 1,4. Мы провели сравнение дисперсии оценок для конкретных и абстрактных слов. Для выделения конкретных и абстрактных слов все слова упорядочены по рейтингам

и разделены на три равные по величине части, причем часть с промежуточными рейтингами не рассматривается. Средняя дисперсия конкретных слов равна 0,9, абстрактных – 1,15. Для абстрактных слов оценки респондентов имеют больший разброс, т.е. респонденты чаще оценивают их по-разному. Это хорошо согласуется с результатом недавнего психологического исследования (Wang & Bi 2021), показавшем, что респонденты указывают больше значений абстрактных слов, чем конкретных.

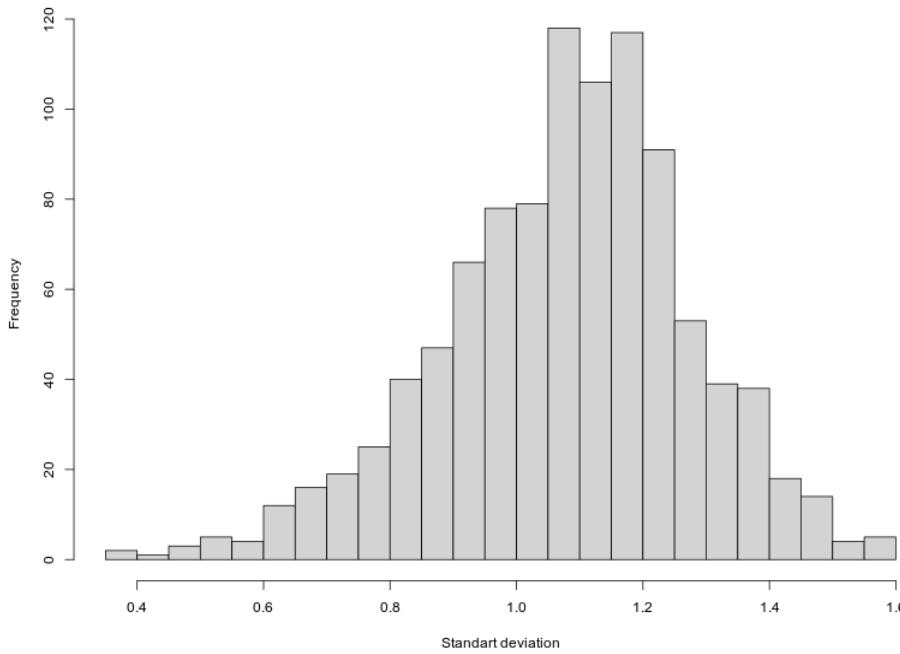

Рис. 3. Гистограмма распределения дисперсии оценок
Fig. 3. Histogram of distribution of ratings dispersion

На рис. 4 приведен график рассеяния с наложенной линией, скользящего среднего. Стандартное отклонение уменьшается при малых и больших значениях конкретности и достигает максимума для приблизительно средних значений.

На рис. S1 в Приложении приведен график распределения оценок, упорядоченных по величине. Характерные распределения оценок для типичных конкретного и абстрактного слова приведены на рис. S2 в Приложении.

В ходе второго опроса в настройках Яндекс.Толóка применялась отложенная приемка, что позволяло оценить ответы пользователей в соответствии с критериями контроля качества и в случае необходимости отклонить ответы тех из них, которые нарушали установленные правила. Для контроля качества прохождения опросов использовались следующие критерии, поддерживаемые сервисом Яндекс.Толóка и аналогичные используемым в работах (Brysbaert et al. 2014a, 2014b):

1) в каждый список было включено 10 контрольных слов. Это наиболее частотные единицы, которые уже были оценены ранее и которые демонстрируют весь диапазон проявления степени конкретности/абстрактности: *дверь, рука, книга, машина, место, слово, часть, сила, возможность, отношение*. Если при анализе ответов была обнаружена слабая корреляция между оценками контрольных слов данного пользователя со средними оценками, полученными в ходе первого опроса, ответы данного пользователя отклонялись;

2) не принимались ответы с единообразием оценок;

3) если пользователь выполнял задание быстрее, чем за установленное минимальное время – 4 минуты, то его ответы отклонялись автоматически.

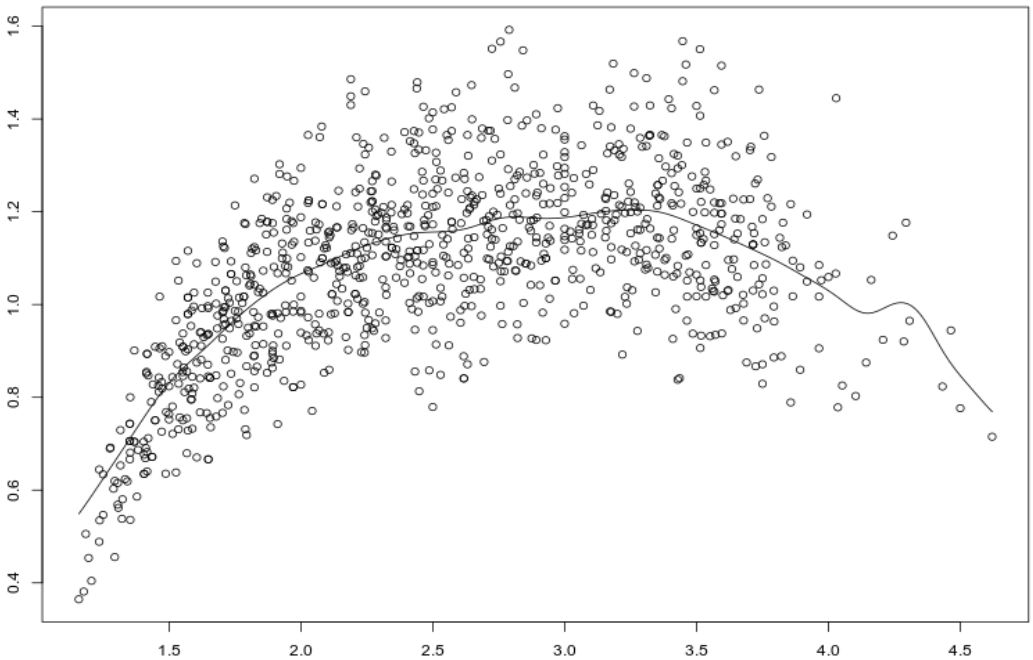

Рис. 4. Распределение дисперсий с наложенной линией скользящего среднего
Fig. 4. Dispersions distribution with superimposed moving average line

3.3. Машинный словарь

Для увеличения количества слов с рейтингами абстрактности/конкретности создан также компьютерный словарь. Причем он существует в трех вариантах. Первый – словарь словоформ (а не лемм) (Solovyev et al. 2019b), содержащий 88 тыс. словоформ существительных и прилагательных, созданный на материале корпуса Google Books Ngram (https://books.google.com/ngrams). При его составлении реализован оригинальный метод, основанный на идее, что конкретные слова встречаются в текстах чаще вместе с конкретными, а абстрактные – вместе с абстрактными. Второй вариант компьютерного словаря создан по технологии

word2vec, модель fastText (ENA, April 17, 2022)[4]. Он содержит 64 тыс. слов (лемм) (Solovyev et al. 2020a). Третий вариант – машинный словарь на 22 тыс. слов, построенный на технологии глубокого обучения, модель BERT (Devlin et al. 2018). Все варианты машинных словарей доступны по адресу (ENA, April 17, 2022)[5].

Оценка качества словарей показала, что наиболее высокий уровень корреляции между машинным и рейтингом респондентов у третьего словаря, созданного при помощи технологии BERT – 0.81 по Спирмену. Это заметно ниже результата, зафиксированного для английского языка – 0,90. Вероятно, это связано с тем, что качество русскоязычной версии BERT ниже англоязычной. В работе (Peti-Stantić et al. 2021) в аналогичном исследовании для хорватского языка полученный результат также оказался заметно хуже итогов сопоставления англоязычных словарей. Однако если ограничиться построением рейтингов высокочастотных слов (встречающихся в Google Books Ngram не менее 1 млн раз), то, как показано в работе (Solovyev et al. 2020b), точность предсказания рейтингов значительно возрастает, примерно до 0,86–0,87 по Спирмену.

4. Анализ словаря

В этом разделе статьи мы проанализируем полученные на основе ответов респондентов данные. Оценки будут рассмотрены под различными углами зрения.

4.1. Сопоставление со словообразовательным критерием

Одним из известных признаков абстрактности слова является наличие в нем определенных суффиксов. К ним относятся следующие: -изм, -аж, -итет (м.р.); -б-а, -от-а, -изн-а, -ин-а, -иц-а, -ура, -к-а, -аци-я, -н-я, -отн-я, -щин-а, -чин-а, -ость, -есть, -ность, -емость, -имость (ж.р., 3 скл.); -ие, -ье (-ьё), -ние, -нье (-ньё), -тие, -тье (-тьё), -ств-о, -еств-о, -тельств-о, -овств-о (ср.р.) (Виноградов 2001).

Из 1500 слов словаря были отобраны 150 слов, которые были оценены респондентами как наиболее абстрактные, и 150 слов, оцененные как наиболее конкретные. Это слова с рейтингами от 3,5 до 5 и от 1 до 2 соответственно. В результате оказалось, что у 94 слов (примерно две трети) с абстрактным значением такие суффиксы присутствуют. Таким образом, наличие суффикса абстрактности является хорошим критерием определения абстрактности слова, но все же он не охватывает примерно треть абстрактных слов.

Из 150 слов с наибольшим индексом конкретности (по данным наших опросов) у 19 слов присутствуют суффиксы абстрактности. Данные имена существительные классифицируем в 4 группы. Во-первых, суффикс -ени-е

[4] https://fasttext.cc/

[5] https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html.

обнаружен в морфемной структуре слов «*стихотворение*» и «*растение*». Однако данный суффикс указывает на абстрактность существительных только в том случае, если оно имеет процессуальное значение и образовано от глагольных основ. В данном случае существительное «*растение*» по словообразовательным признакам нельзя отнести к ЛГР абстрактных существительных. Слово «*стихотворение*» можно трактовать как спорный случай, хотя Семантическим словарем под редакцией Н.Ю. Шведовой (Шведова 1998) оно трактуется как абстрактное.

Во-вторых, в составе двух единиц: *лекарство* и *агентство* – выделяется суффикс -ств-о, указывающий на абстрактность, если производное существительное образовано от имени прилагательного. Указанные лексемы образованы от имен существительных. В-третьих, в одном слове, *прокуратура*, выделяется суффикс -ур-а. В данном случае суффикс указывает на собирательное значение рассматриваемого существительного: «система органов, осуществляющих от имени государства высший надзор за соблюдением законодательства».

В-четвертых, суффикс -к-а выявлен у 14 существительных. Однако данный суффикс может указывать на абстрактность только в том случае, если существительное со значением «действие» образовано от глагола. Из 14 слов только 2 удовлетворяют данному критерию: это существительные «*поставка*» и «*разведка*». В Семантическом словаре Н.Ю. Шведовой слово «*разведка*» в двух значениях трактуется как конкретное и в одном значении – как абстрактное. Существительные «*улыбка*» и «*записка*» также образованы от глаголов, но имеют значение «результат действия, которое указано производящей основой». 10 единиц являются производными от основ имен существительных.

Итак, из 19 слов с высокой степенью конкретности, имеющих суффиксы абстрактности, только два слова по словообразовательным признакам можно отнести к действительно абстрактным существительным (*поставка, стихотворение*).

4.2. Сравнение с данными словаря Н.Ю. Шведовой

Одним из немногих словарей русского языка, содержащим информацию о конкретности/абстрактности лексем, является Семантический словарь под редакцией Н. Ю. Шведовой. Его первый и второй тома посвящены конкретным существительным, третий – абстрактным. Следует отметить, что отглагольные существительные не включены в опубликованную часть словаря. В Семантическом словаре 115 слов из 150, рассмотренных в предыдущем разделе, также классифицируются как абстрактные существительные. 22 слова имеют и абстрактные, и конкретные значения, т.е. присутствуют и во втором, и в третьем томах. 13 слов в словаре Н.Ю. Шведовой отсутствуют (*проведение, осуществление, распространение, ведение, обслуживание, выполнение,*

изучение, основное, принятие, рассмотрение, снижение, увеличение, эффективность). Все они, кроме адъективного существительного *основное*, являются отглагольными существительными.

Из 150 слов с высокими значениями рейтинга конкретности, 144 слова по словарю Н.Ю. Шведовой во всех, либо в некоторых своих значениях являются конкретными 6 слов считаются абстрактными во всех своих значениях: *стихотворение, матч, улыбка, поставка, неделя, надпись.* Если обратиться к сенсорному критерию, предполагающему, что конкретные сущности воспринимаются органами чувств, то отнесение слов *улыбка* и *надпись* к абстрактным можно оспаривать. Таким образом, следует отметить очень высокую степень согласия словаря Шведовой с результатами опроса респондентов. Лишь в 1,3% случаев (*матч, неделя, поставка, стихотворение*) решение Н.Ю. Шведовой и оценки респондентов расходятся.

4.3. Сравнение оценок, полученных в результате опросов респондентов двух возрастных групп

В зарубежных исследованиях возраст респондентов никак не учитывался. Представляется интересным выяснить, есть ли заметные расхождения в оценках конкретности/абстрактности респондентами разных возрастов. Как указывалось ранее, вся выборка респондентов в нашем исследовании разделена на две группы двух возрастных категорий – от 18 до 30 лет (первая группа) и от 31 до 55 лет (вторая группа). В обеих группах оценивались одни и те же слова и респонденты находились в равных условиях. В ходе анализа полученных данных значительных расхождений между ответами обнаружено не было. Коэффициент корреляции Спирмена между оценками обеих групп является очень высоким – 0,933.

Следующая диаграмма (рис. 5) наглядно демонстрирует высокую степень корреляции оценок двух возрастных групп. На диаграмме точками представлены слова, по оси X размещены оценки второй группы, по оси Y – первой.

Разница между оценками варьируется от –1,4 до 1,2. Наибольшая отрицательная разница (от –0,5 до –1,4) обнаружена между оценками 25 слов, приведенных в табл. 1, получивших по ответам респондентов первой возрастной группы оценки от 1,5 до 3,9, а по оценкам респондентов второй группы – от 2,1 до 4,4.

Наибольшая положительная разница выявлена между оценками 20 слов (табл. 2). По ответам респондентов первой возрастной группы данные слова получили рейтинги сот 1,76 до 4,3, по ответам пользователей второй группы – от 1,26 до 3,56.

Среди ответов респондентов первой группы выявлены три лексемы с высокой степенью конкретности (от 1,76 до 2,36), 9 слов со срединным значением (от 2,5 до 3,13) и 8 слов с высокой степенью абстрактности (от 3,46

до 4,3). Среди ответов респондентов второй группы выявлено 11 лексем с высокой степенью конкретности (от 1,26 до 2,4), 8 слов со срединным значением (от 2,46 до 3,46) и слово с высокой степенью абстрактности (3,56). В целом установлено, что возраст (в рассмотренном нами диапазоне) не оказывает заметного влияния на оценку конкретности/абстрактности.

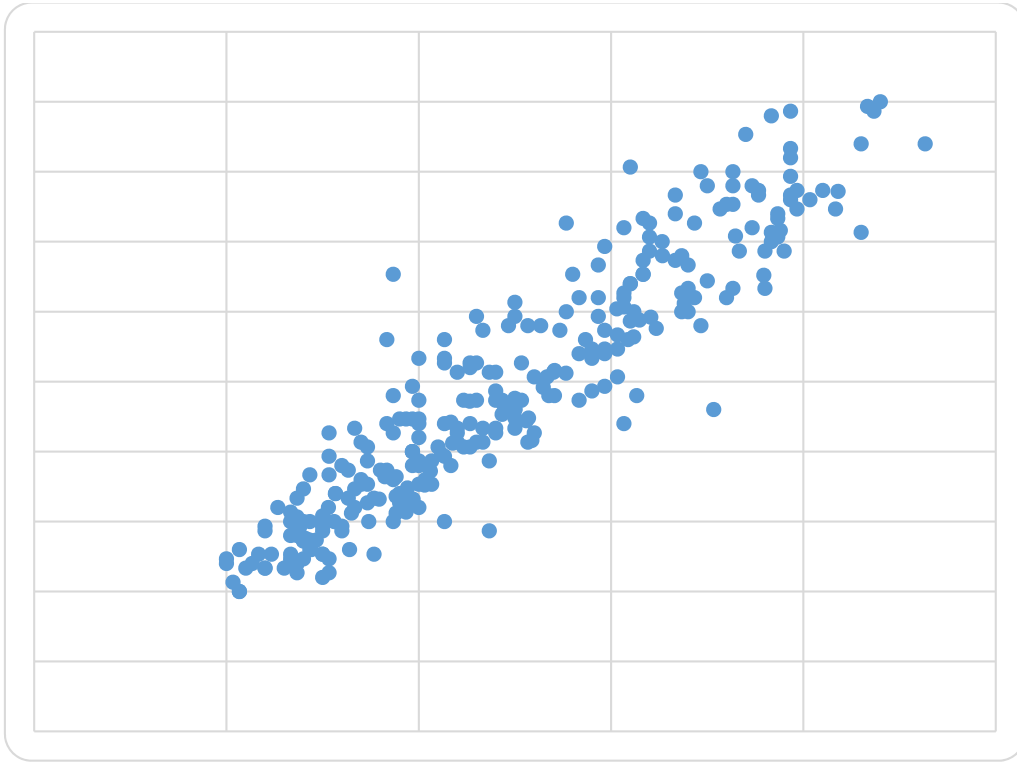

Рис. 5. Диаграмма оценок по двум возрастным группам /
Fig. 5. Ratings plot based on two age groups

Таблица 1. **Слова с наибольшей отрицательной разницей оценок**

Слово	Оценки первой группы	Оценки второй группы	Слово	Оценки первой группы	Оценки второй группы
разведка	1,533	2,133	критерий	1,867	3,267
отчет	1,667	2,167	охота	2,8	3,267
узел	1,867	2,4	дар	2,967	3,467
агент	1,967	2,467	нагрузка	3,067	3,6
указ	2,133	2,633	тариф	2,767	3,633
лекция	2	2,667	методика	3,167	3,667
знакомство	2,133	2,667	намерение	3,333	3,833
интервью	1,833	2,8	глупость	3,467	4
справка	2,133	2,8	жалоба	3,1	4,033
наказание	2,333	2,867	концентрация	3,7	4,267
свадьба	2,3	2,967	возможность	3,833	4,4
воскресенье	2,5	2,967	страдание	3,933	4,433
статистика	2,5	3,067			

Table 1. **Words with major negative difference of ratings**

Word	Group 1 ratings	Group 2 ratings	Word	Group 1 ratings	Group 2 ratings
scouting	1,533	2,133	criterion	1,867	3,267
report	1,667	2,167	hunt	2,8	3,267
knot	1,867	2,4	gift	2,967	3,467
agent	1,967	2,467	exertion	3,067	3,6
decree	2,133	2,633	rate	2,767	3,633
lecture	2	2,667	method	3,167	3,667
acquaintance	2,133	2,667	intention	3,333	3,833
interview	1,833	2,8	dullness	3,467	4
inquiry	2,133	2,8	complaint	3,1	4,033
ordeal	2,333	2,867	concentration	3,7	4,267
wedding	2,3	2,967	opportunity	3,833	4,4
Sunday	2,5	2,967	suffering	3,933	4,433
statistics	2,5	3,067			

Таблица 2. **Слова с наибольшей положительной разницей оценок**

Слово	Оценки первой группы	Оценки второй группы	Слово	Оценки первой группы	Оценки второй группы
шар	1,767	1,267	добыча	2,900	2,433
туман	2,367	1,433	питание	2,967	2,467
салон	2,133	1,500	символ	3,033	2,533
съемка	2,567	2,067	оборот	3,467	2,900
задание	2,588	2,080	перемена	3,600	3,100
совещание	2,600	2,133	карьера	3,633	3,167
раздел	3,067	2,200	секрет	3,800	3,167
темп	3,533	2,300	напряжение	3,794	3,260
свидетельство	2,833	2,367	осуществление	3,900	3,433
стандарт	3,133	2,400	восстановление	4,300	3,567

Table 2. **Words with major positive difference of ratings**

Word	Group 1 ratings	Group 2 ratings	Word	Group 1 ratings	Group 2 ratings
ball	1,767	1,267	prey	2,900	2,433
fog	2,367	1,433	nourishment	2,967	2,467
hall	2,133	1,500	symbol	3,033	2,533
filming	2,567	2,067	turn	3,467	2,900
task	2,588	2,080	change	3,600	3,100
meeting	2,600	2,133	career	3,633	3,167
section	3,067	2,200	secret	3,800	3,167
tempo	3,533	2,300	strain	3,794	3,260
certificate	2,833	2,367	implementation	3,900	3,433
standard	3,133	2,400	restoration	4,300	3,567

4.4. Сравнение оценок, полученных в результате опросов респондентов с разным уровнем образования

В ходе сбора данных посредством сервиса Яндекс.Толóка сохранялись сведения об уровне образования респондентов. Они были использованы для проверки гипотезы о том, что уровень образования может влиять на

вариативность оценок: чем больше значений слова известно респондентам данной группы, тем больше вариативность выбранных оценок по этому слову. Таким образом, предполагалось, что чем выше уровень образования, тем больше значений слова известно пользователю, следовательно, разброс оценок будет шире в группе респондентов с высшим образованием.

Нами был проведен сравнительный анализ оценок пользователей со средним специальным и средним общим образованием (группа 1) с оценками пользователей с высшим и неоконченным высшим образованием (группа 2). На 300 словах второй части опроса было рассчитано среднеквадратическое отклонение оценок респондентов первой и второй группы. У половины слов отклонение оказалось больше у первой группы, у другой половины – у второй. Для первой группы среднее квадратичное отклонение равно 1,039, для второй – 1,046. На рис. 6 приведена диаграмма дисперсии, у которой по оси X размещено среднее отклонение слов у первой группы, по оси Y – у второй. За исключением нескольких выбросов, все остальные точки укладываются вдоль главной диагонали, коэффициент корреляции Пирсона – 0,687. Таким образом, сколько-нибудь значительного различия в дисперсии оценок в зависимости от уровня образования не выявлено.

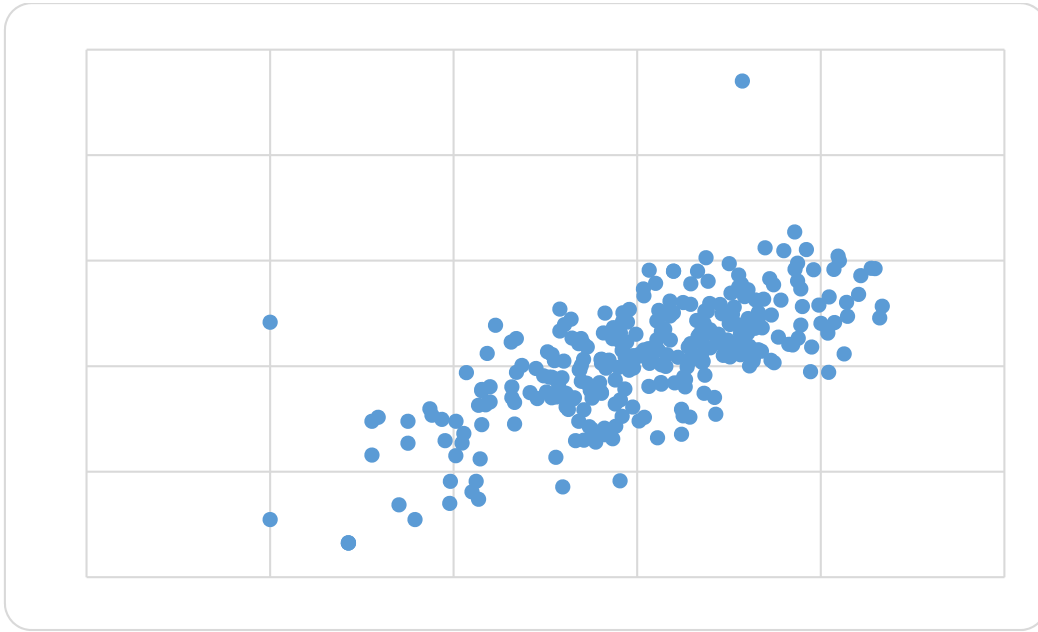

Рис. 6. Дисперсия оценок респондентов по уровням образования /
Fig. 6. Ratings dispersion based on respondents' education

4.5. Сопоставление рейтингов конкретности/абстрактности русских и английских слов

Наличие словарей с рейтингами конкретности/абстрактности для разных языков позволяет провести исследование того, в какой мере концепция конкретности/абстрактности является языково-специфической. В работе

(Solovyev, Ivanov & Akhtyamov 2019a) впервые проведено такое межъязыковое сопоставление: слова из нашего словаря сопоставлены с их английскими эквивалентами и сравнены соответствующие рейтинги. В исследовании использована американская психолингвистическая база MRC (Coltheart 1981), в которой был осуществлен поиск англоязычных аналогов исследуемых русских слов. Межъязыковое сопоставление проведено для 770 слов (из 1000 слов первого опроса). 230 слов из нашего словаря не вошли в сопоставительное исследование по следующим причинам. 1) Отсутствие эквивалентов в английском словаре. Это не только слова, обозначающие этнокультурные реалии, такие как: *милиционер, дача*, но также и названия месяцев, дней недели, по какой-то причине не включенных в MRC. 2) Вторую группу слов составили однозначные слова, которым в английском языке соответствуют разные понятия. Например, *монастырь – monastery* (букв. мужской монастырь) и *convent* (букв. женский монастырь). Рейтинги конкретности/абстрактности русских слов по вышеприведенной формуле конвертированы в формат базы MRC: от 100 (абстрактные) до 700 (конкретные). Фрагмент сопоставления представлен в табл. 3.

Таблица 3. **Рейтинги русских слов и их английских эквивалентов**

#	Слово (рус)	Рейтинг (рус.)	Рейтинг (англ.)	Разница рейтингов	Слово (англ.)
1	сила	340	339	1	strength
2	дерево	606	604	2	tree
3	эффект	288	295	7	effect
	…	…	…	…	…
771	администрация	599	231	268	administration

Table 3. **Russian-English ratings**

#	Russian word	Rating (Rus.)	Rating (Eng.)	Rating difference	English word
1	sila	340	339	1	strength
2	derevo	606	604	2	tree
3	effekt	288	295	7	effect
	…	…	…	…	…
771	administratcija	559	331	268	administration

Коэффициент корреляции Пирсона между рейтингами конкретности/абстрактности русских и английских слов следует признать высоким, он составил 0,78 (Evans 1996). По итогам сопоставления высокая степень различий (выше 67%) рейтингов обнаружена у 46 существительных. При сравнении разницы оценок абстрактных и конкретных слов обнаружено, что большее различие характерно для абстрактных слов. Для проведения такого сравнения слова русского словаря разбиты на 3 равные по величине группы – наиболее конкретных, наиболее абстрактных и слов с промежуточными рейтингами. Для наиболее конкретных слов средняя разница в оценках составила 47 единиц, а для наиболее абстрактных – 56.

Обсудим возможные причины большой разницы между рейтингами конкретности/абстрактности у некоторых пар переводных эквивалентов на

примере слова ***администрация***. В русском языке ***администрация*** имеет только значения: органы управления и должностные лица, возглавляющие организацию (Кузнецов 2006). В то же время в английском, кроме этого слово ***administration*** имеет еще и значение “the activities that are done in order to plan, organize and run a business, school or other institution”, указанное в Oxford Learner’s Dictionaries (ENA, April 17, 2022)[6].

Например: *the day-to-day administration of a company* (там же). Это значение соответствует русскому ***администрирование***. Таким образом, даже у казалось бы точных переводных эквивалентов вполне возможны различия в значениях, причем значения могут различаться именно в аспекте конкретности/абстрактности. Примеры: *Я пошел в администрацию* и *Администрацию университета пора полностью менять* указывают на вполне конкретные значения людей и места. В тоже время ***администрирование*** относится к весьма абстрактному процессу.

Данное исследование позволило сформулировать два основных вывода. Во-первых, русские и английские рейтинги конкретности/абстрактности рассмотренных слов преимущественно расположены в одном и том же сегменте шкалы и во многих случаях весьма близки. Это указывает на то, что концепция конкретности/абстрактности в значительной степени является языковонезависимой, по меньшей мере в пределах культуры западной цивилизации. Во-вторых, важную роль в этой концепции имеет языково-специфический компонент, определяемый разницей культур.

4.6. Многозначные слова

При составлении словарей, подобных нашему, особой проблемой является многозначность слов (Volskaya et al. 2020). Ясно, что разные значения слов вполне могут иметь разные индексы. Однако ранее эта проблема игнорировалась. Нами впервые (Andreeva et al. 2020) предпринята попытка присвоения индексов отдельным значениям слов. С этой целью был проведен отдельный эксперимент. Для каждого заведомо многозначного слова для простоты выбиралось два его разных значения, одно из которых является конкретным, а другое – абстрактным. Значения брались по словарю (Малый академический словарь 1981). Для обоих значений подбирались контексты, в которых реализуются эти значения. Контекст задавался словосочетанием из двух (редко 3–4) слов. Словосочетания составлялись так, чтобы их частотность (по НКРЯ) была примерно одинаковой. В анкеты для оценки включались именно такие словосочетания, в итоге отобраны 206 слов (из 1000). В анкетах словосочетания были сгруппированы по 30 слов (60 сочетаний). Респондентами явились 280 носителей русского языка в возрасте от 18 до 60 лет. Рейтинги нормированы к диапазону 100–700.

[6] https://www.oxfordlearnersdictionaries.com/definition/english/administration?q=administration

Рейтинги конкретности/абстрактности отдельных значений были сопоставлены дважды: (1) друг с другом и (2) с рейтингами слов, оцененных ранее как единое целое. Например, для слова ***дорога*** мы сопоставили (1) оценки двух значений, реализованных в сочетаниях «*проселочная* ***дорога***» (192) и «*собраться в* ***дорогу***» (475); (2) рейтинги обоих этих значений с общей оценкой слова «***дорога***» (199). Как мы видим, в данном конкретном случае два рейтинга, т.е. «*проселочная* ***дорога***» и «***дорога***», близки (192 против 199), в то время как рейтинг сочетания «*собраться в* ***дорогу***» значительно отличается. Первое может свидетельствовать о том, что при восприятии слова ***дорога*** носители языка прежде всего визуализируют физическую дорогу, вроде *проселочной* ***дороги***, а не более абстрактные значения этого слова, такие как "путешествие". Рис. 7 представляет различия в рейтингах многозначных слов, позволяющие оценить степень их разброса на шкале оценок. Средняя разница двух оценок – 204.

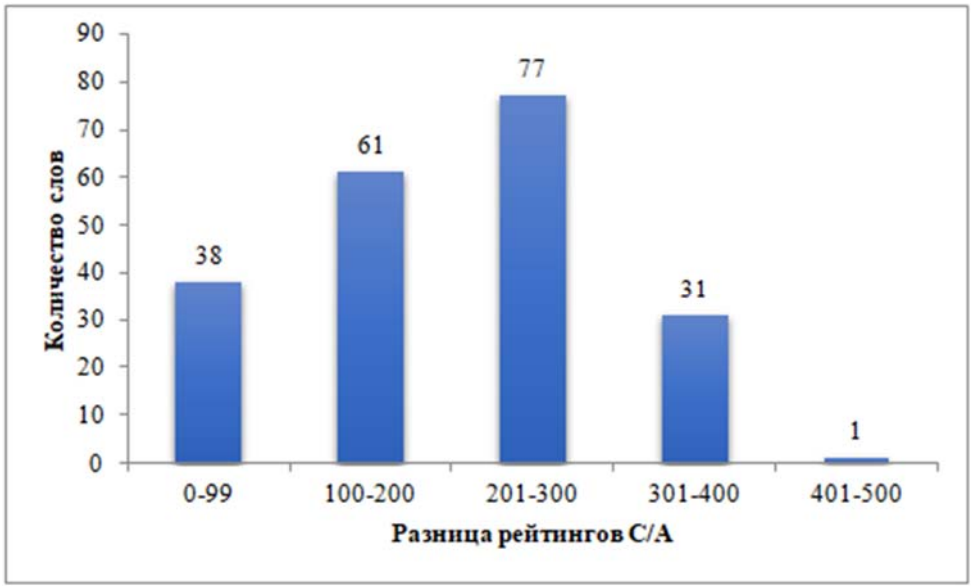

Рис. 7. Разница рейтингов в словосочетаниях / Fig. 7. Ratings difference in word combinations

Максимальная разница в рейтингах (более 400 единиц) обнаруживается, в частности, в слове ***поворот***, определяемом как: 1) «*место, где дорога поворачивает, отклоняется в сторону*»; 2) «*полное изменение в развитии чего-либо*» (Малый академический словарь 1981). Рейтинги сочетаний «***поворот*** *налево от дома*» (129) и «***поворот*** *судьбы*» (540) указывают на различие между оценками респондентов конкретных и абстрактных значений слова. Рейтинг слова ***поворот*** в целом, без разделения на значения, равен 464, т.е. ближе ко второму абстрактному значению. Полученные результаты показывают, что имеет смысл выделять отдельные значения многозначных слов и оценивать степень их конкретности/абстрактности в словосочетаниях, иллюстрирующих только один смысл.

4.7. Сравнение с данными машинного словаря

После создания машинных словарей и до их использования целесообразно проанализировать характер машинных оценок, провести количественное и качественное сравнение их с человеческими. Сопоставим оценки третьей версии машинного словаря с оценками респондентов 1300 слов в нашем словаре. Разница между машинными и человеческими оценками варьируется в диапазоне от 1,28 до 2,1. Отрицательная разница означает, что машинный рейтинг оказался меньше. Большая часть слов (916 из 1300) получила оценки с небольшой разницей в интервале от –0,4 до 0,8 (рис. 8).

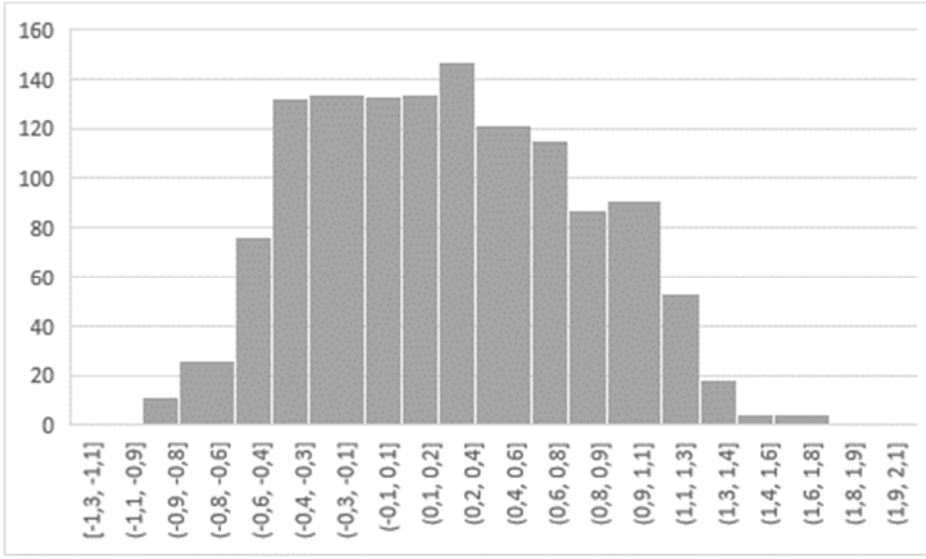

Рис. 8. Разница в оценках / Fig. 8. Ratings' difference

Рассмотрим слова с наибольшей разницей в оценках. Это 114 слов, которые были оценены с разницей от –0,5 до –1,28 (рис. 9), и 155 слов с разницей от 1 до 2,1 (рис. 10). Слова с наибольшей отрицательной разницей – это, как правило, существительные, которые по данным машинного словаря оценивались как более конкретные (с рейтингом меньше 2), а по оценкам респондентов – как менее конкретные. Слова, которые получили наибольшую положительную разницу в оценках, по данным машинного словаря, являются, как правило, более абстрактными (с рейтингом больше 4).

Важно отметить тенденцию, которая обнаруживается при анализе. По оценкам респондентов, большая часть слов получила срединное значение от 2,5 до 3,4 (496 слов), выявлено всего 19 слов со степенью, близкой к 5, и всего 96 слов со степенью от 1 до 1,5, т.е. большая часть опрашиваемых при прохождении опроса не выбирала на шкале крайние значения – 1 или 5. Однако, по данным машинного словаря, единиц со срединными оценками выявлено меньше – 378 существительных, а единиц со степенью, приближенной к крайним значениям, напротив, обнаружено больше, а именно 242 лексемы со степенью от 4 до 5, 211 – со степенью от 1 до 1,5 (рис. 11).

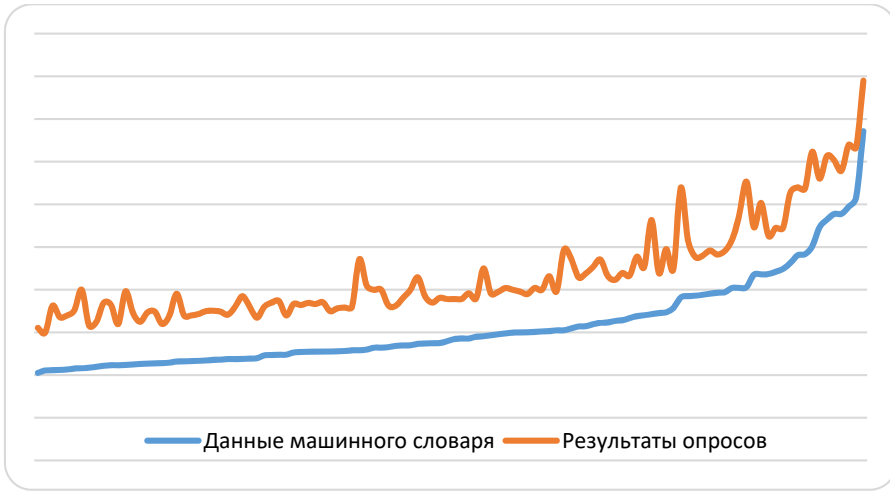

Рис. 9. Отрицательная разница между данными машинного словаря и оценками респондентов /
Fig. 9. Negative difference between machine dictionary and survey results data

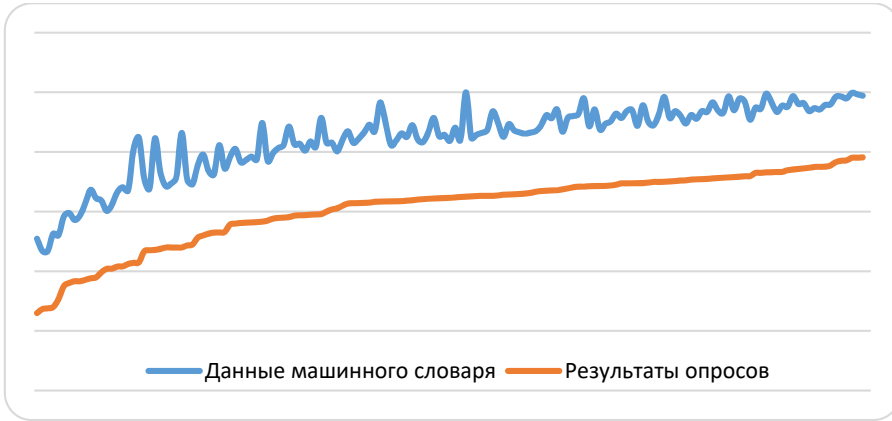

Рис. 10. Положительная разница между данными машинного словаря и оценками респондентов /
Fig. 10. Positive difference between machine dictionary and survey results data

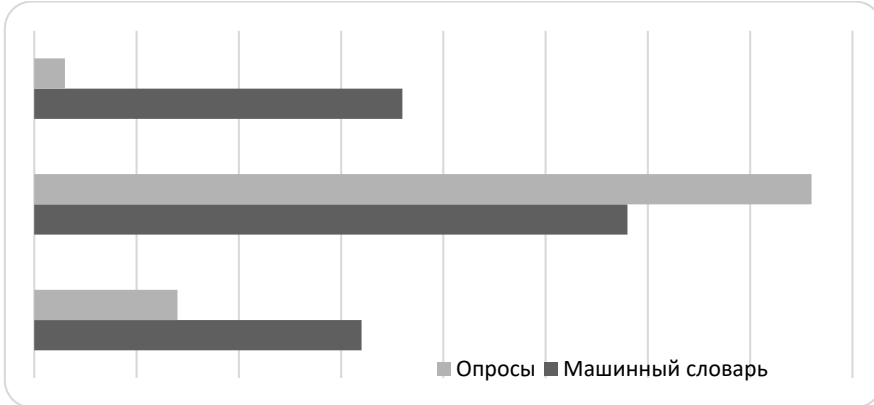

Рис. 11. Количество оценок с крайними и средними значениями /
Fig. 11. Ratings with extreme and mean values

Подобная ситуация означает, что респонденты не склонны к резким оценкам, что является следствием учета ими определенного собственного опыта (Pasquale et al. 2010) либо редких значений многозначных слов, в некоторых из которых слово можно расценивать как абстрактное, а в других – как конкретное. Это приводит к сдвигу оценок к середине шкалы. Такие редкие значения могут быть не представлены в должной мере в корпусах текстов, на которых обучаются нейронные сети.

5. Лингвистические исследования: конкретность vs специфичность

В лингвистике используется семантическая категория, близкая к конкретности, – специфичность. Кажется, что между этими категориями есть корреляция, и поэтому их не всегда различают. Скажем, понятие «*диван*» более специфическое, чем понятие «*мебель*», и одновременно слово *диван* более конкретное, чем *мебель.* Возникает вопрос: в какой мере эти две категории коррелируют? Первое подобное исследование на эмпирическом материале для английского языка было проведено в работе (Bolognesi et al. 2020). В ней показано, что корреляция есть, но умеренная – 0,361 по Спирмену. В нашей работе (Ivanov & Solovyev 2021) ставятся те же цели, что и в указанной публикации, но исследование проводится для русского языка, при этом, естественно, меняются используемые внешние лингвистические ресурсы. Исследование ограничено именами существительными для обеспечения сопоставимости с работой (Bolognesi et al. 2020), а также в связи с тем, что именно для существительных иерархические отношения описаны наиболее подробно.

Категория специфичности/общности интуитивно представляется достаточно понятной. Важный вклад в ее изучение внесли классические работы Рош (Rosch 1975). После создания тезауруса WordNet (Fellbaum 1998) степень специфичности/общности обычно оценивается по положению единицы в иерархии тезауруса в тезаурусе WordNet (Devitt & Vogel 2004). Структура WordNet и ее релевантность лингвистическим фактам представлена в (Miller 1998). Чем понятие, представленное синсетом (синонимическим рядом) WordNet, ближе к нижним уровням тезауруса, тем оно более специфично. Это можно автоматически оценить количественно. Для этого мы используем формулу, предложенную в (Bolognesi 2020): рейтинг специфичности – $(1 + d)/D$, где d – общее число гиперонимов (прямых и непрямых) целевого слова и D – максимальное расстояние от листьев до вершины иерархии. Для WordNet эта величина равна 20. В используемом нами тезаурусе русского языка RuThes (Лукашевич 2011) D = 13. Тезаурус RuThes (ENA, April 17, 2022)[7] содержит более 31,5 тыс. понятий, 111,5 тыс. различных текстовых входов (слов и выражений русского языка). Рейтинг специфичности стандартизирован – приведен к 5-балльной шкале.

[7] http://www.labinform.ru/pub/ruthes/index.htm

Значения конкретности и специфичности всех рассматриваемых нами 14294 слов (общих для RuThes и словаря конкретности) русского языка приведены в файле Concreteness Ratings in RuThes на сайте проекта «Технологии создания семантических электронных словарей» (ENA, April 2017, 2022)[8]. Коэффициент корреляции Спирмена между рейтингами конкретности и специфичности оказался равен 0,264, Пирсона – 0,256 ($p < 0{,}001$). Для английского языка коэффициенты корреляции – 0,361 и 0,354 соответственно (Bolognesi, Burgers & Caselli 2020).

Установленный нами коэффициент корреляции, хотя и является положительным и статистически значимым, классифицируется как слабый (Evans 1996). Таким образом, на материале русского языка подтверждается качественный результат работы (Bolognesi, Burgers & Caselli 2020), что указывает на независимость параметров «конкретность» и «специфичность» и необходимость их самостоятельного изучения. Специфические концепты могут быть как конкретными, так и абстрактными. Рис. 12 визуализирует распределение слов по параметрам конкретность-специфичность. По рисунку видно отсутствие явной корреляции между этими двумя параметрами.

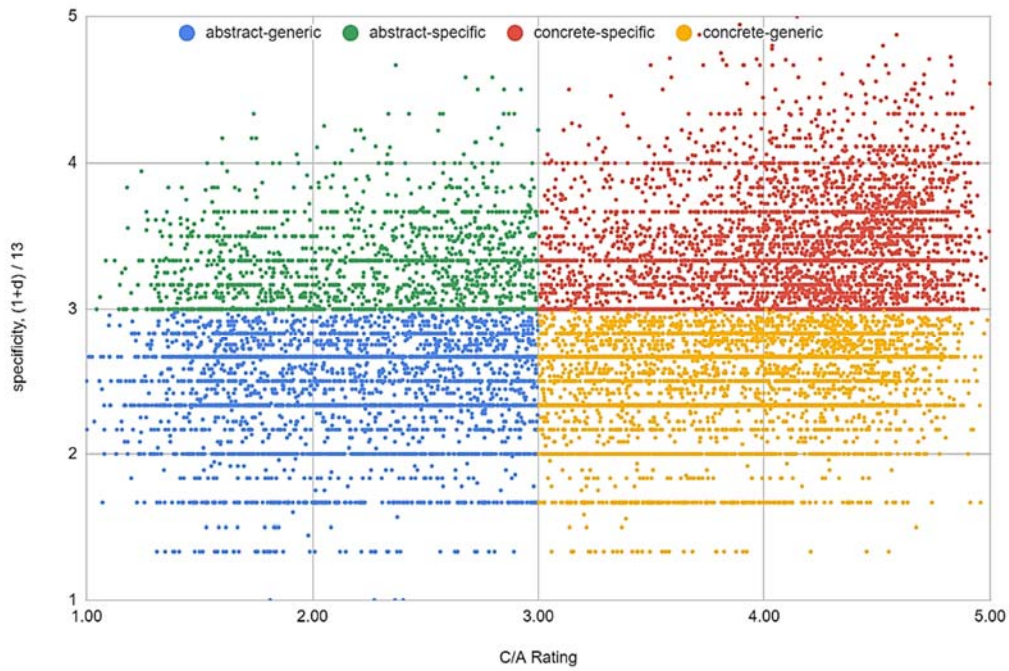

Рис. 12. Распределение слов в двумерном пространстве конкретность-специфичность (Ivanov & Solovyev 2021) /
Fig. 12. Word distribution in two-dimensional space of concreteness-specificity (Ivanov & Solovyev 2021)

[8] https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html

Сопоставление категорий специфичности и конкретности имеет важные импликации для когнитивной науки. В (Bolognesi et al. 2020) выдвинуто предположение, что специфичность отражает характер структурирования мира языком, в то время как конкретность – структурирования мира сознанием в широком смысле, включая перцептивный уровень. Различие этих двух категорий (низкая степень корреляции) является аргументом против «сильной» версии гипотезы лингвистической относительности Сепира–Уорфа, предполагающей, что язык определяет мышление.

В следующем разделе мы перейдем к возможным практическим приложениям словаря и опишем одно уже реализованное его применение – к анализу сложности текстов.

6. Приложения словаря: сложность текстов

Доля абстрактных слов как признак сложности текста рассматривалась рядом исследователей (Taylor & Weir 2012, Fisher et al. 2017). Корреляция между абстрактностью и сложностью текста была также продемонстрирована в работах российских ученых, проводивших исследование на материале русскоязычных текстов (Криони и др. 2008, Solovyev et al. 2019b).

В работе (Solovyev et al. 2020b) изучалась вариативность рейтингов в образовательных текстах. Сопоставлялись учебники: (1) для начальных и старших классов, (2) для гуманитарных и естественнонаучных дисциплин, а также (3) тексты-оригиналы и пересказы. Материал исследования составили Учебный корпус русского языка (УКРЯ) и Корпус пересказов (КП). УКРЯ включает 74 учебника общим объемом 3 млн словоупотреблений (табл. 4). В УКРЯ вошли учебники 2006–2020 гг. выпуска по гуманитарным и естественнонаучным дисциплинам.

Таблица 4. **Учебный корпус русского языка (УКРЯ)**

Класс	Количество словоупотреблений		
	Естественнонаучные дисциплины	Гуманитарные дисциплины	Итого
1	21304	4757	26061
2	29284	28235	57519
3	53565	-	53565
4	51489	24621	76110
5	102467	19527	121994
6	-	159664	159664
7	75205	111788	186993
8	-	273251	273251
9	88335	390821	479156
10	207271	656072	863343
11	-	436322	436322
Итого	628920	2105058	2733978

Table 4. **Russian Academic Corpus**

Grade	Number of words		
	Sciences	Humanities	TOTAL
1	21304	4757	26061
2	29284	28235	57519
3	53565	-	53565
4	51489	24621	76110
5	102467	19527	121994
6	-	159664	159664
7	75205	111788	186993
8	-	273251	273251
9	88335	390821	479156
10	207271	656072	863343
11	-	436322	436322
Итого	628920	2105058	2733978

Корпус пересказов (КП) – это результат исследования, направленного на оценку влияния связности текста на его восприятие читателями (McCarthy et al. 2019). В исследовании участвовали 289 респондентов в возрасте 11–12 лет, ученики 5-го класса. Респондентов индивидуально просили прочитать один из учебных текстов, оригинальный текст (ОТ) и модифицированный текст (МТ), оба состояли примерно из 200 слов. Тексты представляли собой фрагменты главы из учебника Боголюбов Л.Н. Обществознание 5 класс. Учебник для средних школ. 3-е издание. Просвещение, 127 (2013). Пересказы текстов транскрибировались экспертами. Общий размер корпуса составляет 6473 словоупотребления, он доступен на сайте проекта «Технологии создания семантических электронных словарей».

Поскольку в корпусе учебных текстов содержится значительно больше слов, чем в нашем словаре, созданном с помощью опроса респондентов, был использован машинный словарь на 88 тыс. словоформ. По техническим причинам в нем рейтинг принимает значение в диапазоне от –5 (наиболее абстрактные) до +5,5 (наиболее конкретные). Рейтинг текста определяется как среднее арифметическое рейтингов всех входящих в него слов. Если слово из текста отсутствует в словаре, то оно не учитывается. Средние индексы абстрактности текстов учебников, представленные в табл. 5, подсчитаны при помощи онлайн-инструмента RuLingva (Solovyev et al. 2020b), использующего созданный словарь.

Средний индекс указывает на следующее: а) самый высокий индекс конкретности демонстрируют тексты по биологии и учебники для начальной школы, при этом конкретность учебников по биологии для средней школы является самой высокой – +0,49, что даже выше, чем у текстов начальной школы (+0,34); б) учебники по обществознанию имеют самый высокий уровень абстрактности, а учебники по истории расположены в середине шкалы с оценкой «0»; в) индекс абстрактности растет в классах с 1-го по 11-й. Оценка статистической значимости зависимости индекса абстрактности от класса методом линейной регрессии дает значение $p = 1.69e^{-10}$.

Таблица 5. **Индексы текстов учебников и пересказов**

Дисциплина	Класс	Средний индекс
Начальная школа	1-4	+0,34
Биология	5-7	+0,49
Биология	9-10	+0,15
История	10-11	0
Обществознание	5-8	-0,11
Обществознание	9-11	-0,15
Литература	6-8	+0,08
Литература	9-11	-0,14
Тексты МТ	5	0,12
Тексты ОТ	5	0,17
Пересказы	5	0,27

Table 5. **Ratings of recalls and textbooks**

Subject	Grade	Mean rating
Primary school	1-4	+0,34
Biology	5-7	+0,49
Biology	9-10	+0,15
History	10-11	0
Social studies	5-8	-0,11
Social studies	9-11	-0,15
Literature	6-8	+0,08
Literature	9-11	-0,14
MT Texts	5	0,12
OT Texts	5	0,17
Recalls	5	0,27

Тексты ОТ и МТ, предложенные респондентам для пересказа, имеют близкие средние индексы. Как видно из приведенной выше таблицы, средний индекс для пересказов выше, чем у исходных текстов, что подтверждает вывод: респонденты склонны опускать более абстрактные слова и сохранять конкретные при пересказе. Сравнение показателей пересказов и исходных текстов на основе критерия Стьюдента ($p = 0{,}0003$) подтверждает гипотезу о том, что эта разница статистически значима.

Таким образом, созданный нами инструментарий – словари, в том числе машинные, и программа RuLingva – позволяет рассчитывать уровень абстрактности текстов. Полученные на учебных текстах данные подтверждают гипотезы исследования. Аналогичным образом словарь использовался и в других работах по сложности текстов (Солнышкина и др. 2021, Gizatulina et al. 2020).

7. Заключение

Словари являются важным инструментом междисциплинарных исследований концепции конкретности/абстрактности. Словари с рейтингами конкретности/абстрактности созданы для целого ряда языков. В статье описывается первый словарь такого рода для русского языка. Словарь содержит 1500 наиболее частотных существительных. Подробно описана методология создания, которая может быть полезной для создания словарей конкретности/абстрактности для других языков, равно как и для составления других семантических словарей. Методология создания включает многоуровневую систему очистки данных, впервые последовательно примененную в таком масштабе. Выполнено тщательное исследование влияния на результат различных аспектов создания словаря, таких как возраст и уровень образования респондентов. Проведен отдельный эксперимент по оценке многозначных слов. Для других языков подобные исследования не проводились. Нами предложена и опробована методика оценки рейтингов для отдельных значений слов, которая может быть рекомендована к использованию при

построении словарей рейтингов для других языков. Приведено сравнение наших данных с характеристикой конкретности/абстрактности в Русском семантическом словаре Н.Ю. Шведовой, с известными критериями абстрактности.

В дополнение к словарю, созданному путем опроса респондентов, составлен также компьютерный словарь значительно больших размеров, в котором рейтинги конкретности/абстрактности получены путем экстраполяции имеющихся рейтингов респондентов с применением наиболее современных технологий глубокого обучения нейронных сетей. Показано, что качество компьютерных словарей и созданных людьми вполне сопоставимо. Проведенный качественный и количественный анализ данных машинного словаря выявил характер его расхождений с человеческими оценками, что может быть учтено при подборе слов в прикладных исследованиях.

В статье продемонстрированы возможности применения словаря в фундаментальных теоретических исследованиях, направленных на изучение репрезентации в сознании человека таких понятий, как конкретность, абстрактность, специфичность, а также и в прикладных исследований, в частности, для оценки сложности текстов.

В аспекте чисто лингвистических исследований количественно оценена эффективность такого хорошо известного критерия абстрактности, как наличие специфических аффиксов. Проведено сопоставление рейтингов для слов русского языка и их переводных эквивалентов на английском. С одной стороны, имеется высокий уровень корреляции между ними, с другой – выявлены слова с существенным расхождением оценок, что указывает на языковую зависимость категории конкретности/абстрактности. Концепция конкретности близка к концепции специфичности. В связи с этим интересно проанализировать их соотношение. Для русского языка реплицировано исследование, ранее проведенное для английского языка, и показано, что между ними имеется лишь слабая корреляция, поэтому они должны изучаться независимо друг от друга.

Важным приложением словаря конкретности/абстрактности является определение уровня сложности текстов. Чем в тексте больше абстрактных слов, тем он сложнее для восприятия. Поэтому доля абстрактных слов является одним из ключевых параметров, определяющих сложность текстов. С этой точки зрения проанализированы тексты учебников для средней школы.

Подводя итоги серии исследований, можно отметить, что созданные словари имеют высокий уровень качества, достаточный объем и позволяют проводить разнообразные теоретические и прикладные исследования.

Мы планируем проводить дальнейшие исследования в трех направлениях: 1. Повышение качества компьютерных словарей за счет использования более совершенных технологий. 2. Создание словаря с рейтингами позитивности/негативности слов. 3. Лингвистические исследования «эффекта конкретности».

Благодарность

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета.

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Andreeva, Mariia, Marina Solnyshkina, Artem Zaikin, Olga Bukach & Radif Zamaletdinov. 2020. Assessment of comparative abstractness: Quantitative approach. *Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020)*. 132–144.

Black, Paul. 2019. Manhattan distance. *In Dictionary of Algorithms and Data Structures [Online]*. http://www.nist.gov/dads/HTML/manhattanDistance.html. (accessed 19.04.2022)

Bolognesi, Marianna, Burgers Christian & Caselli Tommaso. 2020. On abstraction: Decoupling conceptual concreteness and categorical specificity. *Cognitive Processing* 21 (3). 365–381. DOI: https://doi.org/10.1007/s10339-020-00965-9.

Borghi, Anna M., Ferdinand Binkofski, Cristiano Castelfranchi & Felice Cimatti. 2017. The challenge of abstract concepts. *Psychol. Bull* 143. 263–292.

Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014a. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46 (3). 904–911.

Brysbaert, Marc, Michaël Stevens, Simon De Deyne, Simon De Deyne & Gert Storms. 2014b. Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica* 150. 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010

Chandola, Varun, Arindam Banerjee & Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* 41(3). 1–58.

Charbonnier, Jean & Wartena Christian. 2019. Predicting word concreteness and imagery. *In Proceedings of the 13th International Conference on Computational Semantics-Long Papers.* 176–187.

Cristianini, Nello & John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press.

Coltheart, Max. 1981. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology* 33A. 497–505.

Dallin, J Bailey, Christina Nessler, Kiera N Berggren & Julie L Wambaugh. 2020. An Aphasia treatment for verbs with low concreteness: A pilot study. *American Journal of Speech-Language Pathology* 29 (1). 299–318.

de Groot, Annette M. 1989. Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition 15*(5). 824–845. https://doi.org/10.1037/0278-7393.15.5.824

Devitt, Ann & Vogel Carl. 2004. The Topology of WordNet: Some Metrics. *GWC Proceedings*. 106–111.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv*:1810.04805.

Evans, James D. 1996. *Straightforward Statistics for the Behavioral Ssciences*. Brooks/Cole Publishing, Pacific Grove.

Fellbaum, Christiane. 1998. *Wordnet: An Electronic Lexical Database.* MIT Press. Cambridge, Massachusetts.

Fisher, Douglas, Frey Nancy & Lapp Diane. 2016. *Text Complexity: Stretching Readers with Texts and Tasks*. Corwin Press.

Fliessbach, Klaus, Susanne Weis, Peter Klaver, Christian E. Elger & Bernhard Weber. 2006. The effect of word concreteness on recognition memory. *NeuroImage* 32 (3). 1413–1421. https://doi.org/10.1016/j.neuroimage.2006.06.007

Gizatulina, Diana, Farida Ismaeva, Marina Solnyshkina, Ekaterina Martynova & Iskander Yarmakeev. 2020. Fluctuations of text complexity: The case of Basic State Examination in English. *In SHS Web of Conferences* 88. EDP Sciences.

Ivanov, Vladimir & Solovyev Valery. 2021. The Relation of Categories of Concreteness and Specificity: Russian Data. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021*". URL: http://www.dialog-21.ru/media/5260/ivanovvplussolovyevv049.pdf. (accessed 19.04.2022).

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou & Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv*:1612.03651.

Kousta, Stavroula-Thaleia, Gabriella Vigliocco, David P Vinson & Mark Andrews. 2011. The representation of abstract words: Why emotion matters. *Exp Psychol Gen. Feb*. 140 (1). 14–34. https://doi.org/10.1037/a0021446.

Krioni, Nikolay K., Alexey D. Nikitin & Anastasiya V. Fillipova. 2008. Avtomatizirovannaya sistema analiza slozhnosti uchebnyh tekstov. *Bulletin of Ufa State Technical University of Aviation* 11. 1 (28). 101–107. (In Russ.)

Kuznecov, Sergey A. 2006. *Bol'shoy Tolkovy Slovar' Russkogo Yazyka.* Norint. (In Russ.)

Laming, Donald. 2004. *Human Judgement: The Eye of the Beholder*. London: Thompson Learning.

Lukashevich, Natilia V. 2011. *Thesauruses in Information Search Tasks.* M.: Izd-vo Moskovskogo universiteta. (In Russ.)

Maximilian, Köper & Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2595–2598.

McCarthy, Kathryn Soo, Danielle Siobhan Mcnamara, Marina I. Solnyshkina, Fanuza Kh. Tarasova & Roman V. Kupriyanov. 2019. The Russian language test: Towards assessing text comprehension. *Vestnik Volgogradskogo Gosudarstvennogo Universiteta*. Seriiâ˜2, Iazykoznanie; Volgograd 18 (4). 231–247.

McNamara, Danielle, Arthur C. Graesser, Philip M. Mccarthy & Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge, MA: Cambridge University Press.

Mestres-Missé, Anna, Thomas F. Münte & Antoni Rodriguez-Fornells. 2014. Mapping concrete and abstract meanings to new words using verbal contexts. *Second Language Research* 30 (2). 191–223.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546.*

Miller, George A. 1998. Nouns in WordNet. In Christiane Fellbaum (ed.), *Wordnet: An electronic lexical database mit press*. Cambridge, Massachusetts.

Mkrtychian, Nadezhda, Evgeny Blagovechtchenski, Diana Kurmakaeva, Daria Gnedykh, Svetlana Kostromina & Yury Shtyrov. 2019. Concrete vs. Abstract Semantics: From mental representations to functional brain mapping. *Frontiers in Human Neuroscience* 13. 267. https://doi.org/10.3389/fnhum.2019.00267

Naumann, Daniela, Diego Frassinelli & Sabine Schulte im Walde. 2018. Quantitative semantic variation in the contexts of concrete and abstract words. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, New Orleans, LA*. 76–85.

Paivio, Allan. 1965. Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behaviour* 4. 32–38. https://doi.org/10.1016/s0022-5371(65)80064-0

Paivio, Allan. 1990. *Dual Coding Theory, in Mental Representations: A Dual Coding Approach*. Oxford: Oxford University Press. 53–83. https://doi.org/10.1093/acprof:oso/9780195066661.003.0004

Pasquale, A. Della Rosa, Eleonora Catricalà, Gabriella Vigliocco & Stefano F. Cappa. 2010. Behavior Research Methods Beyond the abstract–concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian. *Behavior Research Methods* 42 (4). 1042–1048. https://doi.org/10.3758/BRM.42.4.1042

Peti-Stantić, Anita, Maja Anđel, Vedrana Gnjidić, Gordana Keresteš, Nikola Ljubešić, Irina Masnikosa, Mirjana Tonković, Jelena Tušek, Jana Willer-Gold & Mateusz-Milan Stanojević. 2021. *The Croatian Psycholinguistic Database: Estimates for 6000 Nouns, Verbs, Adjectives and Adverbs.* 1–18. https://doi.org/10.3758/s13428-020-01533-x

Reilly, Megan, & Rutvik H. Desai. 2017. Effects of semantic neighborhood density in abstract and concrete words*. Cognition* 169. 46–53. https://doi.org/10.1016/j.cognition.2017.08.004

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psycholology: General* 104 (3). 192–233.

Sadoski, Mark, Wiliam A. Kealy, E. T. Goetz & Allan Paivio. 1997. Concreteness and imagery effects in the written composition of definitions. *Journal of Educational Psychology* 89(3). 518–526. https://doi.org/10.1037/0022-0663.89.3.518

Sadoski, Mark. 2001. Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text. *Educational Psychology Review* 13(3). 263–281.

Schmid, Hans-J¨org. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. *Topics in English Linguistics.* Berlin: Mouton de Gruyter.

Schwanenflugel, Paula J. & Edward J. Shoben. 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9 (1). 82–102. https://doi.org/1037/0278-7393.9.1.82

Schwanenflugel, Paula J., Carolyn Akin & Wei-Ming Luh. 1992. Context availability and the recall of abstract and concrete words. *Memory & Cognition* 20 (1). 96–104. https://doi.org/10.3758/bf03208259

Snefjella, Bryor, Michel Généreux & Victor Kuperman. 2019. Historical evolution of concrete and abstract language revisited. *Behavior Research Methods* 51 (4). 1693–1705.

Solnyshkina, Marina I., Radif. R. Zamaletdinov, Ehl'zara Gizzatullina-Gafiyatova, Diana Gizatulina & Maria Begaeva. 2021. Mnogofaktorny analiz slozhnosti teksta. *Inostrannye Yazyki v Shkole*. 28–34. (In Russ.)

Solovyev, Valery D., Vladimir V. Ivanov & Rauf B. Akhtiamov. 2019a. Dictionary of abstract and concrete words of the Russian language: A methodology for creation and application. *Journal of Research in Applied Linguistics* 10. 215–227.

Solovyev, Valery, Mariia Andreeva, Marina Solnyshkina, Radif Zamaletdinov, Andrey Danilov & Dina Gaynutdinova. 2019b. Computing concreteness ratings of Russian and English most frequent words: Contrastive approach. *In the Proceedings of the 12th International Conference on Developments in eSystems Engineering (DeSE)*. 403–408.

Solovyev, Valery D., Vladimir V. Bochkarev & S. V. Khristoforov. 2020a. Generation of a dictionary of abstract/concrete words by a multilayer neural network. *Journal of Physics: Conference Series* 1680 (1). 012046.

Solovyev, Valery, Marina Solnyshkina, Mariia Andreeva, Andrey Danilov & Radif Zamaletdinov. 2020b. Text Complexity and Abstractness: Tools for the Russian

Language. *Proceedings of the International Conference "Internet and Modern Society"*. 75–87.
Solovyev, Valery. 2021. Concreteness/Abstractness Concept: State of the Art. *Advances in Intelligent Systems and Computing* 1358. 275–283.
Spreen, Otfried & Rudolph W. Schulz. 1966. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior* 5. 459–468.
Taylor, Linda & Weir Cyril J. 2012. *IELTS Collected Papers 2: Research in Reading and Listening Assessment* 2. Cambridge University Press.
Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37. 141–188.
Vergallito, Alessandra, Marco Alessandro Petilli & Marco Marelli. 2020. Perceptual modality norms for 1,121 Italian words: A comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks. *Behavior Research Methods*. 1–18.
Vinogradov, Victor V. 2001. Russian language (Grammatical studies of a word). *Russian Language*. (In Russ.)
Vol'skaia, Iulia A. 2020. Creating a dictionary of abstract beings in the Russian language: A criterion for selecting vocabulary. *Philology and Culture* 1 (59). 13–17. (In Russ.)
Volskaya, Yulia A., Irina S. Zhuravkina & Alexander P. Lobanov. 2020. Dictionary of abstract the words of the Russian language: Nouns with high numerical measure of abstractness. *International Journal of Criminology and Sociology* 9. 2398–2405.
Wang, X. & Y Bi. 2021. Idiosyncratic tower of Babel: Individual differences in word-meaning representation increase as word abstractness increases. *Psychological Science 32*(10). 1617–1635.
Yao, Zhao, Jia Wu, Yanyan Zhang & Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behav Res* 49. 1374–1385. https://doi.org/10.3758/s13428-016-0793-2
Zhuravkina, Irina, Valery Soloviev, Alexander Lobanov & Andrey Danilov. 2020. Comparative analysis of concreteness abstractness of Russian words. *In Conference of Open Innovation Association, FRUCT*. 464–470.

Dictionaies and internet resources / Словари и интернет-ресурсы

Lyashevskay Olga N. & Sharoff S.A. 2009. New Russian frequency dictionary. (In Russ.) http://dict.ruslang.ru/freq.php (accessed 28.12.2021).
Small Academic Dictionary. 1981–1984. (In Russ.) https://gufo.me/dict/mas (accessed 28.05.2021).
Russian National Corpus. (In Russ.) http://www.ruscorpora.ru/ (accessed 28.12.2021).
Russian Semantic Dictionary. 1998. In Shvedova N.Yu. (ed.). 'Azbukovnik' (In Russ.)
RuThes Thesaurus. (In Russ.) http://www.labinform.ru/pub/ruthes/index.htm (accessed 28.12.2021).
Technologies of Compiling Semantic Electronic Dictionaries. (In Russ.) https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html (accessed 28.12.2021).
Cohmetrix. http://cohmetrix.com/ (accessed 28.12.2021).
Corpus of Contemporary American English. https://www.english-corpora.org/coca (accessed 28.05.2021).
Google Books Ngram. https://books.google.com/ngrams (accessed 28.12.2021).
FastText. Library for efficient text classification and representation learning. https://fasttext.cc/ (accessed 28.12.2021).

Приложение / Application

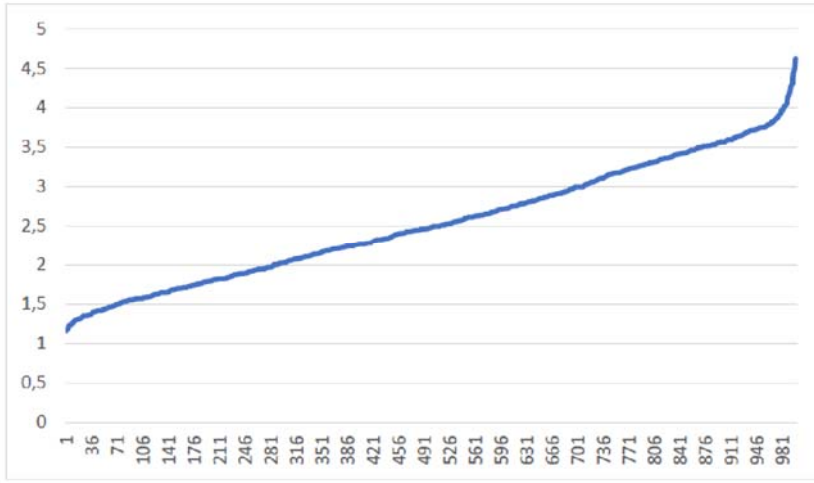

Рис. S1. График распределения оценок, упорядоченных по величине / Fig. S1. Graph of ratings distribution sorted by value

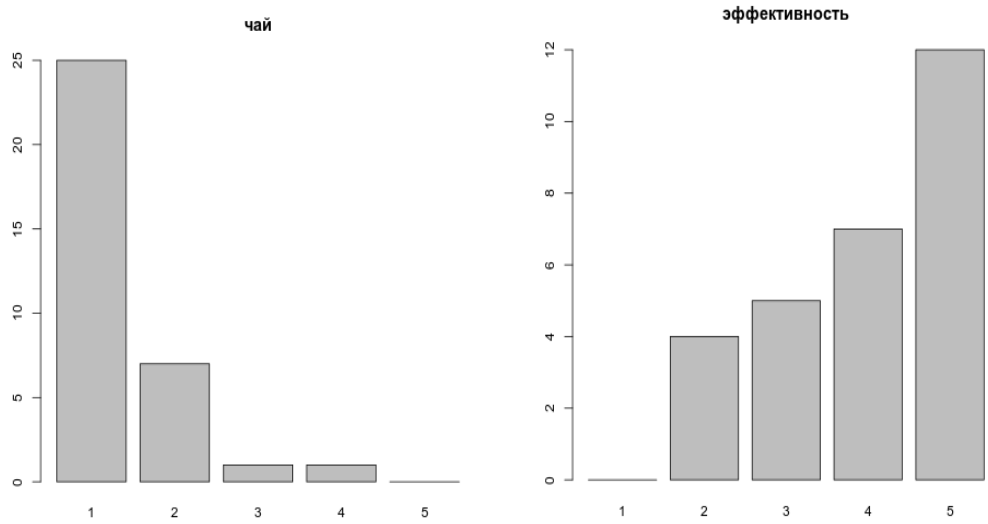

Рис. S2. Характерные распределения оценок для типичных конкретного и абстрактного слова / Fig. S2. Characteristic distributions of typical ratings of concreteness and abstractness of a word

Article history:
Received: 19 November 2021
Accepted: 21 January 2022

Bionotes:
Valery D. SOLOVYEV is Doctor Habil. of Physics and Mathematics, Professor, Chief Researcher of the Text Analytics Research Laboratory at Kazan (Volga Region) Federal University. His research interests embrace cognitive sciences, computer linguistics and text complexity.

Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
e-mail: maki.solovyev@mail.ru
ORCID: 0000-0003-4692-2564

Yulia A. VOLSKAYA is Assistant Professor of the Department of Applied and Experimental Linguistics, and Junior Research Fellow of the Neurocognitive Research Laboratory at Kazan (Volga Region) Federal University. Her research interests include applied linguistics and semantics.
Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
e-mail: kovaleva95julia@mail.ru
ORCID: 0000-0001-8276-5864

Mariia I. ANDREEVA holds a PhD degree in Philology and is Associate Professor of the Department of Foreign Languages at Kazan State Medical University. She is also Junior Research Fellow of the Text Analytics Research Laboratory at Kazan (Volga Region) Federal University. Her research interests are focused on semantics, sociolinguistics and text analysis.
Contact information:
Kazan State Medical University
49 Butlerov St., Kazan, 420012, Russia
email: lafruta@mail.ru
ORCID: 0000-0002-5760-0934

Artem A. ZAIKIN is Doctor of Physics and Mathematics and Associate Professor of the Department of Mathematical Statistics at Kazan (Volga Region) Federal University. He is also Research Fellow of the Research Laboratory investigating the state and evolution of underground tanks and Junior Research Fellow of TRIZ Modeling Center of Research and Education. His research interests are focused on mathematical statistics.
Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
e-mail: kaskrin@gmail.com
ORCID: 0000-0002-5596-3176

Сведения об авторах:
Валерий Дмитриевич СОЛОВЬЕВ – доктор физико-математических наук, профессор, главный научный сотрудник научно-исследовательской лаборатории «Текстовая аналитика» Казанского (Приволжского) федерального университета. Сфера его научных интересов охватывает когнитивную науку, компьютерную лингвистику и сложность текстов.

Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, г. Казань, ул. Кремлевская, 18
e-mail: maki.solovyev@mail.ru
ORCID: 0000-0003-4692-2564

Юлия Александровна ВОЛЬСКАЯ – ассистент кафедры прикладной и экспериментальной лингвистики, младший научный сотрудник научно-исследовательской лаборатории «Нейрокогнитивные исследования» Казанского (Приволжского) федерального университета. В сферу ее научных интересов входят прикладная лингвистика и семантика.

Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, г. Казань, ул. Кремлевская, 18
e-mail: kovaleva95julia@mail.ru
ORCID: 0000-0001-8276-5864

Мария Игоревна АНДРЕЕВА – кандидат филологических наук, доцент кафедры иностранных языков Казанского государственного медицинского университета, младший научный сотрудник научно-исследовательской лаборатории «Текстовая аналитика» Казанского (Приволжского) федерального университета. Сфера ее научных интересов включает семантику, социолингвистику и анализ текста.

Контактная информация:
Казанский государственный медицинский университет
Россия, 420012, г. Казань, ул. Бутлерова, 49
e-mail: lafruta@mail.ru
ORCID: 0000-0002-5760-0934

Артем Александрович ЗАИКИН – кандидат физико-математических наук, доцент кафедры математической статистики Казанского (Приволжского) федерального университета, научный сотрудник Научно-исследовательской лаборатории изучения состояния и эволюции подземных резервуаров, младший научный сотрудник научно-образовательного центра «Моделирование ТРИЗ». Основная сфера его научных интересов – математическая статистика.

Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, г. Казань, ул. Кремлевская, 18
e-mail: kaskrin@gmail.com
ORCID: 0000-0002-5596-3176

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 550–557
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30209

Book review

Review of Sean Wallis. 2021. *Statistics in Corpus Linguistics: A New Approach*. New York/Oxon, Routledge. ISBN 9781138589384 ISBN 9780429491696 (eBook)

Irina PRIVALOVA[1] ✉ **and Mariia KAZACHKOVA**[2]

[1]*Kazan (Volga region) Federal University, Kazan, Russia*
[2]*Moscow State Institute of International Relations (University), Moscow, Russia*
✉angladkova@gmail.com

For citation:
Privalova, Irina & Mariia Kazachkova. 2022. Review of Sean Wallis. 2021. Statistics in Corpus Linguistics: A New Approach. New York/Oxon, Routledge. *Russian Journal of Linguistics* 26 (2). 550–557. https://doi.org/10.22363/2687-0088-30209

Рецензия

Рецензия на книгу Sean Wallis. 2021. *Statistics in Corpus Linguistics: A New Approach*. New York/Oxon, Routledge. ISBN 9781138589384 ISBN 9780429491696 (eBook)

И.В. ПРИВАЛОВА[1] ✉, **М.Б. КАЗАЧКОВА**[2]

[1]*Казанский федеральный университет, Казань, Россия*
[2]*Московский государственный институт международных отношений (университет), Москва, Россия*
✉*an*gladkova@gmail.com

Для цитирования:
Privalova I., Kazachkova M. Review of Sean Wallis. 2021. Statistics in Corpus Linguistics: A New Approach. New York/Oxon, Routledge. *Russian Journal of Linguistics.* 2022. Vol. 26. № 2. P. 550–557. https://doi.org/10.22363/2687-0088-30209

In present-day scholarship, there exist various approaches to the study of the material included in language corpora. The majority of the corpora studies are focused on practical issues of foreign language acquisition and provide samples of classroom activities that engage corpus material (Friginal 2018, Perez-Paredes 2020, Timmis 2015). There are also works devoted to particular linguistic issues,

for instance, the use of corpus linguistics in grammar analysis (Jones & Waller 2015), stylistics, writing and thought presentation (Semino & Short 2011) or special aspects of vocabulary (Szudarski 2017). The study of corpus linguistic data within the frame of non-linguistic disciplines is a rare phenomenon. One of the most successful attempts is an in-depth analysis of corpora use in sociolinguistics. This has become possible due to easier access to data in different languages through websites, social networking sites and blogs (Friginal & Hardy 2014).

The statistical approach to the study of linguistic corpora is of utmost importance since statistical methods are essentially helpful in assessing the representativeness of the language material and the reliability of the obtained results. In addition, statistical methods have proved to be quite effective in building statistical models and making forecasts. "Corpus linguistics is a powerful quantitative methodology, which heavily relies on frequency data and statistical procedures. It is difficult to talk about corpus linguistics without mentioning statistic measures based on frequency and distribution" (Cunxin 2020: 379). No wonder that we have recently witnessed the publication of a number of monographs debating the possibilities of statistical analysis. A bird's-eye view on basic statistics for corpus linguistics including both inscriptive and inferential techniques was presented almost two decades ago by Oakes in his book "Statistics for Corpus Linguistics" (Oakes 1998). It is also worth mentioning the edition "Quantitative Corpus Linguistics with R. A Practical Introduction" (Gries 2016), which demonstrates how to process corpus-linguistic data with the open-source programming language and environment R. Along with data and text processing, the author dwells upon basic aspects of statistical analysis and visualization. Vaclav Brezina's monograph "Statistics in Corpus Linguistics*:* A Practical Guide" (Brezina 2018) was the first attempt not only to provide the theoretical underpinning of statistical analysis but also to consider some practical examples of statistical techniques application. This volume contains some self-study material on a special companion website with exercises, keys and datasets. Cunxin Han, who carried out the analysis of all the three above-mentioned monographs on statistics in linguistic corpora, writes: "Brezina's book is a timely contribution because hands-on books are rarely found in the literature. The first book dedicated to this field "Statistics for Corpus linguistics" is already 20 years old while a more recent book "Corpus linguistics and statistics with R", caters more to the needs of advanced researchers" (Cunxin 2020: 379).

Amidst the body of similar studies, the monograph under review "Statistics in Corpus Linguistics: A New Approach" (2021) by Sean Wallis stands out. This volume is in a league of its own since it unites the achievements of statistics and corpus linguistics. The title of the book declares the *new approach*, and it is indeed presented in this edition. In the preface, the author poses the question: "Why do we need another book on statistics?" which he answers in the following way: "This book arose from the realisation that conventional approaches to the teaching and discussion of statistics – in the field of linguistics at least – are not working" (VIII).

It is worth saying a few words about the author of this book and about the history of its creation. Professor Sean Wallis is the principal research fellow at the Department of English Language and Literature at the University College London and the Head of the first Corpus Linguistics research centre established in Europe. Professor Sean Wallis has been preoccupied with compiling corpora for more than 25 years and he is one of the leading experts in building corpora whose work has international acclaim. Interestingly enough, this monograph (as the author reveals) was written in stages over ten years. Some early versions of two chapters appeared more than ten years ago and were afterwards refined; later new chapters were added. As such, this book is a multi-year and well thought-out research.

The narration of the book "Statistics in Corpus Linguistics: A New Approach" is rather well-elaborated as the book consists of six parts, nineteen chapters, preface, glossary, reference list, and index. Each chapter in its turn has sub-chapters and paragraphs. The following practical issues are put in the spotlight: the constraint of a research question, the correct employment of confidence intervals, the selection of optimal significance tests, the effect of variables on each other, the estimation of distribution patterns similarity, and comparison of the results of two experiments (2). Some new issues are debated in this book, specifically, the role of ideal binomial and normal distributions in various tests as well as the possibilities of the Wilson interval compared to the Gaussian (normal) interval. The requirements to the linguistic experiment that produce valid language data are also described. In the section about the needs of linguists in statistics Wallis points out that the main problem is the correct choice of the test – whether it should be a chi-square or log-likelihood. Also, it is stated that the relevance of the obtained data matters as well as the effect of one variable on another.

A remarkable feature of the book is its discursive style; the author seems to be in a dialogue with his readers. For example, in the Preface the titles of paragraphs are formulated as questions: "Why do we need another book on statistics?", "Why is statistics difficult?", "What do linguists need to know about statistics?" (XIII–XXII). However, the introduction that looks like a conversation should not create an illusion of an easy read, because the author poses some research questions that cannot be tackled without any basic knowledge in statistics: how to employ confidence intervals to correctly measure the size of the effect of one variable on another, how to estimate the similarity of distribution patterns; how to evaluate whether the results of two experiments significantly differ, and so on. A big number of new special terms might scare off amateurs in statistics. To prevent this from happening, the author provides a Glossary (pp. 329–341) wherein the explanation of the most important concepts can be found. There are definitions of some universal concepts such as "algorithm – a complete description of a computational process" (p. 329) or "an axiom – a rule or principle of a mathematical system" (p. 330) or "research question – a general question motivating a piece of research that may require refinement in order to be translated into testable hypothesis" (p. 366). In addition, there are many terms related to statistics, for instance:

"Student's t test – a comparable test to the $_z$ test, properly for Real data (usually termed on a Ratio scale) but sometimes used for Interval data" (p. 340).

Part 1 "Motivations" discusses general considerations on motivations, along with the issues of experimental design and data collection in corpus linguistics. Wallis clarifies the principles of material collection for corpora building. At the very beginning, the author emphasizes that the number of words does not necessarily lead to information quality. The size of the corpus should be in agreement with its annotation richness. In this section, the readers should pay attention to three important issues. First, a clear demarcation between such notions as "corpus" and "dataset": "In the most general sense, corpora are simply collections of language data processed to make them accessible for research purposes. In contrast to experimental datasets sampled to answer a specific research question, corpora are sampled in a manner that – as far as possible – permits many different types of research question to be posed. Datasets extracted from corpora are not obtained under controlled conditions but under 'naturalistic' or 'ecological' ones" (p. 4).

Second, the author distinguishes corpora that include written texts and corpora of spoken data. A major part of the corpora is drawn from written sources, "…a corpus of spoken data, ideally in the form of recordings aligned with orthographic transcriptions" (p. 5). When data come from real world sources rather than controlled laboratory conditions, they are more informative and relevant. Speech corpora represent a pioneering approach to make the recordings of discourses, dialogues and responses. It is an approach to register the multimodal essence of the language with sound evidences and prosodies (Svenja & Carter 2013). Linguistic data obtained from psycholinguistic experiments can be included in language corpora; however, these experiments must be carried out with little (or no) guidance from the experimenter. There are essentially three distinct classes of empirical evidence that may be obtained from any linguistic data source, and they are: factual evidence, frequency evidence and interaction evidence (p. 6). Wallis argues that the same statistical methods can be applied to different types of corpora in linguistics, including text corpora, speech corpora, lexical corpora and inter-lingual corpora. He considers monolingual, bilingual and even trilingual corpora, which are widely used as electronic dictionaries, as examples for statistical analysis. The necessity to use statistical methods is also determined by the fact that modern linguistic corpora are a repository of big data. The successful development of corpus linguistics goes hand in hand with the development of new computer technologies. Digitalisation of the language provides for the centralisation of information in many forms, one of which is language corpora. Therefore, linguistic corpora can be viewed as examples of a structured representation of big data, the validity of which can be proved by the application of statistical techniques.

Part 2 "Designing Experiment with Corpora" presents an analysis of the statistical methods that may be applied in corpus experiments. Since many of the studies in corpus linguistics come from literary disciplines, the researchers may be

unfamiliar with mathematics and statistics methodologies. The author asserts that experimental methods in corpus linguistics have their specifics and require a creative approach. Corpus linguists should be careful about controlled 'laboratory' experiments that use stimuli, 'cues', or artificial conditions to encourage particular behaviours: "The dominant trend in corpus linguistics is to build ever-larger 'flat' tagged corpora and employ greater reliance on computation" (29). This part analyses such techniques as obtaining data, extracting data, visualising proportions and applying testing (the Chi-Square test). An example of a linguistic interaction experiment is considered on pp. 40–42. In the chapter "The Vexed Problem of Choice", the author underlines the axiomatic character of models in sociolinguistics and cognitive linguistics research and the non-axiomatic one in corpus linguistics. It would be an exaggeration to say that experimental models in corpus linguistics are limited to mere 'counting surface phenomena.' Linguistic choice corpus research requires the inference of the counterfactual. Alongside what participants wrote or said, the researcher needs to infer what they could have written or spoken instead. To take a simple example, consider the study of *that*-omission in relative causes, as in *The man [that] I saw*. We must be able to reliably identify 'zero-relative' clauses (p. 51). As for exposure rates, the author affirms that corpora are exceptional resources for estimating overall likelihood of readers or hearers encountering a form. There are different formulas that present exposure rate (p(x I word) = f(x) / f(words) and choice rate (p(x I X = / f(X) (51).

Before the readers pass on to Parts 3, 4 and 5, we would strongly recommend them to take a closer look at the definitions in the Glossary. Special statistical terms are in abundance in Parts 3, 4 and 5, and it is challenging to read the text without the knowledge of such terms as: 'distribution', 'interval', 'probability', 'variable', 'baseline', 'contingency', 'uncertainty', etc. In Part 3 "Confidence Intervals and Significance Tests", Wallis makes an introduction into inferential statistics and contemplates on the 'naïve knowledge' that people get through experience. The same is true about the knowledge in statistics which the author calls 'naïve statistics'. He gives the definition of a probability and distinguishes three types: the observed probability (or proportion), the 'true' population probability and the probability that an observation is unreliable (p. 98). Also, the phenomena of binominal, normal and skewed distributions are presented in the possible applications towards texts. Optimum methods of calculation of linguistic data are presented on p. 212.

In Part 4 "Effect Sizes and Meta-Tests", Wallis compares different versions of the same experiment in order to see how a change to the experimental design may influence the results. He also examines the effect of an experimental design upgrading on reported results. The author introduces such notions as 'point test' and 'multi-point test' for contrasting the distribution of linguistic data across a dependent variable in homogeneity tables, as well as 'gradient test' methods for comparing sizes of effect in homogeneity tables, commencing with intervals and tests with a single degree of freedom. He also describes the application of the test

for Cramer's effect sizes larger table. One more novelty is the consideration that linguistic variables may often be measured as Binomial proportions expressing the probability that in a random case drawn from a sample the people might find a particular linguistic phenomenon (p. 231).

Part 5, "Statistical Solutions for Corpus Samples", addresses particular problems, such as conducting research with imperfect data and adjusting intervals for random-text samples. This chapter considers situations and statistical principles that can be used in their relation to linguistic matter. The author considers variations of observed proportions between text subsamples utilizing two different models: one that analyses each text as a random sample, and another that examines the distribution of actual subsample scores. The discussion of the possibilities of the new method brings the author the conclusion that the 'Binomial' per-text distribution is really the sum of multiple Binomial distributions, one for each sample size (pp. 278–279).

Part 6 "Concluding Remarks" shows how to calculate distribution curves, for instance, how to plot the Wilson distribution and Clopper-Pearson distribution. The Wilson score interval is a member of a class of confidence intervals characterising expected variation about an observed Binomial proportion. The author concludes that the entire point of a statistical method is to understand the implications of their data. "Research concerns the structure of the physical world: in linguistics, the structure of language" (p. 314). The chapter ends with a brief summary of the author's ideas.

In conclusion, it is necessary to note that the availability of corpora and the technological advancements of corpus tools have recently increased dramatically. At present, digitalisation of the language, globalisation of information processes and complication of text models require new methods for studying linguistic material. In order to show groundbreaking results, scientists have to go beyond the borders of one linguistic discipline, and the monograph "Statistics in Corpus Linguistics: A New Approach" by Sean Wallis shows how this can be done. It demonstrates a new non-trivial approach towards statistical methods and the way to apply it when studying quantitative and qualitative linguistic variables, such as a specific lexical unit in various syntactic constructions, the distribution of word length, the distribution of sentence length, and the distribution of specific vocabulary in the text. No doubt, this book will not go unnoticed by "pure" linguists since the author claims it is "…accessibly written to those with little to no statistical background" (i). Indeed, this book is written in such a way that a set of specific knowledge and terms would not scare away a beginner. Moreover, researchers of a "new generation" working in the field of corpus linguistics cannot do without any basic information in the field of programming and statistics. Thus, the advantages of the Chi-Square testing or t-Student's testing in their application towards corpus linguistics are obvious. By and large, the monograph "Statistics in Corpus Linguistics: A New Approach" by Sean Wallis is not a book for one-time reading, rather it is a guide that scientists will refer to on a regular basis when exploring some kind of linguistic material presented in corpora.

Acknowledgements

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

REFERENCES

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899

Cunxin, Han. 2020. Statistics in Corpus Linguistics: A Practical Guide. *Journal of Quantitative Linguistics* 27(4). 379–383. https://doi.org/10.1080/09296174.2019.1646069

Friginal, Eric & Jack Hardy. 2014. *Corpus-Based Sociolinguistics. A Guide for Students*. N.Y.: Routledge. https://doi.org/10.21283/2376905X.2.32

Friginal, Eric. 2018. *Corpus Linguistics for English. Teachers. Tools, Online Resources, and Classroom Activities*. N.Y.: Routledge. https://doi.org/10.4324/9781315649054

Gries, Stefan Th. 2016. *Quantitative Corpus Linguistics with R. A Practical Introduction*. N.Y.: Routledge. https://doi.org/10.4324/9781315746210

Jones, Christian & Daniel Waller. 2015. *Corpus Linguistics for Grammar. A Guide for Research*. N.Y.: Routledge. https://doi.org/10.4324/9781315713779

Oakes, Michael. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University press. ISBN 0-7486-103204

Perez-Paredes, Pascual. 2020. *Corpus Linguistics for Education. A Guide for Research*. N.Y.: Routledge. DOI: 10.4324/9780429243615

Semino, Elena & Mick Short. 2011. *Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writin*g. N.Y.: Routledge. https://doi.org/10.4324/9780203494073

Szudarski, Pawel. 2017. *Corpus Linguistics for Vocabulary. A Guide for Research*. N.Y.: Routledge. https://doi.org/10.4324/9781315107769

Svenja, Adolphs & Ronald Carter. 2013. *Spoken Corpus Linguistics. From Mono-modal to Multimodal.* N.Y.: Routledge. https://doi.org/10.4324/9780203526149

Timmis, Ivor. 2015. *Corpus Linguistics for ELT. Research and Practice*. N.Y.: Routledge. https://doi.org/10.4324/9781315715537

Wallis, Sean. 2021. *Statistics in Corpus Linguistics: A New Approach*. N.Y./Oxon: Routledge. https://doi.org/10.4324/9780429491696

Book review history:
Received: 20 December 2021
Accepted: 25 February 2022

Bionotes:

Irina V. PRIVALOVA is Doctor Habil. of Philology, a Leading Research Felllow of the Research Laboratory "Text Analytics" at the Institute of Philology and Intercultural Communication of Kazan (Volga Region) Federal University. Her research interests embrace psycholinguistics, intercultural and mass communication, as well as corpus linguistics.

Contact information:
Kazan (Volga Region) Federal University
building 33, 2 Tatarstan street, Kazan, 420021, Russia
e-mail: ivprivalova@mail.ru
ORCID: 0000-0002-7740-2185

Maria B. KAZACHKOVA is an Associate Professor at Moscow State Institute of International Relations (University). Her research interests include corpus linguistics, text analytics, and discourse studies.
Contact information:
Moscow State Institute of International Relations (University)
3 Novo-Sportivnaya st., Odintsovo, 143071, Russia
e-mail: mbkazachkova@yandex.ru
ORCID: 0000-0002-0357-3010

Сведения об авторах:
Ирина Владимировна ПРИВАЛОВА – доктор филологических наук, ведущий научный сотрудник НИЛ «Текстовая аналитика» Института филологии и межкультурной коммуникации Казанского федерального университета. Ее исследовательские интересы включают психолингвистику, межкультурную и массовую коммуникацию, корпусную лингвистику.
Контактная информация:
Казанский федеральный университет
420021, Казань, ул. Татарстан 2, здание № 33, комната № 46
e-mail: ivprivalova@mail.ru
ORCID: 0000-0002-7740-2185

Мария Борисовна КАЗАЧКОВА – кандидат филологических наук, доцент Московского государственного института международных отношений (университета). Ее научные интересы включают корпусную лингвистику, аналитику текста и исследование дискурса.
Контактная информация:
Московский государственный институт международных отношений (университет)
Россия, 143071, Одинцово, ул. Ново-Спортивная, д. 3
e-mail: mbkazachkova@yandex.ru
ORCID: 0000-0002-0357-3010

Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 558–564
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30307

Book review

Review of A.Ya. Shajkevich, V.M. Andryushchenko, N.A. Rebeckaya. 2021. *Distributive-statistical analysis of the language of Russian prose of the 1850–1870s,* vol. 3. Publishing House YaSK, Moscow. ISBN 978-5-907290-61-7

Venera R. BAYRASHEVA

Kazan (Volga Region) Federal University, Kazan, Russia
✉vbayrasheva@gmail.com

For citation:
Bayrasheva, Venera R. 2022. Review of A.Ya. Shajkevich, V.M. Andryushchenko, N.A. Rebeckaya. 2021. *Distributive-statistical analysis of the language of Russian prose of the 1850–1870s,* vol. 3. Publishing House YaSK, Moscow. *Russian Journal of Linguistics* 26 (2). 558–564. https://doi.org/10.22363/2687-0088-30307

Рецензия

Рецензия на книгу: А.Я. Шайкевич, В.М. Андрющенко, Н.А. Ребецкая. 2021. *Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг.* Т. 3. М.: Издательский Дом ЯСК. ISBN 978-5-907290-61-7

В.Р. БАЙРАШЕВА

Казанский (Приволжский) федеральный университет, Казань, Россия
✉vbayrasheva@gmail.com

Для цитирования:
Байрашева В.Р. Рецензия на книгу А.Я. Шайкевич, В.М. Андрющенко, Н.А. Ребецкая. 2021. *Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг.* Т. 3. М.: Издательский Дом ЯСК. *Russian Journal of Linguistics.* 2022. Т. 26. № 2. С. 558–564. https://doi.org/10.22363/2687-0088-30307

Третий том завершает публикацию фундаментального исследования «Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг.», выполненного в Институте русского языка РАН коллективом под руководством авторов издания. Статистические методы систематически применяются в

лингвистических исследованиях с середины прошлого века, и их роль продолжает возрастать. Это объясняется, прежде всего, возможностью быстрой автоматической обработки огромных массивов текстов, появившейся благодаря необычайному прогрессу вычислительной техники, в том числе созданию Интернета. Лингвисты практически сразу после появления компьютеров осознали их возможности и приступили к созданию электронных корпусов текстов. В настоящее время корпусы размером 1 млрд слов и более уже не редкость. Созданный недавно Генеральный интернет-корпус русского языка (ENA, May 22, 2022)[1] содержит 20 млрд слов.

Для анализа корпуса текстов основным инструментом является дистрибутивно-статистический анализ (ДСА), которому авторы дают следующее широкое определение: «Дистрибутивно-статистический анализ есть сумма формальных алгоритмических процедур, направленных на описание языка и опирающихся только на распределение (дистрибуцию) заданных элементов в тексте». Цель всего исследования двоякая – описать и развить ДСА, а также изложить полученные с его помощью конкретные научные результаты, представляющие интерес для русистики. Исследование проводится на материале специально собранного корпуса русской прозы объемом 15 млн слов.

Основным объектом представленной рецензии является вышедший в 2021 г. последний том трехтомного труда, однако методология исследований, важные технические аспекты описаны в первых двух томах, поэтому по необходимости будет затронуто и содержание предыдущих томов. Авторы начинают исследование с противопоставления дескриптивной и теоретической лингвистики. В середине прошлого века проявилось разочарование в дескриптивной методике, явно выраженное в высказывании Н. Хомского: «было бы абсурдным пытаться построить грамматику, которая непосредственно описывала бы наблюдаемое лингвистическое поведение» (Хомский 1962). Однако, несмотря на всю продуктивность развитых Н. Хомским и его последователями абстрактных моделей грамматики, дескриптивный подход нельзя считать отвергнутым. Более того, в последнее время дескриптивный подход получил мощный импульс благодаря развитию компьютерных технологий. Одним из инструментов дескриптивного метода является ДСА, представленный в этой книге.

Авторы развивают интервальный подход в рамках ДСА, состоящий в том, что текст определенным образом разбивается на сегменты, а распределение языковых элементов анализируется в пределах сегментов. В первом томе такими сегментами являются слова (рассматривается распределение букв), во втором томе – биграммы, в третьем томе – отрезки текста длиной 40 слов. На каждом уровне анализа авторы обнаружили интересные закономерности.

[1] http://www.webcorpora.ru/

Первый том посвящен трем вопросам: описанию истории ДСА, результатам ДСА на микроинтервалах (словах) и частотному словарю языка русской прозы 1850–1870-х гг.

История применения статистических методов в лингвистике возводится к работам Т. Менденхолл конца XIX в. Приведенный в первом томе подробный анализ эволюции ДСА представляет несомненный интерес для истории лингвистических учений. Обсуждаются как внутренние мотивы эволюции ДСА, так и внешнее влияние. Основное внимание уделено подходам, развитым в отечественной лингвистике. Подробно изложен статистико-комбинаторный метод Н.Д. Андреева, работы В.В. Шеворошкина и Б.В. Сухотина.

Уже в первом томе авторы вводят математический аппарат, используемый в последующих томах. Для оценки статистической значимости отклонений от математического ожидания события используется следующая формула: $S = (f - m - 1)/ \sqrt{m}$, где f – наблюдаемая частота данного события, m — математическое ожидание этого события, подсчитанное на основе какой-то нулевой гипотезы (пуассоновского распределения). В ДСА событием является совместная встречаемость двух языковых элементов в пределах рассматриваемого интервала. Чем величина S больше, тем совместная встречаемость больше отклоняется от случайной, т.е. можно говорить об их сущностной связи. Экспериментально авторы установили, что при $S > 4$ можно говорить о неслучайности совместной встречаемости пары языковых элементов. Применяя ДСА на уровне слов, авторы выделили 43 агрегированных парадигмы с соответствующими наборами суффиксов. Также в первом томе полностью приведен частотный словарь русской прозы 1850–1870-х гг., содержащий 51 тыс. слов.

Во втором томе завершается описание сочетаемости элементов на уровне слов и приводится полный анализ на уровне биграмм. При анализе на уровне биграмм и бóльших интервалов величина m для пары слов a, b из вышеприведенной формулы принимается равной $m_{ab} = (F_a \times F_b) / N$, где F – частота слова в корпусе, N – объем корпуса (в словоупотреблениях). В используемом для исследования корпусе встретилось почти 6 млн биграмм, изучаются 235 тыс. наиболее частотных.

Авторы показывают, как ДСА биграмм приводит к выявлению элементов грамматики, таких как число, падеж, род, некоторых дистрибутивных классов – предлогов, адвербов, компаративов и др., а также указания на возможность некоторой коррекции классов. Например, формы на -о (*хорошо*) целиком вошли в парадигму адъективов (*хороший*).

Отметим, что авторы не ограничиваются только русским языком и только одним применением ДСА – интервальным анализом корпусов. Рассматривая хорошо известный Брауновский корпус английского языка, они рассчитывают по формулам взаимную близость между 15 его подкорпусами и обнаруживают два кластера подкорпусов – деловой (включающий также научные тексты) и подкорпус художественной литературы.

По степени детальности и полноты описания морфологии русского языка данная монография вполне может быть поставлена в один ряд с Грамматическим словарем русского языка А.А. Зализняка. Причем важно, что описание вытекает из формального квантитативного анализа корпуса текстов.

Третий том посвящен ДСА на средних интервалах и, пожалуй, является наиболее интересным. Прежде чем излагать полученные результаты, авторы предпринимают серьезный методологический анализ. Подробно описываются проблемы, возникшие при лемматизации, и выбранные способы их решения. Как и во многих задачах обработки текстов, проблему представляют омонимы. Хотя потенциально разрешение омонимии возможно на основе дистрибуции, оно представляет большие трудности, и в данной работе использован экспертный семантический анализ.

Принципиально важной проблемой стал выбор длины интервала, сделать который можно было только эмпирически. Авторы рассматривают три длины интервала: в 40 слов, в 200 слов и в 1000 слов, а затем проводят содержательный анализ результатов каждого из экспериментов. Лучшие результаты были получены при выборе интервала длиной 40, при котором совместная встречаемость слов в пределах интервалов отражает семантические (авторы используют термин ‘текстуальные’) связи между словами. Для больших интервалов (1000 слов) обнаруживаются сюжетные связи анализируемых произведений. Например, на материале произведений И.С. Тургенева выявляются связи *резать – лягушка – нигилист*, не характерные для русского языка в целом. Таким образом, ДСА на больших интервалах может быть полезным для литературоведческого анализа. Это, однако, выходит за рамки рассматриваемой работы.

В электронной версии монографии приведены все пары слов, для которых $S > 3$. Разумеется, на выбор величины S влияет размер обрабатываемого корпуса. Выбор $S > 3$ релевантен для корпуса размера около 15 млн слов. При этом для конкретных групп слов при выявлении связей в монографии рассматриваются и другие пороговые значения. Интересны следующие приводимые авторами иллюстрации. В множестве слов {*год весна лето осень зима*} наибольшая сила связи ($S = 76$) обнаруживается у пары *лето – зима*, что, вероятно, отражает антонимический характер семантических отношений между ними. А далее идут пары соседних времен года со значительно меньшей силой связи. Например, для пары *осень – зима* $S = 51$. Что же касается пары *весна – осень,* то для нее $S = 15$, т.е. связь менее сильная. Этот простой пример демонстрирует возможности ДСА; результат, который не может быть получен иными методами. Авторы также рассматривают ряд других примеров: месяцы, дни недели, числительные и проч.

Однако более интересны результаты применения ДСА ко всем словам используемого корпуса. В монографии приводится перечень из 3000 пар слов с аномально высоким значением S. Это перечень трактуется как шаг на пути к созданию частотного словаря фразеологических единиц. Первыми

в алфавитном порядке указаны следующие пары: *ааронов жезл, авторское самолюбие, ад вымощенный*. Существующие фразеологические словари русского языка создаются вручную, и представляется интересным сопоставить их со словарем, созданным компьютерным способом.

Для многозначных слов (омонимия и полисемия) указывается, что полностью проанализировать и компактно представить все их устойчивые связи представляется затруднительным, и приводится лишь некоторое число примеров. Например, выявлены следующие устойчивые текстуальные связи со словом *медведица*: 1) *медвежонок* (45[2]), *медведь* (18), *объятие* (9); 2) *Орион* (101), *Южный крест* (79), *звезда* (19), *небо* (6). Формальный ДСА без учета семантики не всегда дает хорошие результаты. Так у слова *мир* выявлено лишь одно значение – ‘peace’. Другие значения не показывают значимых ($S > 3$) сочетаний в рассматриваемом корпусе текстов. В других корпусах ситуация может быть иной, но этот пример демонстрирует определенные ограничения представленного в монографии варианта ДСА. Разумеется, чем больше корпус текстов, к которому применяется ДСА, тем полнее и точнее будут результаты.

Множество слов и их текстуальные связи представляют собой сеть, содержащую около 26 тыс. слов и 500 000 связей между ними. Между двумя словами (лемматизированными) из корпуса устанавливается связь, если для них $S > 3$. Причем с этой связью ассоциируется число S. Сеть напоминает популярные в последнее время карты слов (см., например, ENA, May 22, 2022)[3]. Однако карты слов обычно строятся на основе ассоциаций, описанная же в монографии сеть текстуальных связей построена по корпусу ДС методом. Было бы интересно провести специальное исследование по их сопоставлению.

В сети выделяются кластеры слов – группы слов, внутригрупповые связи которых преобладают над внешними. Описан метод выделения кластеров. Сначала ручным способом выделяется небольшая группа слов (центр кластера) с большими коэффициентами связи S. Затем она пополняется словами, имеющими хотя бы две связи со словами центра. Затем еще раз пополняется словами, имеющими хотя бы одну связь со словами ранее построенного множества. В монографии применяется подход case study, анализируется состав нескольких построенных таким способом кластеров. Один из примеров – кластер «Дуэль». Оказалось, что в него входят глаголы, преимущественно совершенного вида, что объясняется краткосрочностью дуэлей. Представляется интересным довести формальный подход до конца, применив для выделения кластеров современные алгоритмы кластеризации. Однако с учетом огромного объема сети и ее неоднородностью применение алгоритмов кластеризации нетривиально и требует отдельного исследования.

[2] В скобках указано значение *S*.

[3] https://wordassociations.net/ru/

Авторы обращают особое внимание на проблемы, возникающие при неоднородности текстов. Например, во входящих в исследуемый корпус произведениях А.И. Герцена аномально часто по сравнению с корпусом в целом встречаются слова *революция* и *республика*. Возможность существования подобных неоднородностей следует учитывать как на стадии формирования или выбора корпуса, так на стадии компьютерной обработки данных и на стадии анализа результатов.

Сопоставляя приведенный в монографии вариант ДСА с другими аналогичными методами, следует обратить внимание на следующее. По большому счету, этот метод сближается с латентным семантическим анализом (Landauer 1998), который также направлен на выявление контекстно-зависимых значений слов при помощи статистической обработки больших корпусов. Принципиальная разница состоит в том, что в методе ДСА, используемом в монографии, рассчитывается только близость слов, а в латентном семантическом анализе одновременно рассчитывается еще и близость фрагментов корпуса, при этом близость фрагментов является основной. Дистрибутивная близость слов является и предметом векторной семантики, классическая реализация которой представлена в (Mikolov 2013). С появлением в последние годы нейронных сетей глубокого обучения именно векторная семантика стала основным инструментом дистрибутивного анализа. При этом следует учесть сложность реализации таких задач, требующих серьезно квалификации в компьютерных технологиях. С этой точки зрения ДСА, предложенный в рассматриваемой монографии, выгодно отличается простотой реализации, что делает его доступным большому числу лингвистов.

В заключение отметим, что рецензируемая монография является прекрасным ответом на известное высказывание Н.С. Трубецкого «Язык лежит вне меры и числа» (Трубецкой 1960: 15). Несмотря на, казалось бы, сухой предмет монографии – статистику – книга написана живым языком, не характерным для современных научных публикаций, и, несомненно, привлечет внимание всех заинтересованных исследователей.

Благодарность

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ-2030).

Acknowledgements

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

Трубецкой Н.С. Основы фонологии. М., 1960. [Trubetskoi, Nikolai S. 1960. Osnovy fonologii. Moscow. (In Russ.)].

Хомский Н. Синтаксические структуры // Новое в лингвистике. 1962. Вып. 2. М. [Chomsky, Noam. 1962. Syntactic structures. *Novoe v Lingvistike*. 412–528. Moscow. (In Russ.)].

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Landauer, Thomas K., Peter W. Foltz & Darrell Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes* 25. 259–284.

Book review history:
Received: 20 December 2021
Accepted: 25 February 2022

Bionote:
Venera R. BAYRASHEVA holds a doctoral degree in physics and mathematics. She is Associate Professor of the Theoretical Cybernetics Department at Kazan (Volga Region) Federal University. Her research interests include theoretical cybernetics and computer linguistics.

Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
e-mail: vbayrasheva@gmail.com
ORCID: 0000-0002-1728-034X

Сведения об авторе:
Венера Рустамовна БАЙРАШЕВА – кандидат физико-математических наук, доцент кафедры теоретической кибернетики Казанского федерального университета. Сфера научных интересов: теоретическая кибернетика, компьютерная лингвистика.

Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, Казань, ул. Кремлевская, 18
e-mail: vbayrasheva@gmail.com
ORCID: 0000-0002-1728-034X