

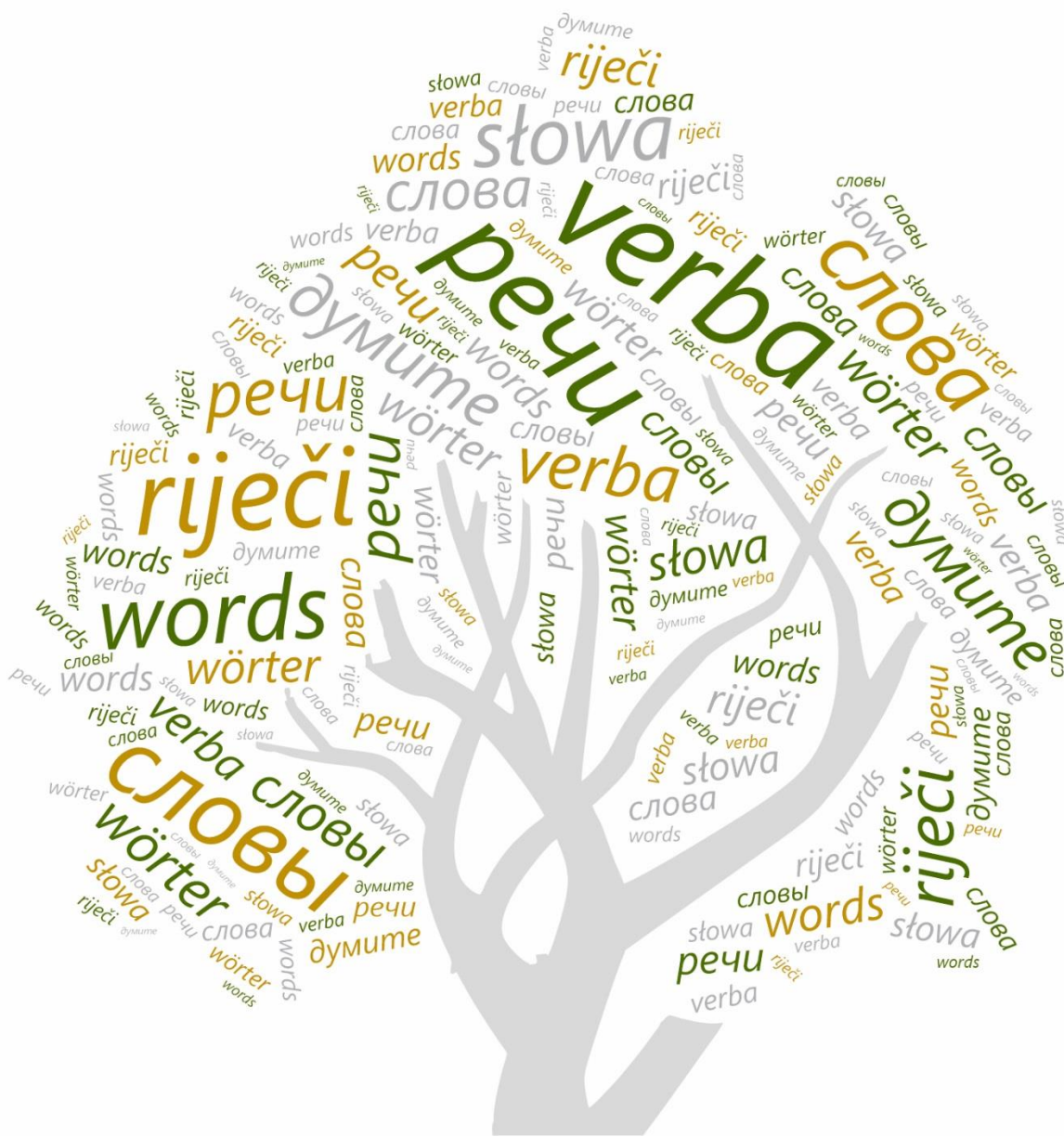


НОВГОРОДСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ
ИМЕНИ ЯРОСЛАВА МУДРОГО

Verba

Северо-Западный
лингвистический журнал

№ 3(13) 2024



ISSN 2713-0665 (Online)

Verba

Северо-Западный лингвистический журнал



(16+)

Verba. Северо-Западный лингвистический журнал

Сетевое периодическое научное издание

3(13) 2024

ISSN 2713-0665 (Online)

Выписка из реестра зарегистрированных СМИ

Эл № ФС77-80208 от 22.01.2021 г. Издание зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор)

Издается с 2021 г.

Периодичность: не менее 4 раз в год

УЧРЕДИТЕЛЬ И ИЗДАТЕЛЬ

ФГБОУ ВО «Новгородский государственный университет имени Ярослава Мудрого» (НовГУ)

АДРЕС УЧРЕДИТЕЛЯ И ИЗДАТЕЛЯ

173003, Россия, Великий Новгород,
ул. Большая Санкт-Петербургская, д. 41
тел.: +7 (8162) 62-72-44
e-mail: novsu@novsu.ru

АДРЕС РЕДАКЦИИ

173003, Россия, Великий Новгород,
ул. Большая Санкт-Петербургская, д. 41, ауд. 1216
тел.: +7(8162) 33-88-30 (доб. 2294)
E-mail: verba@novsu.ru

Сайт издания: <https://verba.press>

Редакторы перевода: О. Наволоцкая

Дизайн обложки: В. Фромов

Макет, верстка: Д. Ванюшкин

Разработка сайта: А. Ни

Дата выхода: 30.10.2024

© НовГУ, 2024

© Авторы статей, 2024

Все права защищены

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

Главный редактор:

Т.В. Шмелева, доктор филологических наук, профессор кафедры журналистики Новгородского государственного университета им. Ярослава Мудрого, Великий Новгород, Россия

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

В.И. Макаров, кандидат филологических наук, доцент; Новгородский государственный университет им. Ярослава Мудрого, Великий Новгород, Россия

А. Бирих, доктор филологических наук, профессор; Трирский университет, Трир, Германия

Х. Вальтер, доктор филологических наук, профессор; Университет им. Эрнста Морица Арндта г. Грайфсвальда, Грайфсвальд, Германия

В.Л. Васильев, доктор филологических наук, доцент; Новгородский государственный университет им. Ярослава Мудрого, Великий Новгород, Россия

В.И. Заика, доктор филологических наук, доцент; Новгородский государственный университет им. Ярослава Мудрого, Великий Новгород, Россия

В.И. Коваль, доктор филологических наук, профессор; Гомельский государственный университет им. Ф. Скорины, Гомель, Беларусь

К. Кусаль, доктор филологических наук, профессор; Гуманитарно-экономическая Академия, Лодзь, Польша

В.И. Мокиенко, доктор филологических наук, профессор; Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Т.Г. Никитина, доктор филологических наук, профессор, профессор; Псковский государственный университет, Псков, Россия

Б.Ю. Норман, доктор филологических наук, профессор; Белорусский государственный университет, Минск, Беларусь

М. Рак, доктор филологических наук, профессор; Ягеллонский университет, Краков, Польша

Ж. Финк, доктор филологических наук, профессор; Загребский университет, Загреб, Хорватия



Verba

Северо-Западный лингвистический журнал



(16+)

Verba. North-West Linguistic Journal

online journal

3(13) 2024

ISSN 2713-0665 (Online)

Extract from the register of registered mass media

El № FS 77-80208 of 22.01.2021. The edition is registered by the Federal Service for Supervision in the Sphere of Telecommunication, Information Technologies and Mass Communications (Roskomnadzor)

Founded: 2021

Frequency: at least 4 times a year

FOUNDER AND PUBLISHER

FSBEI HE "Yaroslav-the-Wise
Novgorod State University" (NovSU)

ADDRESS OF THE FOUNDER AND PUBLISHER

173003, Russia, Veliky Novgorod,
B. Sankt-Peterburgskaya St., 41
tel.: +7 (8162) 62-72-44
e-mail: novsu@novsu.ru

CORRESPONDING ADDRESS

173003, Russia, Veliky Novgorod,
Yaroslav-the-Wise Novgorod State University,
B. Sankt-Peterburgskaya St., 41, of. 1216
tel.: +7(8162) 33-88-30 (ext: 2294)
E-mail: verba@novsu.ru

Website of edition: <https://verba.press>

Translation editors: O. Navolotskaya

Cover design: V. Fromov

Layout: D. Vanyushkin

Website creation: A. Nee

Release date: 30.10.2024

© NovSU, 2024

© Authors of articles, 2024

All rights reserved

EDITORIAL BOARD

Editor-in-Chief:

T.V. Shmeleva, Doctor of Philology, Professor
Yaroslav-the-Wise Novgorod State University,
Veliky Novgorod, Russia

Members of Editorial Board

V.I. Makarov, Candidate of Philology, Associate
Professor of the Department of Philology,
Yaroslav-the-Wise Novgorod State University,
Veliky Novgorod, Russia

A. Bierich, Doctor of Philology, Professor
University of Trier, Trier, Germany

H. Walter, Doctor of Philology, Professor
Ernst-Moritz-Arnst University of Greifswald,
Greifswald, Germany

V.L. Vasiliev, Doctor of Philology, Associate Professor
Professor of the Department of Philology,
Yaroslav-the-Wise Novgorod State University,
Veliky Novgorod, Russia

V.I. Zaika, Doctor of Philology, Associate Professor
Professor of the Department of Philology,
Yaroslav-the-Wise Novgorod State University,
Veliky Novgorod, Russia

V.I. Koval, Doctor of Philology, Professor
Francisk Skorina Gomel State University, Gomel, Belarus

K. Kusal, Doctor of Philology, Professor
Humanitarian and Economic Academy in Lodz,
Lodz, Poland

V.I. Mokienko, Doctor of Philology, Professor
St. Petersburg State University, St. Petersburg, Russia

T.G. Nikitina, Doctor of Philology, Professor
Pskov State University, Pskov, Russia

B.Yu. Norman, Doctor of Philology, Professor
Belarusian State University, Minsk, Belarus

M. Rak, Doctor of Philology, Professor
Jagiellonian University, Krakow, Poland

Zh. Fink, Doctor of Philology, Professor
University of Zagreb, Zagreb, Croatia



Содержание

Content

От главного редактора.....4 From Editor-in-Chief.....4

Теоретическое осмысление новаций, проблем и перспектив / Theoretical Comprehension of Innovations, Challenges and Prospects

Компьютерные технологии в лингвистике Computer Technologies in Linguistics
В. А. Белов.....8 *V. A. Belov*.....8

Сравнительно-сопоставительный анализ лингвистических ресурсов для проведения корпусного анализа текстов Comparative-Contrastive Analysis of Linguistic Resources for Corpus Analysis of Texts
А. В. Дмитриев, Е. С. Крупнова.....24 *A. V. Dmitrijev, E. S. Krupnova*.....24

Язык поэзии сквозь призму базы данных / Poetry Language Through Database Lens

Преимущества использования баз данных в лингвистических исследованиях Advantages of Using Databases in Linguistic Research
А. А. Лебедев.....36 *A. A. Lebedev*.....36

Искусственный интеллект / Artificial Intelligence

Язык промптов, или особенности формулирования запросов к генеративным нейросетям для создания изображений Prompt Language, or Features of Queries to Generative Neural Networks for Image Creation
Л. Г. Алексева, П. С. Алексеев.....50 *L. G. Alexeeva, P. S. Alexeev*.....50

Человеческое и компьютерное в научном тексте Human and Computer in Scientific Text
А. С. Савельев.....62 *A. S. Savelyev*.....62

От главного редактора

V

Дорогие коллеги,
наши авторы и читатели!



Тема этого номера журнала – «Компьютерные технологии на службе у лингвистики» – собрала молодых авторов, имеющих собственный опыт использования компьютерных технологий в лингвистических исследованиях и опыт теоретического осмысления этой проблемы, фундаментальность которой для современной лингвистики не вызывает сомнений.

В первой рубрике объединены статьи коллег из Санкт-Петербургского Политехнического института Петра Великого – Вадима Алексеевича Белова «**Компьютерные технологии в лингвистике**» и совместный труд Александра Владиславовича Дмитриева и Елены Сергеевны Крупновой «**Сравнительно-сопоставительный анализ лингвистических ресурсов для проведения корпусного анализа текстов**». Хотя, судя по их названиям главные объекты анализа – это технологии, на самом деле авторов интересует, каков круг исследовательских задач, к решению которых привлекаются компьютерные технологии и специальные лингвистические ресурсы. Анализ публикаций – отечественных и зарубежных, представленный в этих статьях, убеждает в том, что лингвистикой уже достигнуты значительные достижения с помощью компьютерных технологий и открываются еще более многообещающие перспективы. Значимость содружества лингвистики и компьютерных технологий так велика, что говорят об особых лингвистиках – компьютерной и корпусной, с чем, конечно, можно поспорить, что не исключает убедительности всех приведенных фактов и соображений.

Подтверждает убедительность высказанных теоретических суждений вторая рубрика «**Язык поэзии сквозь призму базы данных**», где публикуется статья доцента Петрозаводского государственного университета Александра Александровича Лебедева «**Преимущества использования баз данных в лингвистических исследованиях**». Не отрицая классических методов изучения языка поэзии, отделившейся от нас тремя столетиями, автор демонстрирует результативность в ее исследовании технологии баз данных, приводя таблицы, статистические данные, новые аспекты текстов, которые открываются исследователю, вооруженному новейшими технологиями.

В третьей рубрике номера «**Искусственный интеллект**» представлены статьи молодых авторов номера – аспирантов и научных сотрудников Новгородского государственного университета. Еще недавно искусственный интеллект казался делом туманного будущего, а сегодня, как убеждают статьи этой рубрики, это уже

факты нашей повседневности – учебной и научной. Интересно, что обе статьи построены на экспериментах.

Любовь Геннадьевна и Петр Сергеевич Алексеевы (кстати, молодожены, с чем мы их поздравляем) организуют педагогический эксперимент, предлагая своим студентам сформулировать запрос нейросети, а затем анализируют результаты, формулируя в итоге принципы языка таких запросов. Их статья так и называется – **«Язык промптов, или особенности формулирования запросов к генеративным нейросетям для создания изображений»**, что убеждает: овладеть этим языком нам придется, видимо, довольно скоро, и в этом интерес к статье и ее практическая ценность.

Алексей Сергеевич Савельев в статье **«Человеческое и компьютерное в научном тексте»** предлагает результаты проведенной им серии экспериментов, изменяя реальные научные тексты и даже сочиняя абсурдные научные тексты. Все это для того, чтобы показать, как нейросеть «оценивает» научные тексты. Анализируя разные стороны современных и классических научных текстов, автор показывает, что нейросеть принимает за научность и как часто эти признаки оказываются далеки от реальной научной ценности и оригинальности текста. В плане оценочности эта статья вносит ноту трезвости и даже скепсиса в отношении к компьютерным технологиям и их надежности в оценке научных статей.

Таким образом, публикуемые в номере журнала статьи представляют проблему отношений лингвистики и компьютерных технологий и фронтально с большими надеждами, и конкретно с осмыслением сложностей и даже опасностей этого объединения.

Стоит сказать, что география этого номера соответствует названию журнала: все авторы работают на Северо-Западе, наряду с лингвистами из Великого Новгорода и Петрозаводска, в журнале публикуются петербуржцы. Стоит отметить, что это представители классических университетов встретились на страницах этого номера с сотрудниками знаменитого Политехнического университета Петра Великого, в составе которого есть Высшая школа лингвистики и педагогики.

Благодарю сердечно авторов номера и рецензентов – докторов филологических наук Владимира Ивановича Заику из Великого Новгорода, Белова Вадима Алексеевича из Петербурга.

До новых встреч на электронных страницах нашего журнала!

Т. В. Шмелева

Letter from the Editor-in-Chief

Dear colleagues, our authors and readers!

The theme of this issue of the journal – *Computer Technologies in the Service of Linguistics* – has brought together young authors who have their own experience of using computer technologies in linguistic research and experience of theoretical understanding of this problem, the fundamental nature of which for modern linguistics is beyond doubt.

In the first section, there are papers by colleagues from Peter the Great St. Petersburg Polytechnic University — *Computer Technologies in Linguistics* by Vadim A. Belov and the joint work of Alexander V. Dmitrijev and Elena S. Krupnova *Comparative-Contrastive Analysis of Linguistic Resources for Corpus Analysis of Texts*. Although, judging by the titles, the main objects of analysis are technologies, in fact, the authors are interested in the range of research problems, to the solution of which computer technologies and special linguistic resources are applied. The analysis of publications, both domestic and foreign, presented in these papers, convinces us that linguistics has already achieved significant results with the help of computer technologies and even more promising prospects are opening up. The importance of the concord of linguistics and computer technologies is so great that special linguistics fields are being discussed, including computer and corpus linguistics, which, of course, can be argued, but does not exclude the persuasiveness of all the facts and considerations presented.

The second section *Poetry Language Through Database Lens* confirms the persuasiveness of the theoretical judgments expressed, where the paper by Associate Professor of Petrozavodsk State University Aleksandr A. Lebedev *Advantages of Using Databases in Linguistic Research* is presented. Without denying the classical methods of studying the language of poetry separated from us by three centuries, the author studies database technology, adding tables, statistical data and new aspects of texts that are revealed to the researcher armed with the latest technologies.

The third section of the issue *Artificial Intelligence* features papers by our young authors — post-graduate students and research fellows of Yaroslav-the-Wise Novgorod State University. Until recently, artificial intelligence seemed to be a matter of the uncertain future, but today, as the papers convince us, it is already a fact of our everyday life — educational and scientific. It is interesting that both papers are based on experiments.

Liubov G. Alexeeva and Peter S. Alexeev (by the way, newly married couple, with which we congratulate them) have organized a pedagogical experiment, asking their students to formulate a request for a neural network, and analyze the results, ultimately formulating the principles of the language of such requests. Their paper is called *Prompt Language, or Features of Queries to Generative Neural Networks for Image Creation*, which convinces us: we are probably going to have to master this language quite soon, and this is the interest in the paper and its practical value.

Alexey S. Savelyev in his paper *Human and Computer in Scientific Text* offers the results of a series of experiments he has conducted, changing real scientific texts and even

writing absurd scientific texts. All this in order to show how a neural network “evaluates” scientific texts. Analyzing different aspects of modern and classical scientific texts, the author shows what a neural network accepts as scientific and how often these features are far from the real scientific value and originality of the text. In terms of evaluation, this paper brings a note of sobriety and even skepticism to the attitude towards computer technologies and their reliability in evaluating scientific articles.

Thus, the papers of this issue present the problem of the relationship between linguistics and computer technologies, both frontally with great hopes and specifically with an understanding of the complexities and even dangers of this unification.

It is worth mentioning that the geography of the issue corresponds to the title of the journal: all the authors work in the North-West, and, along with linguists from Veliky Novgorod and Petrozavodsk, the journal publishes researchers from St. Petersburg. It should be noted that representatives of the classical universities have met on the pages of the issue with employees of the famous Peter the Great St. Petersburg Polytechnic University, which includes the Higher School of Linguistics and Pedagogy.

I would like to express my heartfelt gratitude to the authors of the issue and reviewers — Doctors of Philological Sciences Vladimir I. Zaika from Veliky Novgorod, Vadim A. Belov from St. Petersburg.

Until we meet again on the electronic pages of our journal!

T. V. Shmeleva

ТЕОРЕТИЧЕСКОЕ ОСМЫСЛЕНИЕ НОВАЦИЙ, ПРОБЛЕМ И ПЕРСПЕКТИВ / THEORETICAL COMPREHENSION OF INNOVATIONS, CHALLENGES AND PROSPECTS

Компьютерные технологии в лингвистике

В. А. Белов

Computer Technologies in Linguistics

V. A. Belov

Вадим Алексеевич Белов – доктор филологических наук, доцент; Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

E-mail: belov.vadim.a@gmail.com

Статья поступила: 01.10.2024. Принята к печати: 20.10.2024.

В статье представлен анализ современных исследований в области компьютерной и корпусной лингвистики. Актуальность работы связана с тем, что данные сферы бурно развиваются, поэтому важно представить на русском языке анализ возможностей и достижений компьютерной лингвистики. В работе используются теоретические методы исследования. Статья состоит из двух разделов: в первой рассматриваются основные исследования в области корпусной лингвистики, во второй – кратко представлены достижения компьютерной лингвистики. Отмечается, что корпусные данные стали важным источником информации для лингвистических работ разной проблематики: они используются в работах по изучению лексической семантики, грамматики, дискурса, истории языка, идиостиля автора, а также для решения практических задач, связанных с переводом и обучением языку. В целом работы, выполненные с применением корпусных данных, можно отнести к функциональным. Они часто основываются на дистрибутивном (тезаурном) подходе к значению. Компьютерная лингвистика представляет широкую область исследования, находящуюся на стыке лингвистики, математики и информационных технологий. Достижения современной компьютерной лингвистики используются для решения практических задач (автоматическое порождение и восприятия текста, индексация и анализ информации). Для автоматизации речи используются формальные модели описания, предполагающие последовательный графематический (фонологический), морфологический, синтаксический, семантический и дискурсивный анализ. Современные языковые модели, которые чаще всего обучаются на специальных корпусах, также применяются для решения лингвистических задач.

Vadim A. Belov – Doctor of Philological Sciences, Associate Professor; Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russian Federation

ORCID: 0000-0002-4173-2000

Received: 01.10.2024. Accepted for publication: 20.10.2024.

The article presents an overview of modern research in the field of computational and corpus linguistics. The relevance of the work is due to the fact that these areas are rapidly developing, so it is important to present an overview in Russian of the possibilities and achievements of computational linguistics. The work uses theoretical research methods. The article consists of two sections. The first examines the main studies in the field of corpus linguistics, the second briefly presents the achievements of computational linguistics. It is noted that corpus data have become an important source of data for linguistic works on various issues. Corpus information is used in studies of lexical semantics, grammar, discourse, history of language, author's individual style, etc., as well as for solving practical problems related to translation and language teaching. In general, work carried out using corpus data can be classified as functional and is often based on a distributive (thesaurus) approach to meaning. Computational linguistics is a broad field of research located at the intersection of linguistics, mathematics and information technology. The achievements of modern computational linguistics are used in practical tasks (automatic generation and perception of text, indexing and analysis of information). For the automation of speech, formal models of description are used, which assume consistent graphematic (phonological), morphological, syntactic, semantic and discourse analysis. Modern language models, which are most often trained on special corpora, are also used to solve linguistic problems. This work is addressed to linguists, specialists in the field of information technology, as well as students of philological and information sciences.

Ключевые слова: компьютерная лингвистика, корпусная лингвистика, коллокация, языковая частотность, дистрибутивные модели, мера ассоциативности, языковые модели, анализ тональности

Keywords: computational linguistics, corpus linguistics, collocation, language frequency, distributional models, associativity measure, language models, sentiment analysis

УДК 81'322

OECD: 6.020Y

V

Постановка проблемы. В настоящее время информационные технологии переживают этап бурного развития: регулярно появляются новые технологии и практики в этой области. Так, появление нейронных сетей и языковых моделей, произошедшее в последние десять лет, расширило арсенал информационных средств. Безусловно, подобное интенсивное развитие затрагивает разные научные сферы, в том числе лингвистику. В настоящей работе ставится проблема – выявить круг направлений лингвистических исследований, в которых сложилась практика использования компьютерных технологий, на основе анализа отечественных и зарубежных публикаций по теме.

История вопроса. Лингвистическая наука во многом стояла у истоков современных IT-технологий. Так, когнитивный лингвист и психолог Джордж Миллер, который во многом заложил основы современного когнитивного подхода к языку, большое внимание уделял информационным технологиям; примером такого внимания является известная лексическая база данных «WordNet» [Miller, Beckwith, Fellbaum, 1988]. Похожий пример – научная биография Марвина Минского, который считается пионером в области искусственного интеллекта и нейронных сетей. Однако в лингвистике он прежде всего известен разработкой понятия фрейма, которое плотно вошло в современную когнитивную науку; сегодня фрейм один из важнейших инструментов обработки когнитивной и языковой информации.

Анализ направлений компьютерной лингвистики проводился неоднократно в отечественной науке; см. работы [Захаров, Богданова, 2020; Ляшевская, 2016; Рюкова, 2024; Чилингарян, 2021].

Актуальность настоящей работы заключается в следующем: во-первых, компьютерная лингвистика – многоаспектная область, которая включает разнообразные направления, поэтому аналитические исследования обычно охватывают лишь некоторые аспекты вопроса, тогда как цель настоящего опыта анализа – выявить полный круг направлений лингвистических направлений теоретического и прикладного характера; во-вторых, эта сфера развивается интенсивно, и некоторые предшествующие уже потеряли актуальность, рассмотрение проблемы нуждается в дополнительной информации.

Методология и методика исследования. Так как цель настоящей статьи – выявить направления современных лингвистических работ с применением компьютерных технологий во всей их полноте, следовательно, работа носит теоретический характер, она основывается прежде всего на теоретических методах исследования, включающих анализ, обобщение, дедукцию, а также на исследовательском опыте автора и рефлексии по поводу изучаемых проблем.

Теоретические положения иллюстрируются примерами, для работы с которыми используются методы компьютерной лингвистики.

Анализ материала. Как показывает анализ публикаций, можно говорить о двух ключевых направлениях лингвистики, использующих компьютерные технологии – компьютерной и корпусной лингвистике, в рамках каждой из них сложились конкретные направления, что и определяет порядок представления в настоящей работе.

Компьютерная лингвистика

Чаще всего компьютерной лингвистикой называют изучение языка с помощью различных компьютерных технологий. Теоретическое осмысление понятия затрудняет то, что компьютерная лингвистика оказывается сопряжена с активно развивающимися информационными технологиями, которые в большей степени ориентированы на практические, а не теоретические задачи. Например, в англоязычной традиции больше распространён термин Natural Language Processing, NLP (обработка естественного языка), связанный с автоматическим анализом языка. Конечным результатом анализа языка должны стать автоматическое порождение и восприятие речи. Ключевым этапом развития NLP стало появления языковых моделей, которые с помощью обучения на представленном материале (чаще всего корпусе) формирует вероятностные модели.

В настоящее время языковые модели используются для решения большого количество практических лингвистических задач: распознавание и порождение речи, грамматическая и синтаксическая разметка (анализ) текстов, поиск информации, исправление ошибок, машинный перевод и пр. Так, создателя НКРЯ отмечают, что внедрение нейросетевых языковых моделей позволило провести качественную грамматическую разметку текстов [Савчук, Архангельский, Бонч-Осмоловская, Дони́на, Кузнецова, Ляшевская, Орехов, Подрядчикова, 2024]. В повседневной жизни эти технологии позволяли создать чат-боты, голосовое управление устройствами, выявлять спам (нежелательную информацию) по содержанию, подбирать актуальную рекламу, исправлять языковые ошибки, генерировать тексты в различных жанрах и пр. [Jurafsky, Martin, 2024].

В целом работа автоматических речевых систем строится по традиционной схеме, где первыми уровня анализа являются графематический (фонологический), морфологический, синтаксический, семантический и дискурсивный анализ текста. Однако сами процедуры анализа часто построены по формальным моделям, что предполагает их автоматизацию. Например, морфологический анализ предполагает тегирование (индексирование) всех входящих слов по частям речи с их грамматическими характеристиками и установление леммы (начальной, словарной формы) для словоформ. Сложным случаем является омонимия, где система вынуждена давать несколько вариантов грамматического описания; преодоление омонимии происходит только на семантическом уровне анализа. Для формы *потом* будет представлено два варианта: первый вариант от леммы *потом* (наречие, неизменяемое), второй вариант от существительного *пот* (единственное число,

творительный падеж). Как правило, морфологический анализ текстов строится на основе грамматического словаря, хотя возможен бессловарный способ, предполагающий поиск возможных окончаний (завершающих аффиксов) слов. Также сложности вызывают различные имена и названия объектов, организаций, для которых создается отдельный словарь.

Наиболее сложными для автоматических анализаторов речи является работа со смыслом высказываний: эти проблемы решаются в рамках многоаспектного семантического анализа, а также анализа тональности. В компьютерной лингвистике так называется анализ эмоционально-оценочной информации, также используются термины *сентимент-анализ*, *анализ мнений*. Однако даже в этих направлениях получены значимые результаты.

Так, анализ тональности текста, необходимый для интерпретации отзывом, комментариев, новостей и других текстов, которые трудоемко анализировать вручную, показывает высокую эффективность. Теоретически существующие модели основываются на положениях лингвистики, где выделяется три типа оценки: положительная, отрицательная и нейтральная (без видимой оценки). Первые системы были построены на основе правил на основе размеченного с точки зрения оценки словаря: в тональном словаре содержался набор слов, которым приписывалась ручная оценка (например, -2 или -1 для отрицательной оценки, 0 – для нейтральной и +2 или +1 – для положительной). Подобные системы в целом показывают высокую эффективность, но они достаточно трудоемки. Примером такой системы является «Русентилекс»: этот словарь 12 тысяч слов и выражений, которые были проанализированы экспертами-лингвистами в рамках контекстного употребления.

В настоящее время используются модели, основанные на машинном обучении, которые в целом показывают хорошие результаты. Однако наиболее сложной задачей при работе с этими системами является подбор корпуса текстов, на основе которых можно обучить систему. Так, в работе [Рубцова, 2012] был использован корпус сообщений в социальной сети «Твиттер», где были представлены 400 тыс. позитивных сообщений и 300 тыс. негативных «твиттов». В работе [Романов, Васильева, 2017] для обучения использовались рецензии на сайте «Кинопоиск», где система ориентируется на оценки фильмов. В исследовании [Софронова, 2024] применялась сложная база для обучения системы, включающая выборку из корпуса, размещенного лингвистом, и результаты психолингвистического эксперимента с оценками носителей языка.

Таким образом, в рамках компьютерной лингвистики решается масса практических вопросов, связанных с теоретическими проблемами семантики, стилистики и другими.

Корпусная лингвистика

Наиболее распространёнными являются корпусные исследования, которые достаточно плотно вошли в лингвистическую науку. Корпусом называют коллекцию текстов в устном или письменном виде, которая обработана компьютерными

средствами. Современные корпуса прежде всего снабжены разметкой: лингвистической (информация о лексической, грамматической, просодической и пр. организации) и нелингвистической (сведения об авторе и тексте).

В зарубежной лингвистике выделяются два направления работ в этой области [Tognini-Bonelli, 2001]. Первое, именуемое *corpus-based linguistics*, представляет исследования, предполагающие использование корпуса для подтверждения некоторых гипотез и идей. Второе направление (*corpus-driven linguistics*) предполагает, что корпус становится источником гипотез о языке, представляя собою новую теорию, философия языка. Наверное, наиболее удачно прокомментировал второе направление российский ученый В. А. Плунгян: «В современной теоретической лингвистике корпус – это не только мощный инструмент исследования языка, но и новая идеология, ориентирующая исследователя на текст как главный объект теоретической рефлексии. Корпус в каком-то смысле вернул лингвистам их подлинный объект – тексты на естественном языке в максимально полном объеме» [Плунгян, 2008]. Отметим, что обозначенное разграничение часто критикуется: так, в авторитетном издании «*Corpus Linguistics: Method, Theory and Practice*» указывается: «All corpus linguistics can justly be described as corpus-based» [McEnery, Hardie, 2011, p. 6]. Действительно, корпусные методы не предполагают нового объекта исследования и обновления общего подхода к языковым явлениям, а они только кардинальным образом расширяют эмпирическую базу исследования, что, безусловно, стимулирует развитие функционального подхода в лингвистике.

Анализ литературы позволил определить конкретные направления корпусной лингвистики, сосредоточенных на исследованиях особых языковых явлений.

Сейчас корпусные данные используются в разнообразных исследованиях. Прежде всего они дают информацию о примерах употреблений единиц в контексте. В «докорпусную» эпоху сбор данных, на основе которых проводится любое лингвистическое исследование, занимал большое количество времени и сил: автор исследования был вынужден самостоятельно провести сбор и ручную обработку материала. В целом доступ к обширному языковому материалу стимулирует научные исследования по разным проблемам. При этом можно сказать, что корпус способствует усилению функционального направления в лингвистике, для которых характерно «объяснение языковой формы через ее функции» [Кибрик, Плунгян, 2002, с. 276] и опора на эмпирические данные, в том числе корпусные [Там же]. Приведем некоторые примеры исследований, основанных на корпусных данных.

Лексическая семантика

В лексической семантике с помощью корпуса можно проследить, какие значения (лексико-семантические варианты) слова реализуются в языке. Так, многозначное слово *чистый*, по сведениям Малого академического словаря [Евгеньева, 1999], имеет 15 значений; однако они представлены неравномерно в современной речи. Корпусные данные позволяют выделить наиболее актуальные для современного языка употребления. В современных текстах чаще всего реализуются такие значения: первое значение «Не загрязненный, не запачканный, не

имеющий грязи или пятен»; пример употребления в (1); третье (переносное) значение «Имеющий свободную, открытую, ничем не занятую поверхность», пример употребления в (2); седьмое переносное значение «Отличающийся хорошей отделкой; тщательный. Аккуратно и искусно выполненный», пример употребления в (3);

(1) *Но боюсь брать из шкафа **чистую** одежду* (НКРЯ¹: Н. Б. Черных, журнал «Волга», 2015);

(2) *Луна в вечернем **чистом** небе висела полная, видная сквозь ветви клён* (НКРЯ: М. А. Булгаков. Мастер и Маргарита, 1929–1940);

(3) *Но помните, что там надо трудиться, а вы с **чистым** почерком найдете работу* (НКРЯ: Ф. М. Решетников. Между людьми, 1864).

В подобных исследованиях корпусные сведения позволяют, во-первых, представить количественные данные о реализации того или иного значения; во-вторых, выявить случаи семантической деривации в современных текстах; в-третьих, определить функциональные особенности употребления единиц в контексте. Полученные таким образом результаты могут быть представлены в современных словарях.

Контекст употребления может быть использован для определения эмоционально-оценочного содержания единицы. В работе [Радбиль, 2024] производится исследование имплицитной оценочности глагола *случиться*, по данным поэтического корпуса НКРЯ. На основе анализа делается вывод, что этот глагол имеет негативно-отрицательную «семантическую ауру». «Семантической аурой», по мнению британских ученых Дж. Р. Фёрта и Дж. Синклера, является ассоциативно-смысловой фон слова (часто связанного с оценкой), который может не осознаваться носителями языка и фиксируется на основе контекстов употребления [Firth, 1957; Sinclair, 1991]. Как правило, информация об эмоционально-оценочном содержании недостаточно полно представлено в словарях, за исключением современных словарей, которые опирают на корпусные данные (см., например, Активный словарь русского языка)².

Дистрибутивные модели

Корпус является важным источником данных о сочетаемости единиц, что стимулирует развитие дистрибутивных моделей в семантике. Подобный подход трактует значение слова через его употребление и позволяет автоматизировать семантический анализ: значение слова представляется как совокупность его контекстных употреблений. Здесь можно вспомнить известную цитату Дж. Фёрса, основоположника Лондонской лингвистической школы, который возродил интерес

¹ Здесь и далее принято сокращение: НКРЯ – Национальный корпус русского языка (<https://ruscorpora.ru>).

² Активный словарь русского языка / В. Ю. Апресян, Ю. Д. Апресян, Е. Э. Бабаева, О. Ю. Богуславская, Я. М. Бухаров, И. В. Галактионова, М. Я. Гловинская, Б. Л. Иомдин, Т. В. Крылова, И. Б. Левонтина, А. А. Лопухина, А. В. Птенцова, А. В. Санников, Е. В. Урысон. Редакторы тома: В. Ю. Апресян, И. В. Галактионова, Б. Л. Иомдин. Под общим руководством академика РАН Ю. Д. Апресяна. – Т. 4, ч. 1. – Электрон. текстовые данные. – М.: МЦНМО, 2023. – 256 с.

лингвистики к контексту и коллокациям: «You shall know a word by the company it keeps» («Вы поймете слово по его окружению») [Firth, 1957, p. 11].

Таким образом, на основе данных о сочетаемости можно построить дистрибутивные модели и словари, которые моделируют семантические связи в лексиконе, представляя семантическое сходство единиц. Значение в рамках этой модели представляет собой сеть взаимосвязанных единиц (без определенного толкования). Подробнее о тезаурусных моделях значения [Белов, 2020].

Наиболее распространенной научной метафорой семантической связи является вектор – условное семантическое расстояние между единицами в многомерном пространстве [Landauer, Foltz, Laham, 1998; Burgess, Lund, 2000]. Этот показатель может рассчитываться на основе совместной встречаемости в корпусе и/или результатов ассоциативных связей. Когнитивным основанием подобного подхода является то, что семантические связи формируются на базе регулярно встречающегося языкового окружения [Landauer, Foltz, Laham, 1998]. Семантическая связанность в рамках дистрибутивных моделей может устанавливаться, во-первых, с помощью определения контекстов, где слова могут употребляться вместе; во-вторых, с помощью установления одинаковых контекстов, в которых единицы употребляются.

Так, слова *утконос*, *опоссум*, *сумчатые*, *млекопитающее* часто встречаются в одинаковых контекстах, поэтому их можно признать семантически близкими. Таким образом организована работа дистрибутивных моделей по установлению семантически близких слов: синонимами становятся единицы, которые способны употребляться в одинаковых контекстах (то есть оказываются взаимозаменяемы в контексте). По такому алгоритму работают известные системы «WordNet», «Latent Semantic Analysis» (скрытые семантический анализ) [Landauer, Foltz, Laham 1998; Miller, Beckwith, Fellbaum, 1988; Rogers, 2010], которые выявляют латентные семантические связи с помощью создания вычислительной модели.

Достижением последнего десятилетия можно считать появление дистрибутивных моделей, основанных на технологиях самообучающихся нейронных сетей: самая известная система «word2vec», способная работать на базе многомиллионных корпусов текстов. В исследовании [Литвинова, Паничева, 2024] данная языковая модель использовалась для реконструкции ассоциативных связей слов и для оценки эмоционально-психологического состояния респондентов.

Коллокация

Единицы, регулярно встречающиеся вместе (в одном контексте), называются коллокацией. В корпусных исследованиях большое внимание уделяется этому понятию: английский лингвист, один из основателей корпусной лингвистики Джон Синклер (John Sinclair) считал коллокацию центральным понятием современной лингвистики.

Выделяется два подхода к определению коллокаций: в рамках широкого подхода, представленного в вычислительных науках, коллокацией называется любая повторяющееся сочетание звуков. Однако более распространенным в лингвистике оказывается узкий подход к коллокациям как многословные единицы (multi-

wordunits): «Collocation – theco-occurrenceoftwoormorewords» [Teubert, Cermáková, 2007]. Дж. Синклер связал коллокацию с диапазоном (span), в рамках которого существует коллокация. Обсуждаемый диапазон связности можно проиллюстрировать НКРЯ при работе со сочетаниями: например, чаще всего в сочетании *повесить нос* составляющие элементы употребляются последовательно (то есть у них минимальный диапазон), однако отмечаются примеры, когда возможны и другие варианты, где компоненты разделены (тогда диапазон более широкий): *повесив горбатый нос, повесили сморщенные носы, уж не повесили ли они носы?*

На основе коллокаций можно составить конкорданс, то есть перечень контекстов употребления слов, который может служить основой для словаря. Например, с глаголом *бить* чаще всего употребляются существительные *морда, кнут, палка* и пр., что позволяет говорить, что чаще всего реализуется не первое значение (*то же, что ударять*) [Евгеньева, 1999], а четвертое значение, которое связано с причинением боли кому-нибудь.

Корпусные данные дают возможность рассчитать меру ассоциативности сочетания, которые показывают силу синтагматической связи между элементами в составе коллокации. Преимуществом такого вычислительного подхода является объективность результатов и доступ к обширному материалу, так как интуитивное определение степени спаянности вариативно, и не может быть надежным источником данных.

В лингвистике разные подходы к определению типов несвободных сочетаний, которые запоминаются носителями языка в готовом виде. В отечественной науке наиболее известной является классификация фразеологизмов В.В. Виноградова (фразеологические сращения, или идиомы, фразеологические единства, фразеологические сочетания). На практике это деление часто оказывается спорным и вариативным, поэтому актуальной является разработка объективных средств верификации сочетаний.

С помощью корпусных данных можно определить меру устойчивости сочетания. Разработаны несколько статистических инструментов (показателей) для этих целей, но чаще всего используются MI (MutualInformation) [Church, Hanks, 1996], Dice [Smadja, McKeown, Hatzivassiloglou, 1996], T-score и прочие, которые рассчитываются с помощью таких частотных данных (частотность совместного употребления, частотность каждой единицы и т. д.). Обзор данных показателей представлен в [Залеская, 2014].

Покажем, как это работает на примере нескольких сочетаний. Сравним два сочетания *бить баклуши* и *бить палкой*. Несмотря на то, что первое сочетание интуитивно ощущается несвободным, а второе – свободным, степень синтагматической связности у второго сочетания выше: *бить палкой* (MI 12,05), *бить баклуши* (MI 7,07). Такие значения показателей объясняются тем, что в корпусе сравнительно небольшое количество употреблений сочетаний *бить баклуши* (41 пример точного совпадения, 92 примера в разных грамматических формах в основном корпусе НКРЯ). Подобные статистические показатели в целом демонстрируют хорошие результаты.

Таким образом, традиционные подходы к несвободным сочетаниям (фразеологизмам, коллокациям, фраземам) могут контрастировать с новыми данными, полученными в ходе корпусных исследований. Как представляется, проблема несвободных сочетаниям требует дальнейшего изучения с учетом новых теоретических и эмпирических достижений.

Частотность

Возможно, одним из важнейших ресурсов, которые открывает корпус, являются сведения об языковой частотности, полученные на больших объемах данных. Феномен частотности в языке связан с вероятностным прогнозированием в речи, то есть опережением носителя языка в процессе восприятия речи [Sinclair, 1991; Венцов, Касевич, 2003]: слушающий строит гипотезы (предположения) о содержании и форме речи, основываясь на знаниях о типичном речевом поведении в этой ситуации. Частотные единицы воспринимают значительно быстрее, о чем свидетельствует, например, результаты эксперимента по праймингу, в рамках которого испытуемых просят установить, является ли предъявленная им цепочка букв (звуков) реальным словом. Распознавание частотных слов происходит значительно быстрее. Этот процесс описывают российские психолингвисты И. Горелов и К. Седов следующим образом: «Частотный словарь можно представить себе в виде пирамиды, на вершине которой располагаются немногие самые часто встречающиеся единицы, у широкого основания пирамиды – большинство встречающихся относительно редко» [Горелов, Седов, 2001, с. 86]. Как правило, частотные лексемы обладают большим количеством значений: такие единицы более активно подвержены семантической деривации (см., например, работы [Bybee, 2002; Hilpert, Gries, 2009]).

Важным преимуществом корпуса является возможность выявить частотность в диахронической перспективе. Например, с помощью корпусных данных можно определить, когда в русский язык пришло то или иное слово: так, слово *стёб* впервые фиксируется в текстах 1992 года, но период более активного использования приходится на период с 2010-х годов (пример употребления в (4)).

(4) *И Леню хлопали и по плечам, и по спине, и кто-то волосы ему взъерошил, а после трубоч сдохнул удачно, и снял несвойственный и даже вредный коллективу чересчур серьезный стёб* (НКРЯ: С. Солоух. Клуб одиноких сердец унтера Пришибеева, 1991–1995)

Благодаря морфологической разметке можно выделить особенности развития определенных частей речи. Например, во второй половине XX века фиксируется увеличение употребления вводных слов: на рисунке (см. Рис. 1) видно, что с 1960-х годов возрастает частотность вводного (дискурсивного) слова *естественно*, а пиковые значения частотности приходятся на период с 1995 по 2011 гг. Причем подобного изменения частотности не отмечается для слова *естественно*, употребляемого в роли предикатива, наречия, частицы.

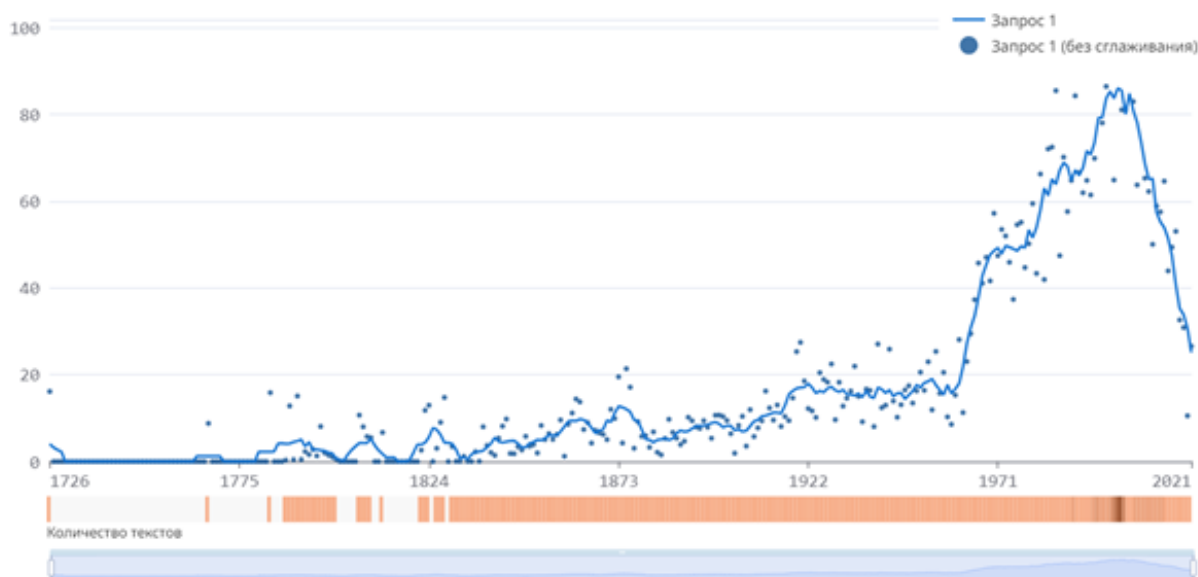


Рис. 1. Частотность употребления вводного слова *естественно* по временной шкале в Национальном корпусе русского языка (Основной корпус)

Идиостиль

Весьма любопытные возможности открывает корпусные технологии в изучении художественного текста и идиостиля отдельных авторов. Несмотря на то, что изучение идиостилей имеет прочную традицию в отечественной науке, современные корпусные данные значительно дополняют представления о языке писателя. Так, опубликована коллективная монография «Корпусная модель идиостиля Достоевского» [Баранов, Добровольский, 2021], построенная на основе сравнения данных корпуса текстов писателя с данными НКРЯ; подобное сопоставление выделяет особенности стиля Достоевского на общем фоне языка эпохи. Исследование проводится на разных уровнях – лексическом (то есть особенности словаря произведений писателя), синтаксическом (анализ синтаксических конструкций), нарративном (изучаются особенности построения сюжета) и интертекстуальном (находятся отсылки к другим текстам). Особое место уделяется идиоматике автора, например, для произведений оказываются частотны употребления дискурсивных слов *кстати* и *между прочим* по сравнению с фоновой частотностью употребления в языке эпохи.

Таким образом, корпусные технологии могут быть плодотворно использованы в качестве инструмента изучения авторского идиостиля.

Перевод

Широкие возможности открывают корпусные и компьютерные технологии для автоматического перевода (так называемого машинного перевода), а также для совершенствования практики и техники перевода [Baker, 1995; Zanettin, 2014; Камшилова, Беляева, 2023].

Системы автоматического (машинного) перевода, в том числе основанного на нейронном обучении, базируются на параллельных корпусах, которые являются уникальным инструментом для перевода. Параллельный корпус представляет собой

набор текстов на языке оригинала и его перевод, позволяющий найти соответствие между оригинальным и переводческим текстом: «Использование параллельных корпусов позволяет, например, найти точные переводные эквиваленты для каждого слова и выражения в конкретных текстах, причем все значительные факторы, обусловившие выбор того или иного способа перевода, могут быть изучены на аутентическом текстовом материале» [Добровольский, 2003, с. 13].

Современные системы нейронного обучения, которые используются для машинного перевода компанией «Google» с 2016 года для 30 мировых языков, используют обширные параллельные корпуса для обучения. Подобные алгоритмы, которые работают с целыми фразами, позволяют значительно повышать качество перевода. Предшествующие системы машинного перевода, называемые статистическими, также основывались на параллельных корпусах, однако использовали статистический анализ для поиска соответствий между текстом и его переводом, поэтому были неэффективны, например, при работе с низкочастотными единицами (было недостаточно данных для определения статистических вероятностей) и переводе сложных (целостных) высказываний, где нужно было учесть большое количество факторов.

Кроме параллельных корпусов, системы машинного перевода используют разные типы словарей (в том числе грамматические, словари идиом и пр.), морфологическую разметку и пр.

В целом параллельный корпус открывает широкие возможности для сопоставительных исследований. Так, значительно упрощается работа с так называемыми лингвоспецифичными словами, для которых трудно найти однозначные эквиваленты в других языках [Зализняк, Левонтина, Шмелев, 2005]. Примером может стать высокочастотное русское слово *нет*, которое на немецкий язык чаще переводится как *Doch* (устоявшийся перевод на русский язык *но*), а английском – *Yes* (*да*) [Добровольский, Левонтина, 2009].

Лингводидактика

Современные корпуса часто используют в учебных целях. Наиболее распространённой практикой является подбор языкового материала по необходимой цели. Подобные образовательные методики используются как для преподавания иностранного языка, так и для обучения родному. Например, в диссертационном исследовании [Чеботырева, 2024] корпус рассматривается как образовательная технология; в рамках работы она применяется для изучения паремий иностранного языка.

Для обучающих целей используются основной и параллельные корпуса, которые содержат не представленную в словарях информацию. Сведения корпусов позволяют подобрать широкий иллюстративный материал и проверить (расширить) данные учебников и словарей, разработать упражнения и задания, они могут стать основой для самостоятельной исследовательской работы обучающихся [Добрушина, 2009]. Корпусные данные могут быть использованы при обучении разных тем,

связанных лексической и грамматической системой языка, и текстовой организацией [Рычкова, Киеня, 2008].

Для образовательных целей создан особый вид корпуса – учебный (Learnercorpora), представляющий собой аннотированное собрание ошибок, которые допускают школьники или студенты в процессе изучения языка. Подобные корпуса создаются для изучения наиболее распространенных ошибок в изучаемых языках. На материале русского языка, например, создан Корпус русских учебных текстов (КРУТ) объемом более 2,6 млн слов, состоящий из текстов, написанных студентами разных вузов. Корпус сопровождается морфологической разметкой и разметкой по типам ошибок, что упрощает работу с данными. Выделяются следующие типы ошибок: лексические (пример: *обгоняешь и начинаешь спускаться вниз* (лишнее слово)); словообразовательные (*главная пешая (пешеходная) улица Стамбула просто наводнена людьми*); стилистические (*непомерно выпячивая на первый план свою личную жизнь, мы **подчас** забываем, что человек существо общественное*), грамматические (*не находя смысл за словами (слов), мы упускаем важную часть развития нашего прошлого*); дискурсивные (*чтобы стать специалистом в гражданской специализации нужно для начала изучить ее составляющие* (тавтология)).

Учебные корпуса можно использовать в научно-исследовательских целях, анализируя механизмы организации лексикона и грамматики носителя языка. Например, на основе интерпретации речевых ошибок построена известная модель ментального лексикона В. Левельта, которая предполагает декларативный компонент («знания что», знания о фактах) и процессуальный компонент («знания как», информация о языковых действиях) [Levelt, 1989].

Выводы. Итак, в настоящей работе представлен анализ ключевых направлений компьютерной лингвистики. Отдельно рассматриваются корпусные исследования, которые стали неотъемлемой частью современных лингвистических работ.

На основании проведенного анализа можно сделать вывод, что компьютерные технологии являются важной частью современных лингвистических исследований, расширив научный инструментарий и сферы практического применения лингвистических знаний. Компьютерная лингвистика сегодня – хороший пример синтеза достижений разных наук: лингвистики, математики, информационных технологий и когнитивных наук. Интегрируя методы и достижения разных наук, учёным удастся получать не только новые знания о языке и человеке в целом, но и добиваться значимых практических результатов. Можно сказать, что в этой сфере практические исследования, нацеленные на решение бытовых и утилитарных задач, опережают теоретическое осмысление полученных результатов.

При этом потенциал корпусных и компьютерных методов в лингвистике не освоен должным образом: как представляется, количество исследований с использованием названных технологий будет расти и станет обязательной частью любой языковедческой работы.

Литература

Балашов, Е. А., Баранов, А. Н., Добровольский, Д. О., Киселева, К. Л., Козеренко, А. Д., Коробова, М. М., Михайлов, М. Н., Осокина, Е. А., Фатеева, Н. А., Федорова, Л. Л., Шарапова, Е. В. (2021). *Корпусная модель идиостиля Достоевского*. Москва: ЛЕКСРУС.

Белов, В. А. (2020). Семантические исследования организации и функционирования ментального лексикона. *Научный диалог*, 8, 29–51. DOI: 10.24224/2227-1295-2020-8-29-51

Венцов, А. В., Касевич, В. Б. (2003). *Проблемы восприятия речи*. Москва: Едиториал УРСС.

Виноградов, В. В. (1977). Фразеология. Семасиология. *Лексикология и лексикография: избранные труды*. Москва: Наука. 118–16.

Горелов, И. Н., Седов, К. Ф. (2001). *Основы психолингвистики: учебное пособие*. 3-е изд., перераб. и доп. Москва: Лабиринт.

Добровольский, Д. О. (2003). Корпус параллельных текстов и литературный перевод. *Научно-техническая информация. Серия 2: Информационные процессы и системы*, 10, 13–18.

Добровольский, Д. О., Левонтина, И. Б. (2009). Русское нет, немецкое nein, английское no: сопоставительное исследование семантики на базе параллельных корпусов. *Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог 2009»*. 8(15). Москва: РГГУ. 97–101.

Добрушина, Н. Р. (2009). Корпусная методика обучения русскому языку. *Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы*. Санкт-Петербург: Нестор-История. 338–351.

Евгеньева, А. П. (ред.) (1999). *Словарь русского языка: в 4-х т.* Москва: Русский язык.

Залеская, В. В. (2014). Программа выявления в тексте двучленных статистически значимых осмысленных коллокаций (на материале русского языка). *Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2014), Санкт-Петербург, 19–20 ноября 2014 года*. Санкт-Петербург: Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики. 283–289. URL: <https://ojs.itmo.ru/index.php/IMS/article/download/267/263>

Зализняк, А. А., Левонтина, И. Б., Шмелев, А. Д. (2005). *Ключевые идеи русской языковой картины мира*. Москва: Языки славянской культуры.

Захаров, В. П., Богданова, С. Ю. (2020). *Корпусная лингвистика: учебник*. 3-е изд., перераб. Санкт-Петербург: Издательство Санкт-Петербургского университета.

Камшилова, О. Н., Беляева, Л. Н. (2023). Машинный перевод в эпоху цифровизации: новые практики, процедуры

References

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target: International Journal of Translation Studies*, 7(2), 223–243. 10.1075/target.7.2.03bak.

Baranov, A. N., Dobrovolsky, D. O. (eds.) (2021). *Corpus Model of Dostoevsky's Idiostyle*. E. A. Balashov, A. N. Baranov, D. O. Dobrovolsky, K. L. Kiseleva, A. D. Kozerenko, M. M. Korobova, M. N. Mikhailov, E. A. Osokina, N. A. Fateeva, L. L. Fedorova, E. V. Sharapova. Moscow: LEXRUS Publ., 2021. (In Russian).

Belov, V. A. (2020). Semantic Studies of the Organization and Functioning of the Mental Lexicon. *Scientific Dialogue*, 8, 29–51, 10.24224/2227-1295-2020-8-29-51. (In Russian).

Burgess, C., Lund, K. (2000). The dynamics of meaning in memory. *Cognitive dynamics: Conceptual and representational change in humans and machines*. Mahwah: Lawrence Erlbaum Associates Publishers, 117–156.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14, 261–290, 10.1017/S0954394502143018.

Chebotyeva, K. A. (2024). *Application of corpus technology in the process of teaching paremiological units to schoolchildren of specialized classes: Abstract of diss... Candidate of Pedagogical Sciences*. Nizhny Novgorod. (In Russian).

Chilingaryan, K. P. (2021). Corpus linguistics: theory VS methodology. *Bulletin of Peoples' Friendship University of Russia. Series: Language Theory. Semiotics. Semantics*, 1, 196–218, 10.22363/2313-2299-2021-12-1-196-218. (In Russian).

Church, K., Hanks, P. (1996). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29, 10.3115/981623.981633.

Dobrovolsky, D. O., Levontina, I. B. (2009). Russian no, German nein, English no: a comparative study of semantics based on parallel corpora. *Computational linguistics and intelligent technologies. Proceedings of the international conference "Dialogue 2009"*. Moscow, Russian State University for the Humanities Publ., 97–101. (In Russian).

Dobrovolsky, D.O. (2003). Corpus of Parallel Texts and Literary Translation. *Nauchno-tehnicheskaya informatsiya. Seriya 2: Informatsionnyye protsessy i sistemy*, 10, 13–18. (In Russian).

Dobrushina, N. R. (2009). Corpus-based methods of teaching Russian. *National Corpus of the Russian Language. 2006–2008. New results and prospects*. St. Petersburg: Nestor-Istoriya Publ., 338–351. (In Russian).

Evgenyeva, A. P. (ed.) (1999). *Dictionary of the Russian Language: In 4 volumes*. Moscow: Russkiy Yazyk Publ., 1999. (In Russian).

Firth, J. R. (1957). *Papers in Linguistics: 1934–1951*. Oxford: Oxford University Press.

- и ресурсы. *Terra Linguistica*, 14 (1), 41–56. DOI: 10.18721/JHSS.14105
- Кибрик, А. А., Плунгян, В. А. (2002). Функционализм. *Современная американская лингвистика: фундаментальные направления* / под редакцией: А. А. Кибрика, И. М. Кобозевой, И. А. Секериной. 2-е изд, испр. и доп. Москва: Едиториал УРСС. 276–339.
- Литвинова, Т. А., Паничева, П. В. (2024). Индивидуальные различия в ассоциативном значении слова сквозь призму языковой модели и семантического дифференциала. *Научный результат. Вопросы теоретической и прикладной лингвистики*, 10(1), 61–93. DOI: 10.18413/2313-8912-2024-10-1-0-5
- Лукашевич, Н. В., Левчик, А. В. (2016). Создание лексикона оценочных слов русского языка RuСентилекс. *Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2016): материалы VI международной научно-технической конференции, Минск, 18-20 февраля 2016 года*. Минск: БГУИР. 377–382.
- Ляшевская, О. Н. (2016). *Корпусные инструменты в грамматических исследованиях русского языка*. Москва: Языки славянской культуры: Рукописные памятники Древней Руси.
- Плунгян, В. А. (2007). Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики. *Национальный корпус русского языка и проблемы гуманитарного образования: материалы международной научной конференции, Москва 19-20 апреля 2007 г.* Москва: Высшая школа экономики. 64–66.
- Радбиль, Т. Б. (2024). Выявление оценочного потенциала нейтрального слова в поэзии (по данным поэтических интернет-корпусов). *Критика и семиотика*, 1, 138–157. DOI: 10.25205/2307-1753-2024-1-138-157
- Романов, А. С., Васильева, М. И., Куртукова, А. В., Мещеряков, Р. В. (2018). Анализ тональности текста с использованием методов машинного обучения. *R. Piotrowski's Readings in Language Engineering and Applied Linguistics: Proceedings*, Saint Petersburg, November 27, 2017. Saint Petersburg: Creative Commons ССО. 86–95.
- Рубцова, Ю. (2012). Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора. *Инженерия знаний и технологии семантического веба*, 1, 109–116.
- Рычкова, Л. В., Киеня, С. Н. (2010). Корпусные технологии в преподавании РКИ. *Этнокультурный и социолингвистический аспекты в теории и практике преподавания языков в негуманитарных вузах: сборник научных статей*. Минск: Белорусский национальный технический университет. 32–43.
- Рюкова, А. Р. (2024). Корпусно-ориентированные исследования языка: краткий обзор достижений и трудностей. *Russian Linguistic Bulletin*, 1 (49), 24. DOI: 10.18454/RULB.2024.49.17
- Gorelov, I. N., Sedov, K. F. (2001). *Fundamentals of Psycholinguistics*. Moscow: Labirint Publ. (In Russian).
- Hilpert, M., Gries, S. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition Get access Arrow. *Literary and Linguistic Computing*, 24 (4), 385–401, 10.1093/lc/fqn012.
- Jurafsky, D., Martin, J. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Stanford.
- Kamshilova, O. N., Belyaeva, L. N. (2023). Machine translation in the era of digitalization: new practices, procedures and resources. *Terra Linguistica*, 1, 41–56, 10.18721/JHSS.14105. (In Russian).
- Kibrik, A. A., Plungyan, V. A. (2002). Functionalism. *Modern American Linguistics: Fundamental Directions*. Ed. by A. A. Kibrik, I. M. Kobozeva, I. A. Sekerina. Moscow: Editorial URSS Publ., 276–339. (In Russian).
- Landauer, Th., Foltz, P., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25 (2-3), 259–284, 10.1080/01638539809545028.
- Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge: MIT Press.
- Litvinova, T. A., Panicheva, P. V. (2024). Individual differences in the associative meaning of a word through the lens of the language model and semantic differential. *Theoretical and Applied Linguistics*, 10(1), 61–93, 10.18413/2313-8912-2024-10-1-0-5. (In Russian).
- Lukashevich, N. V., Levchik, A. V. (2016). Creation of a lexicon of evaluative words of the Russian language RuSentileks. *Proceedings of the Open Semantic Technologies for Intelligent Systems (OSTIS-2016) conference*. Minsk: Belarusian State University of Informatics And Radioelectronics Publ., 377–382. (In Russian).
- Lyashevskaya, O. N. (2016). *Corpus tools in grammatical studies of the Russian language*. Moscow: Yazyki slavyanskoy kul'tury: Rukopisnyye pamyatnik Drevney Rusi Publ. (In Russian).
- Mcenery, T., Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- Miller, G., Beckwith, R., Fellbaum, C. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), 235–244, 10.1093/ijl/3.4.235.
- Plungyan, V. A. (2007). Corpus as a tool and as an ideology. *National corpus of the Russian language and problems of humanitarian education. Proceedings of the international scientific conference. Moscow, April 19-20, 2007*. Moscow: Higher School of Economics Publ. 64–66. (In Russian).
- Radbil, T. B. (2024). Identifying the evaluative potential of a neutral word in poetry (based on online poetry corpora). *Critique and Semiotics*, 1, 138–157, 10.25205/2307-1753-2024-1-138-157. (In Russian).

- Савчук, С. О., Архангельский, Т. А., Бонч-Осмоловская, А. А., Дони́на, О. В., Кузнецова, Ю. Н., Ляшевская, О. Н., Орехов, Б. В., Подрядчикова, М. В. (2024). Национальный корпус русского языка 2.0: новые возможности и перспективы развития. *Вопросы языкознания*, 2, 7–34. DOI: 10.31857/0373-658X.2024.2.7-34
- Софронова, Е. В. (2024). *Automated Sentiment Analysis of Femininitives in the Russian Language: выпускная квалификационная работа магистра: направление 45.04.04 «Интеллектуальные системы в гуманитарной среде»; образовательная программа 45.04.04_01 «Цифровая лингвистика (международная образовательная программа) / Digital Linguistics (International Educational Program)»*. Санкт-Петербург: Санкт-Петербургский политехнический университет Петра Великого. DOI 10.18720/SPBPU/3/2024/vr/vr24-5826. Авторизованным пользователям СПбПУ.
- Чеботырёва, К. А. (2024). *Применение корпусной технологии в процессе обучения паремнологическим единицам школьников профильных классов: автореферат диссертации на соискание ученой степени кандидата педагогических наук: специальность 5.8.2*. Нижний Новгород: Нижегородский государственный лингвистический университет им. Н. А. Добролюбова.
- Чилингарян, К. П. (2021). Корпусная лингвистика: теория VS методология. *Вестник Российского университета Дружбы народов. Серия. Теория языка. Семиотика. Семантика*, 12 (1), 196–218. DOI: 10.22363/2313-2299-2021-12-1-196-218
- Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target: International Journal of Translation Studies*, 7(2), 223–243. DOI: 10.1075/target.7.2.03bak.
- Burgess, C., Lund, K. (2000). The dynamics of meaning in memory. *Cognitive dynamics: Conceptual and representational change in humans and machines*. Mahwah: Lawrence Erlbaum Associates Publishers. 117–156.
- Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14, 261–290. DOI: 10.1017/S0954394502143018
- Firth, J. R. (1957). *Papers in Linguistics, 1934–1951*. London, etc.: Oxford University Press.
- Church, K., Hanks, P. (1996). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29. DOI: 10.3115/981623.981633
- Hilpert, M., Gries, S. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24 (4), 385–401. DOI: 10.1093/lc/fqn012
- Jurafsky, D., Martin, J. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Stanford.
- Landauer, Th., Foltz, P., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25 (2-3), 259–284. DOI: 10.1080/01638539809545028
- Rogers, T. (2008). Computational models of semantic memory. *The Cambridge Handbook of Computational Psychology*. Cambridge, Cambridge University Press. 226–267. DOI: 10.1017/CBO9780511816772.012.
- Romanov, A. S., Vasilyeva, M. I., Kurtukova, A. V., Meshcheryakov, R. V. (2018). Sentiment analysis of texts using machine learning methods. *Proceedings of the 2nd International Conference “R. Piotrowski’s Readings in Language Engineering and Applied Linguistics” (Saint Petersburg, 2017)*. Saint Petersburg: Creative Commons CCO, 86–95. (In Russian).
- Rubtsova, Yu. (2012). Automatic construction and analysis of a corpus of short texts (microblog posts) for the task of developing and training a tone classifier. *Knowledge Engineering and Semantic Web Technologies*, 1, 109–116. (In Russian).
- Rychkova, L. V., Kienya, S. N. (2010). Corpus technologies in teaching Russian as a foreign language. *Ethnocultural and sociolinguistic aspects in the theory and practice of teaching languages in non-humanitarian universities: Collection of scientific articles*. Minsk: Belarusian National Technical University, 32–43. (In Russian).
- Ryukova, A.R. (2024). Corpus-oriented language studies: a brief summary of achievements and challenges. *Russian Linguistic Bulletin*, 1(49), 10.18454/RULB.2024.49.17. (In Russian).
- Savchuk, S. O., Arkhangelsky, T. A., Bonch-Osmolovskaya, A. A., Donina, O. V., Kuznetsova, Yu. N., Lyashevskaya, O. N., Orekhov, B. V., Podryadchikova, M. V. (2024). Russian National Corpus 2.0: New opportunities and development prospects. *Voprosy yazykoznaniiya*, 2, 7–34, 10.31857/0373-658X.2024.2.7-34. (In Russian).
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Smadja, F. McKeown, K., Hatzivassiloglou, V. (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1), 1–38.
- Sofronova, E. V. (2024). *Automated Sentiment Analysis of Femininitives in the Russian Language: Master’s thesis: direction 45.04.04 “Intelligent systems in the humanitarian environment”*. St. Petersburg: Peter the Great St. Petersburg Polytechnic University, 10.18720/SPBPU/3/2024/vr/vr24-5826. (In Russian).
- Teubert, W., Cermakova, A. (2007). *Corpus Linguistics: A Short Introduction*. London: Bloomsbury Academic.
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Philadelphia: John Benjamins Publ., 223.
- Ventsov, A. V., Kasevich, V. B. (2003). *Problems of Speech Perception*. Moscow: Editorial URSS. Publ. (In Russian).
- Vinogradov, V. V. (1977). *Phraseology. Semasiology. Lexicology and Lexicography. Selected Works*. Moscow: Nauka Publ., 118–16. (In Russian).
- Zakharov, V. P., Bogdanova, S. Yu. (2020). *Corpus linguistics*. St. Petersburg: St. Petersburg University Publ. (In Russian).

- Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge: MIT Press.
- Miller, G., Beckwith, R., Fellbaum, C. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), 235–244. DOI: 10.1093/ijl/3.4.235
- Rogers, T. (2008). Computational models of semantic memory. *The Cambridge Handbook of Computational Psychology*. Cambridge: Cambridge University Press. 226–267. DOI: 10.1017/CBO9780511816772.012
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Smadja, F., McKeown, K., Hatzivassiloglou, V. (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1), 1–38.
- Teubert, W., Cermakova, A. (2007). *Corpus Linguistics: A Short Introduction*. London: Bloomsbury Academic.
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Philadelphia: John Benjamins Publishing.
- Mcenery, T., Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
- Zanettin, F. (2014). *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. London; New-York: Routledge. DOI: 10.4324/9781315759661. (Vol. 14: Translation-Driven Corpora)
- Zalesskaya, V. V. (2014). A program for identifying statistically significant meaningful binomial collocations in the text (based on the Russian language). *XVII All-Russian United Conference "Internet and Modern Society" (IMS-2014)*. St. Petersburg. Electronic resource. Retrieved from: <https://ojs.itmo.ru/index.php/IMS/article/download/267/263>. (In Russian).
- Zaliznyak, Anna A., Levontina, I. B., Shmelev, A. D. (2005). *Key ideas of the Russian linguistic picture of the world*. Moscow: Yazyki slavyanskoy kul'tury. (In Russian).
- Zanettin, F. (2014). *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. London; New-York: Routledge Publ. <https://doi.org/10.4324/9781315759661>.

Для цитирования статьи:

Белов, В. А. (2024). Компьютерные технологии в лингвистике. *VERBA. Северо-Западный лингвистический журнал*, 3(13), 8–23. DOI: 10.34680/VERBA-2024-3(13)-8-23

For citation:

Belov, V. A. (2024). Computer Technologies in Linguistics. *VERBA. North-West linguistic journal*, 3(13), 8–23. (In Russian). DOI: 10.34680/VERBA-2024-3(13)-8-23

Сравнительно-сопоставительный анализ лингвистических ресурсов для проведения корпусного анализа текстов

А. В. Дмитриев, Е. С. Крупнова

Comparative-Contrastive Analysis of Linguistic Resources for Corpus Analysis of Texts

A. V. Dmitrijev, E. S. Krupnova

Александр Владиславович Дмитриев – кандидат филологических наук, доцент; Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

E-mail: avd84@list.ru

Елена Сергеевна Крупнова – магистр, специалист по учебно-методической работе; Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

E-mail: krupnalena@mail.ru

Статья поступила: 05.10.2024. Принята к печати: 20.10.2024.

В статье рассматривается основная задача корпусной лингвистики – корпусный анализ письменных текстов на естественном языке с помощью лингвистических ресурсов, которые используются для её решения. Корпусный анализ подразумевает метод исследования языка, который использует большие коллекции текстов или корпуса для получения статистических и лингвистических данных о языке. Лингвистические ресурсы, такие как словари, тезаурусы, грамматические базы данных значительно расширяют возможности и точность корпусного анализа. Помимо этого, корпусная лингвистика занимается созданием корпусных менеджеров, которые обрабатывают тексты и выполняют функции составления конкорданса, поиска ключевых слов, коллокаций и другие. В работе кратко описывается функционал программ WMatrix, WordSmith, GATE, AntConc и Sketch Engine, а также проводится сравнительно-сопоставительный анализ их характеристик. В результате сделан вывод о том, что ряд программ отличается набором функций, параметрами сохранения данных, форматом входного текста и доступностью. Кроме того, перечисляются направления их использования в научно-практической деятельности. Лингвистические ресурсы могут быть полезны для стилистического анализа текстов, изучения лингвистических особенностей авторского стиля, обучения иностранному языку, например, грамматике или лексике, в компьютерной лексикографии, дискурс-анализе и в других направлениях. Рассмотренные инструменты не только повышают точность анализа, но и расширяют возможности, интегрируясь в программные инструменты для автоматизации корпусного анализа. Выбор подходящего инструмента для проведения исследования зависит от объёма и глубины анализа текста.

Alexander V. Dmitrijev – Candidate of Philological Sciences, Associate Professor; Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russian Federation

ORCID: 0000-0003-3632-793X

Elena S. Krupnova – Master's degree, specialist in educational and methodological work; Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russian Federation

ORCID: 0009-0007-3127-2737

Received: 05.10.2024. Accepted for publication: 20.10.2024.

In the last few decades, a scientific field known as computational linguistics has been actively developing. The paper discusses the main task of corpus linguistics – corpus analysis of written natural-language texts with the help of linguistic resources that are used to solve it. Corpus analysis refers to a method of language research that utilizes large collections of texts or corpora to obtain statistical and linguistic data about the language. Linguistic resources such as dictionaries, thesauri, and grammatical databases greatly enhance the capability and accuracy of corpus analysis. In addition, corpus linguistics deals with the building of corpus managers that process texts, perform concordance, search for keywords and collocations, etc. The paper briefly describes the functionality of WMatrix, WordSmith, GATE, AntConc and Sketch Engine programs and makes a comparative-contrastive analysis of their characteristics. It is concluded that the programs differ in feature set, data saving parameters, input text format and accessibility. In addition, directions for their use in research and practice are suggested. Linguistic resources can be useful for stylistic analysis of texts, studying linguistic features of author's style, teaching a foreign language, for example, grammar or vocabulary, in computer lexicography, discourse analysis and other directions. The example of the corpus analysis of the topic *famine* during the blockade of Leningrad with the help of the AntConc program is given. In the course of the mentioned research, 749 fragments of memories of Leningrad citizens were collected on the basis of 15 frequency words and a frequency dictionary of 158 words was compiled. Considered tools not only increase the accuracy of analysis, but also expand the possibilities and integrate into software tools for automation of corpus analysis. The choice of the appropriate tool for the study depends on the scope and depth of text analysis.

Ключевые слова: корпусная лингвистика, лингвистические корпуса, корпусный менеджер, стилистический анализ текста

УДК 81'322:81'42

Keywords: natural language processing, corpus linguistics, linguistic corpora, corpus manager, corpus stylistics, stylistic corpus analysis

OECD: 6.02OT

V

Постановка проблемы. В последние несколько десятков лет активно развивается научная область, компьютерная лингвистика, основным фокусом которой является автоматическая обработка письменных текстов на естественном языке (Natural Language Processing или NLP). Компьютерная лингвистика – междисциплинарная область, которая возникла на стыке лингвистики, информатики, математики и искусственного интеллекта. Основными её прикладными задачами являются машинный перевод, классификация и кластеризация текстов, поиск и извлечение информации, индексирование, реферирование, аннотирование, интеллектуальный анализ данных, формирование ответов на вопросы, анализ тональности текстов, распознавание и синтез звучащей речи и другие.

Анализ языковых данных является непростой задачей и для её решения требуется большой массив данных, в котором содержится несколько сотен тысяч примеров употреблений тех или иных слов. Этим вопросом занимается другой подраздел компьютерной лингвистики – корпусная лингвистика, которая занимается созданием и использованием лингвистических корпусов для решения различных задач в области лингвистики и смежных областей [Николаев, 2016]. Актуальность исследования заключается в необходимости автоматизации процесса обработки текстов на естественном языке и их корпусном анализе для решения различных прикладных задач, как поиск и извлечение информации, анализ кореферентности, машинное обучение, обучение переводу и многие другие. На данный момент существует множество инструментов, с помощью которых можно проанализировать тексты, однако среди них можно выявить ряд отличий в наборе функций и инструментов. Для получения наиболее эффективных результатов проведённого исследования необходимо знать, какой лингвистический ресурс подойдёт лучше всего. Кроме того, программы со временем дорабатываются и в них появляются новые функции, которые не отражены в некоторых существующих работах.

История вопроса. Вопросы использования корпусов и корпусных менеджеров были рассмотрены в [Захаров, 2005]. Лингвистический корпус – «большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [Захаров, 2005, с. 3]. Термин «корпус текстов» также обозначается «корпусным менеджером», специализированной системой поиска, с помощью которой можно искать ключевые слова, фразы, коллокации и контексты словоформ в корпусе, создавать конкордансы, составлять списки слов по заданным критериям, получать статистическую информацию о частоте употребления слова в корпусе и представлять результаты в удобном формате для пользователя. Такие программы быстро обрабатывают результаты и интуитивно понятны в использовании.

Рассматриваемые в статье программы были проанализированы с точки зрения функционала в некоторых работах. Например, сравнение двух программ Antconc и Sketch Engine для исследования коллокаций в английском языке, отдельное описание функционала программы Antconc, Sketch Engine, а также сравнение Antconc, Wordsmith Tools и Sketch Engine на материале текстов кинодискурса [Палийчук, 2022; Котюрова, 2020; Кротова, 2019; Шамова, 2021].

Методология и методика исследования. Цель исследования заключается в сравнительно-сопоставительном анализе пяти наиболее популярных лингвистических инструментов – WMatrix, WordSmith, GATE, AntConc и Sketch Engine, используемых для автоматической обработки и корпусного анализа текстов, и их применении в научно-практической деятельности.

Исследование проводится с применением следующих методов: отбор и анализ данных о функциональности наиболее популярных лингвистических ресурсов, сравнительно-сопоставительный анализ инструментов и метод комплексного описания полученных результатов.

Анализ материала. Для проведения анализа были взяты 5 наиболее популярных лингвистических ресурса WMatrix, WordSmith, GATE, AntConc and Sketch Engine. Кратко охарактеризуем программы.

WMatrix. Программа была разработана в конце XX века Полом Рейсоном в рамках проекта REVERE, цель которого состояла в изучении вопроса извлечения информации из документов, связанных с разработкой программного обеспечения. С помощью данной программы можно выделять ключевые слова, определять часть речи слова, проводить анализ текстов на уровне грамматики и семантики, визуализировать частотность употребления слов в корпусе и анализировать конкордансы, исследуя единицы языка в их контекстуальном окружении. Простой функционал программы WMatrix6 состоит из четырёх функций: поиск списка слов и их частотность, поиск конкретного слова и облака слов, позволяющие увидеть значимость слов [Rayson]. Стоит также отметить, что на данный момент доступ к программе ограничен.

После загрузки текста в программу можно провести анализ данных: увидеть список наиболее частотных слов по частям речи и по семантическому тегу; класс слов, отражающих тему текста; конкорданс слова или тега и просмотр контекста слева или справа; визуализацию частотности слов и тегов (чем крупнее шрифт, тем наиболее значимо слово в тексте или, другими словами, тем чаще оно появляется в тексте). Кроме того, есть возможность сравнить файл со стандартным эталонным корпусом для нахождения ключевых слов и ключевых семантических категорий, связанных с эталонным корпусом.

В WMatrix5 можно также провести более продвинутый анализ данных как токенизацию, получение многословных выражений (семантический теггер автоматически размечает их как единое целое с помощью одного семантического тега) и n-грамм (или их ещё называют кластерами, лексическими связками или устойчивыми выражениями). Кроме того, доступны функции получения списка частотности лемм, частотности тега по части речи и анализа ключевых частей речи, а также коллокаций (словосочетания, которые встречаются либо слева, либо справа от

слова). Ещё одной особенностью программы является анализ метафор с использованием семантических тегов.

Ещё одним инструментом являются коллокации. В программе в колонке “Word” можно увидеть слово, а в “Collocate” словосочетание, которое встречается либо слева, либо справа от него. В других шести колонках (L3 – R3) содержится информация о частотности появления словосочетания в позиции слева от слова – одно слово (L1), два слова (L2) и три слова (L3). Таким же образом в позиции справа от слова – R1, R2 и R3. В колонке “Total” указывается сумма всех позиций, в которых словосочетание встречается со словом. В колонке “Word Freq” – число встречаемости слова в корпусе, в “Collocate Freq” – число словосочетаний. На основе этой информации можно составить таблицу для каждого слова и пары слов и посчитать показатели статистических данных. Двумя основными метриками в программе являются MI (взаимная информация) и LL2 (the two-cell Log-Likelihood или правдоподобие). С помощью первой можно выделить устойчивые коллокации.

Sketch Engine. Проект, разработанный лингвистом Адамом Киллгарриффом и чешским программистом Павлом Рыхли. С помощью этого инструмента можно анализировать аутентичные тексты, для выявления необычных в использовании и редких элементов языка. Он также предназначен для анализа текста или приложений для интеллектуального анализа текста.

Программа обеспечивает репрезентативную выборку языка, поскольку содержит 800 готовых к использованию корпусов на более чем 100 языках, каждый из которых имеет размер до 80 миллиардов слов [программа Sketch Engine].

Sketch Engine предоставляет большой набор функций, основными которого являются: Concordance (поиск примера употребления вводимого слова, леммы, фразы или словосочетания); Word list (список всех слов в выбранном корпусе); Keywords and Terms (ключевые слова и термины); Collocations (коллокации, стоящие слева или справа от заданного слова); Thesaurus (нахождение слов, которые появляются в похожем контексте, как и вводимое слово); Word Sketch (описания возможной сочетаемости слова с другими); Word Sketch Differences (сравнение скетчей для двух лексических единиц); WebBootCaT (создание собственного корпуса текстов) и Trends (изменение частоты слов в корпусе) [Палийчук, 2022].

Лингвистический процессор GATE (General Architecture for Text Engineering) предназначен для решения различных задач: ручная и автоматическая семантическая разметка текстов, извлечение информации, анализ кореферентных связей в тексте, машинное обучение, работа с онтологиями [Рубайло, 2016]. Последняя выпущенная версия на 2024 год – версия 9.0.1.

Первая версия системы GATE была выпущена в 1996 году разработчиками из университета Шеффилда. Она была написана на языке C++ и использовалась в широком спектре контекстов анализа языка, включая извлечение информации на английском, греческом, испанском, шведском, немецком, итальянском и французском языках [Большакова, 2011].

Данная система включает в себя три важные составляющие: архитектура, фреймворк и среда разработки. Под архитектурой понимается абстрактное описание того, как может быть построена система обработки языка, типы обычно используемых

компонентов и так далее. Компоненты – многократно используемые программные блоки с четко определенными интерфейсами. Под фреймворком подразумевается объектно-ориентированная библиотека классов, реализующая архитектуру и предоставляющую ряд сервисов, которые можно использовать в различных контекстах приложений. Одним из таких приложений является среда разработки, построенная на основе фреймворка.

Платформа GATE поддерживает множество текстовых форматов, включая Plain Text, PDF, Email, форматы Microsoft Office и другие. В программе есть возможность сохранить корпуса текстов и документы и продолжить с ними работать в дальнейшем.

Преимуществами данной программы являются следующие:

- удобный графический интерфейс;
- экспорт аннотаций в XLM формате;
- возможность повторного использования лингвистических компонентов;
- представление общей базы для разработки приложений и компонентов.

Недостатками являются:

- сложность и неочевидность графического интерфейса;
- отсутствие возможности добавления новых компонентов для визуализации данных;
- неэффективная реализация базы данных;
- необходимость совместимости схем аннотаций для интеграции различных компонентов;
- отсутствие поддержки компонентов, представляющих источники данных [Большакова, 2011].

WordSmith Tools. Программный пакет, предназначенный в первую очередь для работы с текстовыми файлами. Набор программ был разработан британским лингвистом Майком Скоттом и впервые выпущен в 1996 году. На данный момент доступна бесплатная версия 9.0. Стоит отметить, что WordSmith Tools можно использовать на 80 языках.

Программный пакет состоит из основных трёх модулей:

1. Модуль “Concord” используется для создания конкордансов (списка всех употреблений вводимого слова или фразы в контексте). С его помощью можно отобрать коллокации или кластеры слов, а также их местоположение в тексте.
2. Модуль “WordList” используется для составления списка всех слов или словоформ, включённых в выбранный корпус, а также частотного списка.
3. Модуль “KeyWord” используется для создания списка ключевых слов и грамматических форм в соответствии с определёнными статистическими критериями.

Кроме этого, каждый модуль содержит в себе другие функции анализа текста, как, например, поиск словосочетаний – коллекции слов, которые наиболее часто используются в тексте вместе с одним определённым словом. Также есть ряд

дополнительных модулей, которые полезны для подготовки, очистки и форматирования текстового корпуса. Главным преимуществом программы является выгрузка результатов на экран или их сохранение в отдельный файл.

AntConc. Бесплатный программный инструмент, разработанный Лоуренсом Энтони в 2011 году, для проведения статистических исследований текстов. С его помощью можно идентифицировать и подсчитывать длинные кластеры, составлять конкорданс для заданного слова в пределах контекстного окна, частотный список словоформ или лемм с указанием ранга и абсолютной частоты; выделять ключевые слова художественного текста; выявлять связи между полученным ключевым словом и анализом ключевой семантической области; искать длинные n-граммы для определения их значимости, ценности и связи со значением слова [Николаев, 2016]. Помимо этого, данная программа используется для стилистического анализа текста на основе методов корпусной лингвистики. AntConc выполняет следующие функции:

- 1) построение конкорданса для заданного слова в пределах контекстного окна;
- 2) построение графика к конкордансу;
- 3) построение частотного списка словоформ или лемм с указанием ранга и абсолютной частоты
- 4) выделение ключевых слов художественного текста, которые могут помочь читателям понять смысл текста.
- 5) выявление связи между полученным ключевым словом и анализом ключевой семантической области.
- 6) поиск коллокаций заданного слова на основе мер ассоциации;
- 7) поиск длинных n-грамм для определения их значимости, ценности и связи со значением слова [Николаев, 2016].

Программа состоит из девяти инструментов: KWIC (результаты поиска употребления слов или фраз в корпусе текстов); plot (график конкорданса) для представления результатов поиска в виде штрих-кода, который показывает местоположение, где слово или фраза появляется в корпусе; функция File (просмотр файла) отображает содержание текста; clusters (кластеры) показывает группы слов, которые появляются в тексте рядом с ключевым словом, N-Gram (списки n-gram) для отображения наиболее частотных словосочетаний в корпусе; функция Collocate для нахождения слов, которые сочетаются с другими словами; Word list для получения упорядоченного списка всех слов в корпусе от наиболее частотных к менее, Keyword для определения слов, часто употребляемых в текстах по сравнению с эталонным корпусом, и Wordcloud (визуализация результатов, полученных с помощью инструментов KWIC, File, Cluster, N-Gram, Collocate, Word и Keyword, в виде «облака слов» (например, наиболее частотные слова представлены шрифтом большего размера). Ограничением данной программы заключается в объеме корпуса – допускается небольшой размер.

Сравним в таблице представленные выше программы по функционалу, формату вводимого текста/файла, доступности, сохранении данных и языку интерфейса.

Таблица 1. Сравнительный анализ лингвистических ресурсов

	WMatrix	AntConc	WordSmith Tools	GATE	Sketch Engine
конкорданс	+	+	+	–	+
график конкорданса	–	+	–	–	–
кластеры	–	+	+	–	+
N-граммы	+	+	–	–	+
коллокации	+	+	+	–	+
частотность слов	+	+	+	–	+
список слов	+	+	+	–	+
ключевые слова	+	+	+	–	+
Wordcloud	+	+	–	–	+
разметка текста	+	–	+	+	+
формат файла/ текста	plain text format, html, sgml, xml	txt, srt, csv, tsv, html, xml, docx, pdf	plain text format, html, xml	plain text format, html, sgml, xml, rtf, Email, OpenOffice, pdf, Microsoft Office	txt, doc, docx, pdf, xml, html, htm, zip, tar.bz2, tar.gz, tgz, vert, ps, tmx
доступность	–	+	+	+	+ (30 дней бесплатно)
сохранение данных	–	–	+	+	+
интерфейс на русском языке	–	–	–	–	–

Результаты исследования

Исходя из результатов составленной таблицы, можно сделать следующие выводы:

1. AntConc, Sketch Engine, WMatrix и WordSmith позволяют работать с коллокациями, кластерами, ключевыми словами текстов, n-граммами, списками слов и конкордансом. Их функционал похож и отличается несколькими параметрами.
2. В программах Antconc, GATE и Sketch Engine файлы могут быть загружены в большом ряде форматов, что удобно в их использовании.
3. Стоит также отметить, что из всех представленных программ только WMatrix ограничен в доступе на момент 2024 года. Остальные бесплатные и их можно скачать локально.
4. В программах WordSmith Tools, GATE и Sketch Engine есть возможность сохранять полученные данные и использовать их повторно.
5. Интерфейс всех ресурсов доступен только на английском языке, однако он интуитивно понятен в использовании.

В целом можно отметить, что для подробного стилистического анализа текста наиболее подходящей является программа Sketch Engine или AntConc. Первая программа имеет такой недостаток, как ограниченный период бесплатного доступа, а вторая — не предусматривает функцию сохранения полученных данных, что может быть неудобным и требует переноса результатов в документ.

В связи с тем, что рассмотренные лингвистические ресурсы по обработке и корпусному анализу текстов были созданы несколько лет назад, они были использованы в различных направлениях, как изучение онтологий, обучение чат-ботов на основе сравнения диалогов между людьми и человеком и компьютером, сравнение письменного и устного типов речи английского языке, анализ ключевых слов и тем в литературных текстах, фразеология, метафоры в политическом или бизнес дискурсе, анализ интернет-блогов Синглиша, изучение языковых вариаций с помощью метода извлечения ключевых тем, использование корпусов в процессе обучения специальному переводу и другие. Приведём несколько направлений использования корпусных менеджеров и систем в научной и практической деятельности.

Одним из актуальных направлений является корпусная стилистика – область, которая стала активно развиваться с конца 1960-х годов, но популярным стала в 2000-е годы. С 2007 года отмечается «корпусный поворот» в стилистике, о чём свидетельствует большое количество энциклопедических статей о практике проведения анализа, а также монографий, в которых описывается данный подход [Leech, 2007; Mahlberg, 2012b; McIntyre, 2015].

Помимо изучения стиля текста при анализе слов, фраз и предложений, частей речи, знаков препинания, пауз, выявления языковых особенностей, которые остаются незамеченными при ручном анализе текста, программы используются для стилистического анализа художественных текстов, анализа фразеологии, сегментации текста, связности и когерентности [Fischer, 2010; Mahlberg, 2012a].

Лингвистические инструменты могут быть также использованы для исследования парцелляции в газетном дискурсе, английской абсолютной причастной конструкции, морфологических и грамматических ошибок в письменной речи носителей языка, дискредитирующих тактик в дискурсе социальных медиа и других работах.

В качестве ещё одного направления исследования можно рассмотреть стилистический анализ текстов с применением метода корпусной лингвистики. Например, анализ темы «голод» во время блокады Ленинграда при рассмотрении и описании жизни жителей города, описанная в книге Алеся Адамовича и Даниила Гранина «Блокадная книга». Для проведения корпусного анализа была выбрана программа AntConc и использованы её четыре функции: Wordlist, Concordance, File view и Clusters.

Из полученного списка слов были отобраны 15 по теме «голод» и взяты их основы: голод*, хлеб*, холод*, смерть*, грамм*, карточк*, ужас*, кусоч*, труп*, съе*, испытани*, еда, слаб*, дистрофи* и болезн* (звёздочкой обозначено любое количество символов после основы). Далее были найдены контексты употребления отобранных слов. В результате исследования было собрано 747 отрывков. Большее количество употребления основ слов «хлеб*», «голод*», «смерть*», «карточк*» и «съе*». Кроме того, был составлен частотный тематический словарь, в который вошли 158 словосочетаний. Наиболее частотными оказались следующие: граммов хлеба, от голода, по карточкам, за хлебом и на хлеб [Крупнова, 2024].

Ещё одним возможным применением лингвистического ресурса является отбор лексических единиц для тематического словаря предметной области для системы фразеологического машинного перевода. Для начала необходимо создать корпус рефератов, например, собрать их с сайта базы данных ВИНТИ РАН путем парсинга, указав тематику и запрос. Затем загрузить файл в формате txt и провести анализ корпуса с помощью программы AntConc. В первую очередь отобрать самые частотные значимые существительные, затем провести анализ кластеров с этими словами, которые могут войти в словарь; анализ N-gram и употребление ключевых слов или фраз в контексте.

Корпусные менеджеры также можно использовать на уроках иностранного языка при изучении, например, грамматики и лексики. С их помощью можно познакомить учащихся с тенденциями употребления тех или иных конструкций, а также слов у письменной и устной речи. Кроме того, для сравнения можно взять тексты разных стилей и жанров.

Выводы. В последние несколько лет активно развивается область на стыке нескольких дисциплин, компьютерная лингвистика, в задачи которой входит обработка текстов на естественном языке. Для автоматизации данного процесса и получения быстрого и точного результата разрабатываются лингвистические ресурсы, которые анализируют язык на разных уровнях.

Корпусная лингвистика занимается созданием корпусных менеджеров, которые обрабатывают тексты и выполняют функции составления конкорданса, поиска ключевых слов, коллокаций и другие. В данной работе были рассмотрены

шесть наиболее популярных программы, среди которых выделяются WMatrix, AntConc, WordSmith Tools, GATE и Sketch Engine.

В ходе сравнительного анализа шести наиболее популярных программ были сформулированы выводы о том, что WordSmith Tools и AntConc предоставляют возможность работы с коллокациями, кластерами, ключевыми словами текстов, а также списками слов. Среди всех ресурсов WMatrix является единственной недоступной. В целом, в ряде программ есть ограниченный набор инструментов, отсутствие возможности сохранения данных, а также отсутствие интерфейса на русском языке.

В работе были также сформулированы некоторые возможные пути их использования в научной и практической деятельности. Среди актуальных направлений развития сетевых лингвистических ресурсов можно выделить корпусную стилистику; машинный перевод, компьютерную лексикографию, дискурс анализ и использование в образовательном пространстве.

Таким образом, лингвистические ресурсы в области автоматической обработки и корпусного анализа текста могут быть полезны при решении различных задач. Выбор подходящего инструмента зависит от объёма и глубины анализа текста.

Благодарности: исследование профинансировано Министерством науки и высшего образования РФ в рамках Программы стратегического академического лидерства «Приоритет-2030» (соглашение No 075-15-2024-201 от 6 февраля 2024 г.).

Литература

Большакова, Е. И., Клышинский, Э. С., Ландэ, Д. В., Носков, А. А., Пескова, О. В., Ягунова, Е. В. (2011) Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. Москва: МИЭМ.

Захаров, В. П. (2005) Корпусная лингвистика: учебно-методическое пособие. Санкт-Петербург: Изд-во СПбГУ.

Котурова, И. А. (2020) Корпусные исследования с помощью сервиса Antconc в условиях работы в вузе. *Язык и культура*, 52, 36–50. DOI: 10.17223/19996195/52/3

Кротова, Е. Б. (2019) Sketch Engine для лингвистических исследований. *Германистика сегодня: материалы Международной практической конференции, 16-17 октября 2018 г., Казань*. Казань: Изд-во Казан. ун-та. 107–112.

Крупнова, Е. С. (2024) Корпусный анализ темы «голод» во время блокады Ленинграда и составление частотного словаря. *Второй Международный молодёжный конкурс научных проектов «Стираем границы»: сборник материалов*. Москва: РГУ им. А. Н. Косыгина. 143–146.

Николаев, И. С., Митренина, О. В., Ландо, Т. М. (редакторы) (2016) Прикладная и компьютерная лингвистика: коллективная монография. 2-е изд. Москва: ЛЕЛАНД.

Палийчук, Д. А. (2022) Корпусные технологии в изучении колокаций (на примере сервисов «AntConc» и «SketchEngine»). *Studia Humanitatis*, 2, 13–14. URL: <https://cyberleninka.ru/article/n/korpusnye-tehnologii-v-izuchenii-kollokatsiy-na-primere-servisov-antconc-i-sketchengine>

AntConc: бесплатный набор инструментов для корпусного анализа, позволяющий конкорданировать и анализировать текст // Сайт Лоуренса Энтони: официальный сайт. URL: <https://www.laurenceanthony.net/software/antconc/>

Программа Sketch Engine // Sketch Engine: официальный сайт. URL: <https://www.sketchengine.eu/>

Рубайло, А. В., Косенко, М. Ю. (2016) Программные средства извлечения информации из текстов на естественном языке. *Альманах современной науки и образования*, 12 (114), 87–92.

Шамова, Н. А. (2021) Сравнительно-сопоставительный анализ корпусных инструментов (на примере работы с корпусами кинодискурса). *Вестник Нижегородского государственного лингвистического университета им. Н. А. Добролюбова*, 53, 82–95. DOI: 10.47388/2072-3490/lunn2021-53-1-82-95

Fischer-Starcke B. (2010) *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries*. London; New York: Continuum.

Leech, G., Short, M. (2007) *Style in Fiction: A Linguistic Introduction to English Fictional*.

London; New York: Longman. URL: <https://sv-etc.nl/styleinfiction.pdf>

Mahlberg M. (2012). *Corpus Stylistics and Dickens's Fiction*. New York: Routledge

Mahlberg, M. (2012). The corpus stylistic analysis of fiction – or the fiction of corpus stylistics? *Corpus Linguistics and Variation*

References

AntConc program: official website. Retrieved from <https://www.laurenceanthony.net/software/antconc/>

Bolshakova, E. I., Klyshinsky, E. S. (2011). *Automatic processing of texts in natural language and computational linguistics: textbook*. Moscow: Moscow Institute of Electronics and Mathematics Publ., 272. (In Russian).

Fischer-Starcke, B. (2010). *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries*. London; New York: Continuum.

Kotuyrova, I. A. (2020). Corpus-based studies with Antconc service at the university. *Language and Culture*, 52, 36–50, 10.17223/19996195/52/3. (In Russian).

Krotova, E. B. (2019). Sketch Engine for linguistic research. *Germanistics Today: Proceedings of the International Practical Conference*. Kazan: Kazan Publ. Kazan University, 107–112. (In Russian).

Krupnova E.S. (2024). Corpus analysis of the theme "hunger" during the blockade of Leningrad and the compilation of a frequency dictionary. *Second International Youth Competition of Scientific Projects "Erasing Borders": collection of materials*. Moscow: Kosygin Russian State University Publ., 143–146. (In Russian).

Leech, G., Short, M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional*. London; New York: Longman. Retrieved from: <https://sv-etc.nl/styleinfiction.pdf>

Mahlberg M. (2012). *Corpus Stylistics and Dickens's Fiction*. New York: Routledge.

Mahlberg, M. (2012). The corpus stylistic analysis of fiction – or the fiction of corpus stylistics? *Corpus Linguistics and Variation in English*, 75, 77–95, 10.1163/9789401207713_008.

McIntyre D. (2015). Towards an integrated corpus stylistics. *Topics in Linguistics*, 16(1), 59–69, 10.2478/topling-2015-0011. Retrieved from: <http://dx.doi.org/10.2478/topling-2015-0011>.

Nikolaev, I. S., Mitrenina, O. V., Lando, T. M. (2016). *Applied and computational linguistics*. Collective monograph. 2nd ed. Moscow: LELAND Publ. (In Russian).

Paliychuk, D. A. (2022). Corpus technologies in the study of collocations (by the example of "AntConc" and "SketchEngine" services). *Studia Humanitatis*, 2, 13–14. Retrieved from: <https://cyberleninka.ru/article/n/korpusnye-tehnologii-v-izuchenii-kollokatsiy-na-primere-servisov-antconc-i-sketchengine>. (In Russian).

Rayson, P. (2009). *Wmatrix: a Web-based Corpus Processing Environment*. Retrieved from: <http://ucrel.lancs.ac.uk/wmatrix/>.

Rubaylo, A. V., Kosenko, M. Yu. (2016). Program means of information extraction from natural language texts. *Almanac of Modern Science and Education*, 12(114), 87–92. (In Russian).

Shamova, N. A. (2021). Comparative-comparative analysis of corpus tools (on the example of work with film discourse corpora). *Bulletin of N.A. Dobrolyubov Nizhny Novgorod State Linguistic University*, 53, 82–95, 10.47388/2072-3490/lunn2021-53-1-82-95. (In Russian).

in *English*. Availability: Published, 75, 77–95.
DOI: 10.1163/9789401207713_008

McIntyre, D. (2015) Towards an integrated corpus stylistics. *Topics in Linguistics*, 16 (1), 59–69.
URL: <https://topling.ukf.sk/index.php/topling/article/view/22/22>.
DOI: <http://dx.doi.org/10.2478/topling-2015-0011>

Rayson, P. (2009) Wmatrix: a Web-based Corpus Processing Environment. *Computing Department, Lancaster University*.
URL: <http://ucrel.lancs.ac.uk/wmatrix/>

Wmatrix tutorial (for version 5). Documentation: Step-by-step instructions using a case study of linguistic analysis of political party manifestos for the UK General Election (updated November 2022).
URL: <https://ucrel.lancs.ac.uk/wmatrix/tutorial/>

Wmatrix tutorial (for version 6). Documentation: Step-by-step instructions on the example of linguistic analysis of political party manifestos for the UK General Election (updated in June 2023). URL: <https://ucrel.lancs.ac.uk/wmatrix/tutorial6/>

Introduction to WordSmith Tools. *WordSmith site*. URL: https://lexically.net/downloads/version_64_8/HTML/index.html

Sketch Engine program: official website. Retrieved from: <https://www.sketchengine.eu/>.

WMatrix 5. Documentation: Step-by-step instructions using a case study of linguistic analysis of political party manifestos for the UK General Election (updated November 2022). Retrieved from: <https://ucrel.lancs.ac.uk/wmatrix/tutorial/>.

WMatrix 6. Documentation: Step-by-step instructions on the example of linguistic analysis of political party manifestos for the UK General Election (updated in June 2023). Retrieved from: <https://ucrel.lancs.ac.uk/wmatrix/tutorial6/>.

WordSmith Tools. Retrieved from: https://lexically.net/downloads/version_64_8/HTML/index.html.

Zakharov, V. P. (2005). *Corpus linguistics: Manual*. Saint Petersburg: Saint Petersburg State University Publ. (In Russian).

Для цитирования статьи:

Дмитриев, А. В., Крупнова, Е. С. (2024). Сравнительно-сопоставительный анализ лингвистических ресурсов для проведения корпусного анализа текстов. *VERBA. Северо-Западный лингвистический журнал*, 3(13), 24–35. DOI: 10.34680/VERBA-2024-3(13)-24-35

For citation:

Dmitrijev, A. V., Krupnova, E. S. (2024). Comparative-Contrastive Analysis of Linguistic Resources for Corpus Analysis of Texts. *VERBA. North-West linguistic journal*, 3(13), 24–35. (In Russian). DOI: 10.34680/VERBA-2024-3(13)-24-35

ЯЗЫК ПОЭЗИИ СКВОЗЬ ПРИЗМУ БАЗЫ ДАННЫХ / POETRY LANGUAGE THROUGH DATABASE LENS

Преимущества использования баз данных в лингвистических исследованиях

А. А. Лебедев

Advantages of Using Databases in Linguistic Research

A. A. Lebedev

Александр Александрович Лебедев – кандидат филологических наук, доцент; Петрозаводский государственный университет, Петрозаводск, Российская Федерация

E-mail: perevodchik88@yandex.ru

Статья поступила: 15.09.2024. Принята к печати: 20.10.2024.

В статье рассматриваются особые возможности, которые предоставляют базы данных для лингвистических исследований. Анализируется роль и значимость баз данных на примере четырех конкретных проектов, которые были разработаны или находятся в процессе разработки при активном участии сотрудников кафедры русского языка Петрозаводского государственного университета. Представленные базы данных иллюстрируют, как междисциплинарный подход, включающий тесное взаимодействие лингвистики с информационными технологиями и другими научными дисциплинами, способен обогатить и упростить процесс исследования. В статье подчёркиваются преимущества такого подхода в контексте сбора, систематизации и анализа лингвистического материала, а также в использовании полученных результатов для дальнейших научных изысканий. Одним из ключевых аспектов обсуждения является тот факт, что базы данных позволяют хранить, обрабатывать и анализировать большие объёмы информации, что затруднительно при традиционных методах работы: базы данных упрощают доступ к большому спектру лингвистических данных, позволяют быстро и точно находить необходимые сведения, а также проводить многофакторные анализы, способно увеличить глубину исследований. Приводятся примеры дополнительных возможностей, таких как интеграция данных из разных источников, полуавтоматизированный анализ и возможность быстрого обновления информации, которые практически недоступны при использовании традиционных подходов работы с материалом. Делается вывод о том, что базы данных не только упрощают работу лингвиста, но и открывают новые горизонты для проведения междисциплинарных исследований, предоставляя

Aleksandr A. Lebedev – Candidate of Philological Sciences, Associate Professor; Petrozavodsk State University, Petrozavodsk, Russian Federation

ORCID: 0000-0001-9939-9389

Received: 15.09.2024. Accepted for publication: 20.10.2024.

The paper examines special opportunities that databases provide for linguistic research. The role and significance of databases are analyzed using four specific projects that have been developed or are in the process of development with the active participation of the staff of the Russian Language Department of Petrozavodsk State University. The presented databases illustrate how an interdisciplinary approach, including close interaction of linguistics with information technology and other scientific disciplines, can enrich and simplify the research process. The paper emphasizes the advantages of this approach in collecting, systematizing and analyzing linguistic material, as well as in using the results obtained for further scientific research. One of the key aspects of the discussion is the fact that databases allow storing, processing and analyzing large amounts of information, which is difficult with traditional methods of work: databases simplify access to a wide range of linguistic data, allow to quickly and accurately find the necessary information, as well as to conduct multivariate analyses, which can increase the depth of research. Examples of additional capabilities, such as integration of data from different sources, semi-automated analysis and the ability to quickly update information, which are practically unavailable when using traditional approaches to working with material, are given. It is concluded that databases not only simplify the work of a linguist, but also open up new horizons for interdisciplinary research, providing innovative tools for solving complex problems in the field of linguistics.

инновационные инструменты для решения сложных задач в области лингвистики.

Ключевые слова: база данных, поэтический синтаксис, диахроническая риторика, Феофан Прокопович

УДК 81`322:81`367:808

Keywords: database, poetic syntax, diachronic rhetoric, Feofan Prokopovich

OECD: 6.020Y

V

Постановка проблемы. Современная лингвистика активно использует компьютерные технологии для анализа языкового материала. Одним из ключевых инструментов, способных обеспечить высокую эффективность подобных исследований, являются базы данных. Их применение позволяет существенно расширить возможности

языкового анализа, автоматизировать и алгоритмизировать некоторые рутинные процессы обработки данных, а также обеспечить высокую точность обработки лингвистической информации. При этом преимущества подобного рода работы с материалом известны не всем лингвистам, что и предопределило задачу данной статьи – указать конкретные преимущества использования баз данных на материале тех программных продуктов, которые были разработаны при участии сотрудников кафедры русского языка Петрозаводского государственного университета.

Анализ любых языковых явлений требует обязательного системного подхода, который сложно реализовать без использования современных технологий. Многие современные лингвистические исследования включают в себя необходимость разносторонней обработки больших объемов данных, что значительно усложняет традиционные методы анализа. В связи с этим возникает необходимость детального рассмотрения преимуществ применения баз данных в лингвистических исследованиях на конкретном языковом материале.

История вопроса. В зарубежной лингвистике вопрос использования баз данных в языковых исследованиях был поставлен сравнительно давно (в частности, следует отметить работы [Linguisticdatabases, 1998; Theuseofdatabases, 2009]). В России одним из первых крупных специалистов в этом направлении стали А. С. Герд и А. Н. Баранов. Их работы, как научные [Баранов, 1998; Герд, 2002], так и учебные [Баранов, 2001; Герд, 2005] заложили основы современных представлений о прикладной и компьютерной лингвистике, которые воплощаются в том числе и в современных междисциплинарных исследованиях на стыке лингвистики и информационных технологий. Петрозаводский государственный университет также поддержал еще в начале XXI века эту тенденцию к использованию баз данных в филологических исследованиях; более подробно об этом – в [Захаров, 2006].

Зачастую исследователи описывают использование баз данных в отдельно взятых сферах лингвистической деятельности. Вопрос использования баз данных активно рассматривается в рамках такого направления, как корпусная лингвистика (к числу работ, системно описывающих данный аспект следует отнести [Потапова,

2012]); также подробно исследование аннотированных баз данных в работе корпусного лингвиста представлено в [Пахолкова, 2012].

Помимо корпусных исследований, базы данных играют важную роль в работе по созданию словарей. В статье Н. А. Мишанкиной «Базы данных в лингвистических исследованиях» [Мишанкина, 2013] описаны преимущества применения технологий баз данных в лексикографической практике, к числу которых отнесены эффективность и быстрота получения информации, возможность ее последовательного структурирования, а также перспективы использования полученных материалов как в исследовательской работе, так и в учебном процессе. Также конкретный пример реализации базы данных прагматически маркированной лексики русского языка представлен в статье [Булыгина, 2016].

Методология и методика анализа. При создании любых баз данных (в том числе и предназначенных для лингвистических исследований) применяется ряд принципов и установок, направленных на обеспечение эффективности таких программных продуктов и удобства их использования. Важнейшим принципом выступает систематизация данных: все исследуемые языковые единицы должны быть структурированы таким образом, чтобы обеспечить быстрый доступ к информации и гарантировать возможность выполнения сложных запросов. Достичь реализации данного принципа можно путем создания четкой и непротиворечивой классификации обрабатываемых данных (причем вне зависимости от уровня языка, будь то лексика, морфология, синтаксис и т.п.), а также разработки и последовательного описания метаданных для каждой такой вводимой в систему единицы.

Еще одним важным аспектом выступает обеспечение возможности масштабируемости разрабатываемой базы данных, что должно позволить легко расширить ее по мере поступления новых данных. Если предполагается, что база данных может быть включена в другие проекты, имеет смысл разработать требования к стандартизации данных, что позволяет интегрировать базы между разными проектами, а также обеспечивать их совместимость. Немаловажную роль играет и принцип автоматизации, позволяющий минимизировать вероятность человеческой ошибки в ходе ввода и обработки данных, а также существенно ускорить процесс анализа полученной информации.

Кафедра русского языка Петрозаводского государственного университета активно использует разные типы баз данных в рамках традиционных и современных лингвистических исследований поэтических и прозаических текстов. В работе мы рассмотрим наиболее востребованные из разработанных за последние годы баз данных, связанных с лингвистическими исследованиями, а также опишем те преимущества, которые исследователь получает при работе с такими базами данных в сравнении с традиционным представлением лингвистического материала.

Анализ материала. В ходе работы над «Синтаксическим словарем русской поэзии» в рамках реализации грантов РГНФ 15-04-00180 а – «Синтаксический словарь

языка русской поэзии XVIII – первой половины XIX века» и РФФИ 17-04-00168-ОГН – «Синтаксический словарь русской поэзии XIX века» (руководитель – Н. В. Патроева) с 2015 по 2019 год формировалась база данных синтаксических конструкций, методом сплошной выборки извлеченных из поэтического творчества авторов указанных временных периодов. Немаловажным преимуществом такой базы данных стал принципиально новаторский подход к работе с языковыми единицами. Более подробно методология создания словаря описана в статье [Патроева, 2015], а результатом обработки данных стали два печатных тома синтаксического словаря [Патроева, 2017] и [Патроева, 2019].

Совместный математико-литературоведческо-лингвистический коллектив в ходе работы над проектом РФФИ 18-012-90026 – «Проблема атрибуции анонимных и псевдонимных статей в журналах «Время», «Эпоха» и еженедельнике «Гражданин» (руководитель – А. А. Рогов) сформировал и наполнил базу данных «Размеченный корпус текстов XIX века в дореволюционной и современной графике для решения задач установления авторства» (Свидетельство №2021621353 от 22.06.2021), которая нашла воплощение в электронном ресурсе СМАЛТ (URL: <http://smalt.karelia.ru/>). Преимуществом такой реализации стала общедоступность полученных данных – наличие интернет-версии аналитического блока информационной системы позволяет пользователям работать с данными удаленно. Подробная информация о базе данных представлена в монографии проектного коллектива [Рогов, 2021].

В процессе работы над грантом РФФИ 22-28-00991 «Поэтический синтаксис русского языка XVIII века в риторическом аспекте» (руководитель Н. В. Патроева) была создана и зарегистрирована база данных «Риторический и медитативный вопрос как фигуры речи в лирике XVIII века» (Свидетельство №2023624528 от 11.12.2023). В эту базу данных включены синтаксические конструкции, представляющие собой риторические вопросы, которые были выбраны из поэтического творчества авторов XVIII века – А. Д. Кантемира, М. В. Ломоносова, А. П. Сумарокова, В. К. Тредиаковского.

На момент написания данной статьи научный коллектив под руководством Н. В. Патроевой в ходе работы над проектом «Диахроническая риторика: язык и слог стихотворных произведений Феофана Прокоповича в аспекте тропо- и фигурирования» создает базу данных средств речевой выразительности в творчестве Феофана Прокоповича. Предполагается, что сбор и систематизация лингвистических данных обусловит создание прочного фундамента для формирования русской диахронической риторики в дополнение к уже активно и давно развивающимся областям истории риторической традиции в России.

Использование баз данных оказывается оптимальным решением для многих современных лингвистических исследований по целому ряду причин. К примеру, базы данных позволяют эффективно хранить информацию и управлять большими ее объемами. Лингвистические исследования зачастую предполагают работу с большими объемами данных (тексты стихотворений, прозаические тексты,

словарные данные, особым образом размеченные тексты и т.п.). Традиционные методы хранения информации в текстовых файлах становятся неэффективными при работе с большими объемами данных, особенно в тех случаях, когда исследователь формирует сложные запросы. Базы данных обеспечивают эффективное управление этими данными, позволяя легко масштабировать системы и обрабатывать большие массивы информации за небольшое время. Такое хранение информации, к примеру, существенно упростило подготовку к печати материалов двух томов «Синтаксического словаря русской поэзии (объем первого тома составляет 36 печатных листов, объем второго тома – 38 печатных листов). В качестве примера табличного представления данных рассмотрим отрывок из базы данных «Синтаксического словаря русской поэзии» (Таблица 1). В таблице приняты следующие обозначения столбцов:

Столбец 1 — Репрезентации;

Столбец 2 — Паспортизация контекста (заглавие, страница);

Столбец 3 — Жанр;

Столбец 4 — Метр (размер);

Столбец 5 — Количество слов (знаменательных и незнаменательных);

Столбец 6 — Число строк, которые занимает предложение;

Столбец 7 — Тип композиции (тождественнострофическая – С / астрофическая (без графически выделенного разделения на строфы) – А / вольная композиция – ВК);

Столбец 8 — Тип строфы (для тождественнострофических произведений);

Столбец 9 — Наличие переносов из строки в строку (цифра указывает количество случаев несовпадения стихового и синтаксического членения, о чем сигнализирует отсутствие знака препинания в конце строки);

Столбец 10 — Риторические приемы;

Столбец 11 — Структурная схема;

Столбец 12 — Тип предложения с точки зрения полноты / неполноты состава;

Столбец 13 — Тип расширителя модели (однородного ряда);

Столбец 14 — Тип осложнителя модели;

Столбец 15 — Коммуникативный тип предложения (по целеустановке);

Столбец 16 — Тип предложения по эмоциональной окраске (интонации);

Столбец 17 — Тип предложения по общему модальному значению.

Таблица 1. Осложненное простое двусоставное предложение в поэзии М. В. Ломоносова (фрагмент базы данных). Используемые сокращения и обозначения расшифрованы в [Патроева, 2019, с. 30–38].

1	Европа, утомленна в брани, Из пламени подняв главу, К тебе свои простерла длани Сквозь дым, курение и мглу.	Там равной ревностью пылают Сердца, как стогны(,) все сияют В исполненной утех ночи.	Петрополь, небу подражая, Подобны испустил лучи.	Небесну пищу я вкушаю, На верх Олимпа вознесен!	Но каждую обозревши часть, С веселием сие вещает: [ПР]	Во храм, сияющий металлом, Пред трон, украшенный кристаллом, Поспешно простирает ход; Венцем зеленым уязенной И в виск, вещает, облеченной Владычице российских вод.
2	18, 123	18, 126	18, 126	19, 127	19, 128	19, 128
3	ода похвальная	ода похвальная	ода похвальная	ода похвальная	ода похвальная	ода похвальная
4	Я4	Я4	Я4	Я4	Я4	Я4
5	18	13	6	8	8	22
6	4	3	1	2	2	6
7	С	С	С	С	С	С
8	10-стишие	10-стишие	10-стишие	10-стишие	10-стишие	10-стишие
9	1	2				2
10	инв., олиц.	инв., мт., син.	инв., мт.	инв., рит.вос., мт.	инв., эл.	инв., эл., прф., мт.
11	[сущ.+гл.].	[гл.+сущ.+гл.].	[сущ.+гл.].	[мест.+гл.].	[<сущ.>+гл.]:	[<сущ.>+гл.+гл.].
12	пол	пол	пол	пол	неп	неп
13	о/об					о/ск, о/об
14	пр/об (подл.вн.об.), д/об	срв/об	д/об	пр/об	д/об	пр/об
15	пов	пов	пов	пов	пов	пов
16	Невос	Невос	Невос	Вос	Невос	Невос
17	утв	утв	утв	утв	утв	утв

Использование особых языков запросов позволяет исследователям быстро и точно извлекать необходимую информацию из обширных корпусов данных. К примеру, можно получить сведения о статистике частотности слов в определенной группе текстов, извлечь контексты использования определенных лексических единиц или построить сложные запросы для анализа морфологических и синтаксических структур. Эта гибкость значительно повышает продуктивность исследований, существенно сокращая время, необходимое на предварительную обработку данных. Подобные инструменты, к примеру, встроены в базу данных СМАЛТ, и у пользователя есть возможность формировать группы текстов и применять к ним различные статистические методы для исследования (с этой функцией можно ознакомиться

непосредственно в интернет-версии базы данных, доступной по ссылке URL: <http://smalt.karelia.ru/shower/>, блок «Исследования»).

К примеру, в основе метрики Г. Хетсо, используемой для анализа каждого представленного в базе данных «СМАЛТ» текста, учитывается 15 параметров (подробнее изложено в [Хетсо, 1986]):

1. Общее распределение частей речи в первых двух и в последних трех позициях предложения.
2. Распределение частей речи в первой позиции предложения.
3. Распределение частей речи во второй позиции предложения.
4. Сочетание частей речи в первых двух позициях предложения.
5. Распределение частей речи в третьей с конца позиции предложения.
6. Распределение частей речи в предпоследней позиции предложения.
7. Распределение частей речи в последней позиции предложения.
8. Сочетание частей речи в последних трех позициях предложения.
9. Средняя длина слова в буквах, вычисляемая на основании выборок размером в 500 текстовых слов.
10. Общее распределение длины слова.
11. Средняя длина предложения в словах, вычисляемая на основании выборок размером в 30 предложений.
12. Общее распределение длины предложения.
13. Лексический спектр текста на уровне словаря.
14. Лексический спектр текста на уровне текста.
15. Индекс разнообразия лексики.

Ручной подсчет каждого из этих параметров чрезвычайно затратен по времени, в то время как автоматическая обработка текстов позволяет получить нужные данные за небольшой промежуток времени, минимизируя вероятность возникновения ошибок, связанных с подсчетами количественных характеристик (примеры подобных автоматических формируемых данных представлены в Таблицах 2 и 3). Подобные статистические данные могут быть использованы в решении вопросов, связанных с определением авторства текстов, чему посвящена монография [Рогов, 2021].

Таблица 2. Общее распределение частей речи в первых двух и последних трех позициях предложения в тексте «Ряд статей о русской литературе. Введение» Ф. М. Достоевского.

Глава	1	2	3	4	5
0: Существительное	57	83	94	76	379
1: Прилагательное	15	43	44	96	39
2: Числительное	9	5	1	9	1
3: Местоимение	176	133	128	163	65
4: Глагол	57	125	101	63	112
5: Причастие	1	5	11	6	6
6: Деепричастие	2	2	1	0	1
7: Наречие	37	58	57	39	35
8: Категория состояния	5	7	1	5	3
9: Частица	93	95	41	55	6
10: Предлог	58	57	115	86	0
11: Союз	108	35	56	47	1
12: Модальное слово	18	3	2	3	0
13: Междометие	2	0	0	0	0
15: Иностранное слово	0	0	3	4	4
16: Цитата	4	4	5	5	5
17: Вводное слово	26	14	5	1	0
20: Неязыковой символ	0	0	0	1	1
21: Сокращенное слово	1	0	3	7	8
22: Часть многочленного названия	1	1	2	4	4

Таблица 3. Общее распределение длины слова в тексте «Ряд статей о русской литературе. Введение» Ф. М. Достоевского

Длина слова (в буквах)	Количество
1	944
2	1853
3	1226
4	1612
5	1036
6	1496
7	1101
8	951
9	721
10	572
11	409
12	261
13	152
14	78
15	50
16	28
17	9
18	2
19	4
20	1
21	1
22	1

Некоторые лингвистические задачи подразумевают работу с иерархически организованными данными или сложными сетями связей. Современные базы данных обеспечивают эффективное представление и управление такими структурами. Это позволяет исследователям моделировать сложные лингвистические феномены с учетом всех межуровневых связей и зависимостей. В качестве иллюстрации можно рассмотреть структуру базы данных «Риторический и медитативный вопрос как фигуры речи в лирике XVIII века» (Рис. 1).

Многотабличная структура базы данных позволяет формировать сложные запросы, включающие в себя несколько параметров (например, можно получить список произведений, в которых встречается два и более риторических вопроса, выраженных неосложненными простыми двусоставными предложениями или список риторических вопросов определенного автора, которые имеют нужную синтаксическую структуру, причем только из тех стихотворений, которые принадлежит определенному жанру – см. пример в Таблице 4).

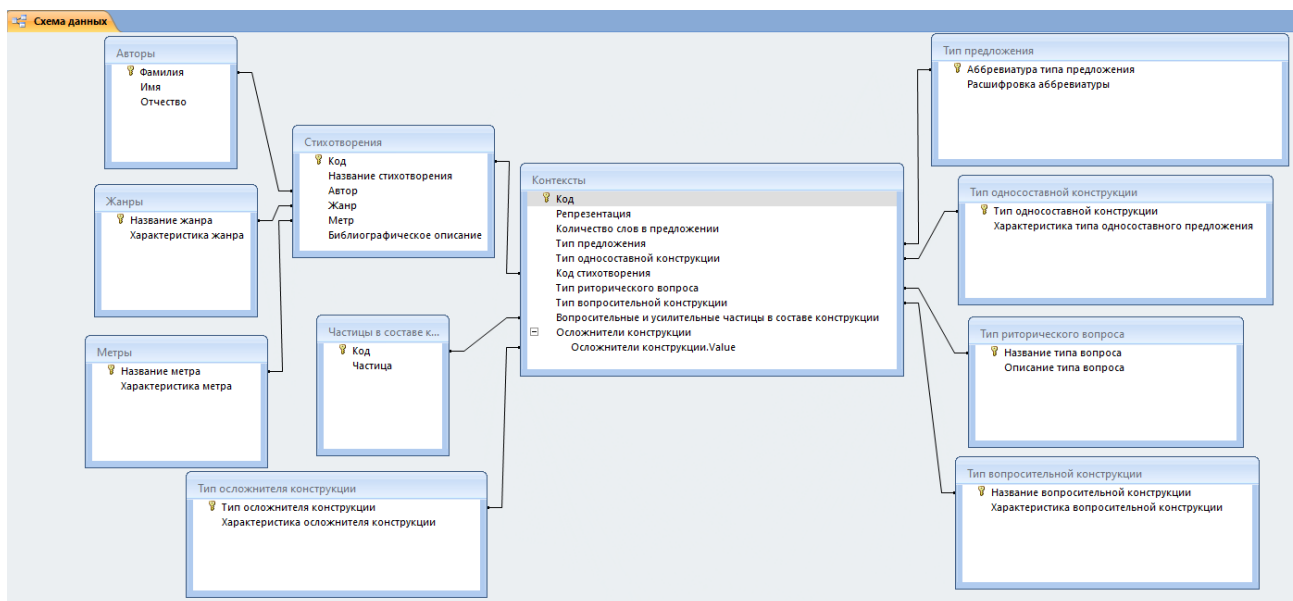


Рис. 1. Структура базы данных «Риторический и медитативный вопрос как фигуры речи в лирике XVIII века»

Таблица 4. Выборка из базы данных «Риторический и медитативный вопрос как фигуры речи в лирике XVIII века», в которой представлены риторические и медитативные вопросы, присутствующие в баснях А. П. Сумарокова и выраженные многокомпонентными сложными предложениями.

Репрезентация	Автор	Аббревиатура типа предложения	Название жанра
От водки голова болит, Но водка сердце веселит, Молошное питье не диво, Его хмельняй и пиво; Какое ж им питье и пить, Коль водки не купить? [Сумароков, 1957, с. 221]	Сумароков	МСП	Басня
О мальчик! узнавать ты был людей искусен, Но знаешь ли теперь, что ты парнасский рак? [Сумароков, 1957, с. 246]	Сумароков	МСП	Басня
Так я тебе скажу об этом поученье: О чем ты сетуешь напрасно, человек, Что твой недолог век И скоро наших тел со духом разлученье? [Сумароков, 1957, с. 222]	Сумароков	МСП	Басня
Не будет никогда чертями там вонять, То правда, и стихи такие пахнут худо, Однако запах сей и истреблять не чудо, Почаще надобно курить, А черт от курева престанет ли дурить? [Сумароков, 1957, с. 237]	Сумароков	МСП	Басня

Использование баз данных способствует автоматизации многих рутинных процессов, что особенно важно при проведении масштабных исследований. Например, можно частично автоматизировать процесс обработки морфологических данных, предлагая в ходе заполнения базы данных наиболее вероятный вариант разбора, что также было реализовано в системе СМАЛТ. Это не только снижает трудозатраты, но и обеспечивает воспроизводимость результатов при всех последующих разборах, что является важным критерием научной достоверности. Возможность хранить и повторно использовать сложные запросы и процедуры в базах данных делает процесс исследования более прозрачным и систематичным. Помимо этого, возможность извлекать статистическую информацию из обработанных данных позволяет получать сведения, которые могут быть использованы для сопоставительного анализа. Примером таких сведений могут стать полученные сравнительные данные об активности риторических приемов в поэзии А. Д. Кантемира, В. К. Третьяковского, М. В. Ломоносова и А. П. Сумарокова (Таблица 5).

Таблица 5. Сравнительные данные, свидетельствующие об активности риторических приемов в поэзии А. Д. Кантемира, В. К. Третьяковского, М. В. Ломоносова и А. П. Сумарокова.

Троп или фигура речи	Количество репрезентаций в стихотворениях				Средняя активность риторического приема на 1 предложение / 1 текст автора			
	Кантемира	Третьяковского	Ломоносова	Сумарокова	Кантемира	Третьяковского	Ломоносова	Сумарокова
Лексический повтор	158	83	276	639	0,2 / 3,6	0,14 / 2,41	0,13 / 2,34	0,21 / 2,49
Метафора	77	141	1275	1025	0,1 / 1,8	0,24 / 4,15	0,6 / 10,81	0,34 / 3,99
Метонимия	39	49	384	207	0,05 / 0,89	0,08 / 1,44	0,18 / 3,25	0,07 / 0,81
Олицетворение	25	36	252	155	0,03 / 0,57	0,06 / 1,06	0,12 / 2,14	0,05 / 0,6
Перифраза	14	14	81	82	0,02 / 0,32	0,02 / 0,41	0,04 / 0,69	0,03 / 0,32
Сравнение	35	21	116	99	0,04 / 0,8	0,04 / 0,62	0,01 / 0,98	0,03 / 0,39
Риторическое восклицание	30	121	267	325	0,04 / 0,68	0,2 / 3,56	0,12 / 2,26	0,11 / 1,26
Риторический вопрос	99	106	227	264	0,12 / 2,25	0,18 / 3,12	0,1 / 1,9	0,09 / 1,03
Риторическое обращение	34	87	304	316	0,04 / 0,79	0,15 / 2,56	0,14 / 2,58	0,1 / 1,23
Гипербола	5	42	28	41	0,01 / 0,11	0,07 / 1,24	0,01 / 0,24	0,01 / 0,16

Еще одно важное преимущество состоит в том, что базы данных позволяют интегрировать лингвистические данные с данными из других дисциплин. Это открывает новые возможности для различных междисциплинарных исследований, где анализ лингвистических данных может быть дополнен другими факторами. В частности, предполагается, что разрабатываемая на данный момент база данных средств речевой выразительности в творчестве Феофана Прокоповича поможет внести вклад в изучение диахронического синтаксиса, исторической лексикологии, лингвопоэтики и стилистики, а также стать основой для лексикографического представления языка Феофана Прокоповича (Таблица 6).

Таблица 6. Фрагмент базы данных «Средства речевой выразительности в творчестве Феофана Прокоповича» (база данных на момент написания статьи находится в процессе разработки)

Автор	Феофан Прокопович
Произведение, год создания	«К Петру Второму», 1728 (?)
Жанр	надпись
Метрика	силлабический 13-сложник
Композиция, строфика	4-стишие
Контекст, выявляющий тропообразование	<i>Бог и Петру Второму вручил стада многа...</i> [Прокопович, 1961, с. 216]
Тип тропа или фигуры	глагольная метафора, узуальная; «фигура смысловая», способствующая «услаждению речи» [Прокопович, 2020, с. 251-259].
Данные словарей русского языка XVIII в.	«Вручаю<...> 1) Отдаю кого или что кому; ввѣрю чьему попеченію, смотрѣнію. <...> 2) Въ руки отдаю...» [Словарь Академии Российской. – СПб.: При Имп. Акад. наук, 1789–1794. Ч. 5. От Р до Т. 1794.] «ВРУЧИТЬ (-ти) <...>. 1. что. Передать из рук в руки, отдать лично кому-л. 2. что. Вверить управление, распоряжение чем-л., поручить. <...> Перен. О духовной пастве» [Словарь русского языка XVIII века / АН СССР. Ин-т рус. яз.; Гл. ред.: Ю. С. Сорокин. Л.: Наука. Ленингр. отд-ние, 1984 (продолж. изд.). Вып. 4. (Воздух — Выпись). Л.: Наука. Ленингр. отд-ние, 1988. 256 с.]

Выводы. Таким образом, следует отметить, что именно использование баз данных в лингвистических исследованиях открывает широкие перспективы для повышения эффективности и качества научных исследований. Они обеспечивают удобное и масштабируемое хранение данных, мощные инструменты для их анализа, поддерживают сложные структуры и связи, способствуют автоматизации и воспроизводимости, а также поддерживают междисциплинарный подход. В будущем, с развитием технологий, базы данных будут играть все более важную роль в лингвистике, открывая новые горизонты для научных исследований.

К числу наиболее значимых преимуществ, которые предлагают базы данных, следует отнести возможность эффективно работать с большими объемами данных, извлекать нужную информацию по запросу пользователя, моделировать сложные многоуровневые лингвистические структуры, автоматизировать некоторые процессы, связанные с обработкой данных, а также интегрировать полученные сведения в рамках междисциплинарных исследований, упрощая совместную работу научных коллективов. Все перечисленные выше возможности нашли свое воплощение в конкретных примерах реализованных баз данных.

Благодарности: исследование выполнено за счет гранта Российского научного фонда № 24-28-00696, <https://rscf.ru/project/24-28-00696/>

Литература

References

- Баранов, А. Н. (1998). Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема. *Русистика сегодня: язык: система и ее функционирование: сборник статей*. Москва: Наука. 1–2, 179–191.
- Булыгина, Е. Ю., Трипольская, Т. А. (2016). База данных прагматически маркированной лексики русского языка: материал, принципы описания, возможности использования. *Вестник Новосибирского государственного педагогического университета*, 6(34), 70–85. DOI: 10.15293/2226-3365.1606.06.
- Герд, А. С. (2002). Базы данных и прикладная лингвистика. *Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных», 5–7 марта 2002 г., Санкт-Петербург*. Санкт-Петербург: Изд-во СПбГУ. 3–6.
- Герд, А. С. (2005). Прикладная лингвистика. Санкт-Петербург, Изд-во Санкт-Петербургского университета.
- Захаров, Н. В. (2006). Филологические науки и информационные технологии. *Знание. Понимание. Умение: фундаментальные и прикладные исследования в области гуманитарных наук*, 2, 166–172.
- Мишанкина, Н. А. (2013). Базы данных в лингвистических исследованиях. *Вопросы лексикографии*, 1(3), 25–33.
- Патроева, Н. В., Лебедев, А. А. (2015). Проект синтаксического словаря языка русской поэзии XVIII – первой половины XIX века. *Ученые записки Петрозаводского государственного университета. Серия. Общественные гуманитарные науки*, 3(148), 53–55.
- Пахолкова, И. А. (2012). Применение методов корпусной лингвистики в традиционном языкознании. *Вестник Московского государственного лингвистического университета*. 646. 125–135.
- Потапова, Р. К. (2012). Дискурсивная составляющая современной корпусной лингвистики (применительно к устно-речевым базам данных). *Вестник Московского государственного лингвистического университета*. 639. 157–167.
- Прокопович, Ф. (1961) *Сочинения / под редакцией И. П. Еремина*. Москва; Ленинград: Издательство Академии наук СССР.
- Прокопович, Ф. (2020) *Об искусстве риторическом десять книг / перевод Г. А. Стратановского; ответственный редактор С. И. Николаев*. Москва; Санкт-Петербург: Альянс-Архео.
- Рогов, А. А., Абрамов, Р. В., Бучнева, Д. Д., Захарова, О. В., Кулаков, К. А., Лебедев, А. А., Москин, Н. Д., Отливанчик, А. В., Савинов, Е. Д., Сидоров, Ю. В. (2021). *Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин»: коллективная монография*. Петрозаводск: Острова.
- Патроева, Н. В. (редактор) (2017). *Синтаксический словарь русской поэзии XVIII века: в 4 т. Т. 1: Кантемир, Тредиаковский*. Санкт-Петербург: Дмитрий Буланин.
- Патроева, Н. В. (редактор) (2019). *Синтаксический*
- Baranov, A. N. (1998). Automation of linguistic research: a corpus of texts as a linguistic problem. *Rusistika segodnya [Russistics today: language: system and its functioning: collection of papers]*, 1–2, 179–191. (In Russian).
- Bulygina, E. Yu., Tripolskaya, T. A. (2016). Database of pragmatically marked lexical items of the Russian language: the content, principles of description, and possibilities of using. *Vestnik Novosibirskogo gosudarstvennogo pedagogicheskogo universiteta*, 6(34), 70–85, 10.15293/2226-3365.1606.06. (In Russian).
- Gerd, A. S. (2002). Databases and applied linguistics. *Proceedings of the Scientific Conference "Corpus Linguistics and Linguistic Databases", March 5–7, 2002, St. Petersburg*. St. Petersburg: St. Petersburg State University Publ., 3–6. (In Russian).
- Gerd, A. S. (2005). *Applied linguistics*. St. Petersburg: St. Petersburg State University Publ. (In Russian).
- Khetso, G. (1986). *Attribution to Dostoevsky: On the Question of the Attribution to F. M. Dostoevsky of Anonymous Articles in the Magazines "Vremya" and "Epokha"*. Oslo: Solum Forlag A. S. 82 (In Russian).
- Linguistic databases (1998). Edited by John Nerbonne. Stanford, Calif.: CSLI Publications.
- Mishankina, N. A. (2013). Databases in linguistic studies. *Voprosy leksikografii*, 2013, 1(3), 25–33. (In Russian).
- Patroeva, N. V. (ed.) (2017). *Syntactic Dictionary of Russian Poetry of the 18th Century: in 4 volumes. Vol. 1: Kantemir, Trediakovskiy*. Saint Petersburg: Dmitrii Bulanin Publ. (In Russian).
- Patroeva, N. V. (ed.) (2019). *Syntactic Dictionary of Russian Poetry of the 18th Century: in 4 volumes. Vol. 2: Lomonosov*. Saint Petersburg: Dmitrii Bulanin Publ. (In Russian).
- Patroeva, N. V., Lebedev, A. A. (2015). Syntactic dictionary of Russian poetry of XVIII — first half of XIX century. *Proceedings of Petrozavodsk state university. Social sciences & Humanities*, 3(148), 53–55. (In Russian).
- Pakholkova, I. A. (2012). Application of the methods of corpus linguistics in traditional linguistics. *Vestnik of Moscow State Linguistic University*, 646, 125–135. (In Russian).
- Potapova, R. K. (2012). Discourse component of modern corpus linguistics (as applied to oral speech databases). *Vestnik of Moscow State Linguistic University*, 639, 157–167. (In Russian).
- Prokopovich F. (1961). *Works*. Edited by I. P. Eremin. Moscow; Leningrad: USSR Academy of Sciences Publ. (In Russian).
- Prokopovich F. (2020). *Ten books on the art of rhetoric*. Translated by G.A. Stratanovsky; executive editor S.I. Nikolaev. Moscow; Sankt-Peterburg: Al'yans-Arkheo Publ. (In Russian).
- Rogov, A. A., Abramov, R. V., Buchneva, D. D., Zakharova O. V., Kulakov, K. A., Lebedev, A. A., Moskin, N. D., Otlivanchik, A. V., Savinov, E. D., Sidorov, Iu. V. (2021). *The*

словарь русской поэзии XVIII века: в 4 т. Т. 2: Ломоносов. Дмитрий Буланин.

Сумароков, А. П. (1957). Избранные произведения. Ленинград: Советский писатель.

Хетсо, Г. (1986). Принадлежность Достоевскому: к вопросу об атрибуции Ф. М. Достоевскому анонимных статей в журналах *Время* и *Эпоха*. Oslo: Solum Forlag A. S.

Nerbonne, J. (edited) (1998). *Linguistic databases*. Stanford (California): CSLI Publications.

Everaert, M., Musgrave, S., Dimitriadis, A. (edited) (2009). *The use of databases in cross-linguistic studies*. Berlin; New York: Mouton de Gruyter.

Problem of Attribution in the Magazines "Vremya", "Epokha" and the weekly "Grazhdanin". Petrozavodsk: Ostrova Publ. (In Russian).

Sumarokov, A. P. (1957). *Selected Works*. Leningrad: Sovetskiy pisatel' Publ., 608. (In Russian).

The use of databases in cross-linguistic studies (2009). Edited by Martin Everaert, Simon Musgrave, Alexis Dimitriadis. Berlin; New York: Mouton de Gruyter Publ.

Zakharov, N. V. (2006). Philological sciences and information technologies. *Knowledge. Understanding. Skill*. 2006, 2, 166–172. (In Russian).

Для цитирования статьи:

Лебедев, А. А. (2024). Преимущества использования баз данных в лингвистических исследованиях. *VERBA. Северо-Западный лингвистический журнал*, 3(13), 36–49. DOI: 10.34680/VERBA-2024-3(13)-36-49

For citation:

Lebedev, A. A. (2024). Advantages of Using Databases in Linguistic Research. *VERBA. North-West linguistic journal*, 3(13), 36–49. (In Russian). DOI: 10.34680/VERBA-2024-3(13)-36-49

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ / ARTIFICIAL INTELLIGENCE

Язык промптов, или особенности формулирования запросов к генеративным нейросетям для создания изображений

Л. Г. Алексеева, П. С. Алексеев

Prompt Language, or Features of Formulation of Queries to Generative Neural Networks for Image Creation

L. G. Alexeeva, P. S. Alexeev

Любовь Геннадьевна Алексеева – аспирант; Новгородский государственный университет имени Ярослава Мудрого, Великий Новгород, Российская Федерация

E-mail: lubov.g.alex@mail.ru

Петр Сергеевич Алексеев – научный сотрудник кафедры журналистики; Новгородский государственный университет имени Ярослава Мудрого, Великий Новгород, Российская Федерация

E-mail: alexeev.dauwalter@gmail.com

Статья поступила: 05.10.2024. Принята к печати: 20.10.2024.

В статье рассматривается проблема формулирования грамотного запроса (промпта) к генеративным нейросетям, позволяющим создать изображение. Данная тема в настоящее время является очень актуальной, т.к. благодаря быстрому развитию технологического прогресса системы искусственного интеллекта внедряются практически во все сферы жизни человека. Многие пользователи не знают, как правильно сделать запрос к генеративной нейросети, поэтому результат часто получается неточный или вообще неправильный. В ходе исследования был проведен эксперимент, в котором приняли участие 67 студентов Политехнического колледжа НовГУ, обучающихся по специальности «Информационные системы и программирование». Студентам было предложено ответить на вопрос: «Как бы вы задали вопрос для Stable Diffusion для получения портрета молодого программиста за столом в современном офисе с панорамными окнами, работающим за ноутбуком. На столе находится кружка кофе». Результаты эксперимента показали, какие ошибки пользователи наиболее часто совершают при написании промптов, как их избежать, а также помогли ввести новый термин – «язык промптов» и описать его главные черты. Нами было отмечено, что для получения желаемого результата при формулировании промпта пользователю необходимо использовать английский язык, так как современные модели обучаются на англоязычных базах данных. Помимо этого, необходимо иногда игнорировать правила грамматики и синтаксиса, так как это может усложнять задачу генеративной нейросети. Также, при создании

Liubov G. Alexeeva – postgraduate student; Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, Russian Federation

ORCID: 0009-0005-2584-7819

Peter S. Alexeev – research fellow, Department of Journalism; Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, Russian Federation

ORCID: 0009-0007-3793-8309

Received: 05.10.2024. Accepted for publication: 20.10.2024.

The article deals with the problem of formulating a competent request (prompt) to generative neural networks that allow to create an image. This topic is currently very relevant, because due to the rapid development of technological progress, artificial intelligence systems are being implemented in almost all areas of human life. Many users do not know how to properly make a request to a generative neural network, so the result is often inaccurate or even incorrect. In the course of the research, an experiment was conducted, in which 67 students of NovSU Polytechnic College, majoring in Information Systems and Programming, took part. The students were asked to answer the following question: *How would you make a request for Stable Diffusion to create a portrait of a young programmer at a desk in a modern office with panoramic windows, working at a laptop? Also there is a mug of coffee on the desk.* The results of the experiment have shown what mistakes users most often make when writing prompts, how to avoid them, and also helped to introduce a new term – *prompt language* and describe its main features. It has been found that in order to get the desired result when formulating a prompt the user should use the English language, since modern models are trained on English language databases. In addition, it is necessary to sometimes ignore the rules of grammar and syntax, as this can complicate the task of the generative neural network. Also, when creating non-existent objects or hybrids, in a query, you can use square brackets and numbers to mix keywords in a certain ratio, which is useful for creating unique or fantasy images.

несуществующих объектов или гибридов в запросе можно использовать квадратные скобки и числа, для смешения ключевых слов в определенном соотношении, что полезно при создании уникальных или фантастических изображений.

Ключевые слова: генеративные нейронные сети, нейронные сети, нейросетевая модель, генерация изображений, искусственный интеллект, промпт-инжиниринг, язык промптов, промпт

УДК 004.032.26:004.8

Keywords: generative neural networks, neural networks, neural network model, image generation, artificial intelligence, prompt engineering, prompt language, prompt

OECD: 1.02EP

V

Постановка проблемы. Научно-технический прогресс развивается с огромной скоростью и, пожалуй, в наше время невозможно отыскать человека, который ничего не слышал об искусственном интеллекте (ИИ). ИИ используется людьми как для решения профессиональных задач (например, для автоматизации процессов на производстве, создания беспилотного транспорта, помощи в диагностике различных заболеваний на основе анализов пациента и др.), так и для личных целей. «По данным на 1 мая 2024 г., аудитория только известного чат-бота ChatGPT достигла 180 млн человек, первые 100 млн подписчиков платформа получила через два месяца после запуска. Для сравнения: Youtube потребовалось 1,5 года, чтобы набрать аудиторию в 100 млн человек» [Поспелова, 2024].

Одним из частых запросов к нейросети является желание пользователя сгенерировать изображение. Например, иногда к этому прибегают художники и дизайнеры, чтобы ускорить процесс своей работы либо вдохновиться. ИИ позволяет экспериментировать с различными стилями, цветами и композициями, помогая создавать уникальные произведения искусства. Люди, не занимающиеся рисованием или дизайном профессионально, напротив могут использовать подобные модели для своих нужд, так как не могут или не умеют самостоятельно создавать необходимые изображения. Генеративные нейросети, помогающие создать картинки, могут использоваться при обучении студентов как творческих, так и технических специальностей. Ну и, конечно, данными технологиями можно пользоваться просто для веселья и развлечения. Самыми популярными генеративными моделями на сегодняшний день можно назвать Stable Diffusion, Midjourney и DALL-E. Нередко, пытаясь создать картинку, пользователи сталкиваются с проблемой: как бы, на их взгляд, конкретно, четко и ясно они ни выражали свои требования, ИИ не выдает им желаемый результат. Иначе говоря, они сталкиваются с проблемой грамотного промпт-инжиниринга.

Промпт – это запрос, с помощью которого ставится задача языковой модели [Глазова, 2024]. Промпт-инжиниринг – это разработка правильных «инструкций» для генеративных нейросетей, позволяющих лучше понять запросы пользователей. Проблемы, связанные с промпт-инжинирингом, возникают по нескольким причинам.

Даже самые популярные и часто используемые модели обучаются на определенных наборах данных, соответственно, их «база информации» может быть ограничена, поэтому и появляются неточности в результате.

Среднестатистический пользователь не всегда понимает особенности и возможности той или иной модели, иногда его, казалось бы, простой запрос может быть слишком сложным и даже технически невыполнимым. Следует отметить, что порой на один и тот же промпт нейросети могут выдавать разный результат в зависимости от случайных факторов или изменений в модели.

В естественных языках многие слова имеют несколько значений, что может привести к неправильной интерпретации промпта и, следовательно, неожиданному результату. Также к нерелевантным ответам может привести то, что модели не всегда учитывают контекст сообщения, а грамматика и синтаксис для нейросети не настолько важна для понимания предложения, как для человека.

В контексте лингвистики нас интересует последний момент. Учитывая связанные с этим проблемы, можно сказать, что для эффективной коммуникации человека с нейросетью необходимо буквально овладеть новым языком, языком промптов. В данной статье ставится проблема обозначить его главные черты, для чего необходимо рассмотреть частые ошибки пользователей, обнаруженные в ходе эксперимента, в котором участвовали студенты 2–4 курса Политехнического колледжа НовГУ специальности «Информационные системы и программирование».

История вопроса. Для начала рассмотрим определения того, что же такое искусственный интеллект и, в частности, генеративные нейросети, а также изучим их историю.

В 1980-е годы ученые Барр и Файгенбаум ввели понятие искусственного интеллекта, которое с тех пор практически не изменилось. Они обозначили ИИ как область информатики, занимающуюся созданием интеллектуальных компьютерных систем. Эти системы выделяются в отдельную категорию, поскольку они способны выполнять функции, аналогичные человеческим, таким как распознавание речи и иностранного языка, логическое мышление и обучение [Вознюк, 2019]. Некоторые ученые считают, что более правильно считать основателями концепции искусственного интеллекта Марвина Мински и Джона Маккарти, которые в 1950-х годах выдвинули идею о том, что «проблема ИИ обычно заключается в необходимости заставить машину действовать подобно разумному человеку» [McCarthy, 1955].

В 1950 британский математик, логик и криптограф Алан Тьюринг выпустил статью «Вычислительные машины и разум», в которой размышлял на тему того, может ли машина мыслить, а также предложил идею теста, ныне известного как «тест Тьюринга» – стандарта для проверки систем искусственного интеллекта [Turing, 1950].

Свою гипотезу относительно искусственного интеллекта выдвинули Аллен Ньюэлл и Герберт Саймон. Они предположили, что «физическая символьная система имеет необходимые и достаточные средства для произведения основных интеллектуальных операций» [Simon, 1970], то есть способности выполнять символьные вычисления достаточно для совершения осмысленных действий. Хоть данная идея и спорна, но стоит отметить, что большая часть разработчиков, работающих над созданием систем ИИ «пошла по пути создания именно символьных систем» [Миндигулова, 2023].

Е. Г. Барщевский выделяет три вида искусственного интеллекта – общий ИИ, ограниченный ИИ и искусственный сверхинтеллект [Барщевский, 2023]. Ограниченный

ИИ специализируется на конкретной задаче, полученной с помощью текстового запроса – промпта. То есть ранее упомянутые нами генеративные нейросети относятся к ограниченному искусственному интеллекту.

Генеративные нейросети можно классифицировать в зависимости от их функции: генерация текста (CharacterAI, Bing AI), создание изображений (Stable Diffusion, Midjourney, Dall-E, Kandinsky, Шедеврум), монтаж видео (Runway Gen-2, Nvidia Eye Contact), работа со звуком (Adobe Enhance, Voice.AI) и т. д. [Малышев, 2024].

Популярность генеративных нейросетей растет с каждым днем. Например, Midjourney, помогающая, как и Stable Diffusion и DALL-E, создавать изображения, с большой скоростью и высоким качеством способна выполнять функции дизайнера, создавая «заготовки», которые позже человек будет обрабатывать под потребности клиента [Соколова, 2024]. Несомненно, это очень выгодно экономически, поэтому очевидно, что со временем интерес к нейросетям будет только расти.

Мы уже упомянули ранее, что для создания требуемого изображения пользователю необходимо правильно сформулировать свой запрос – промпт. Написать правильный промпт, который поможет получить желанный результат, не так уж просто. Одной из распространенных проблем является то, что многие пользователи не понимают механизм работы генеративных нейросетей. Человек часто может написать очевидный для него запрос, не понимая, что нейросеть обучена на ограниченном объеме данных, она не всегда понимает контекст ситуации, абстрактные понятия, не знает какую-либо узконаправленную или специализированную информацию, а также порой чисто технически не может отобразить определенные текстуры или стили. Зная, как правильно сформулировать свой запрос к генеративной нейросети, шанс получить желаемый результат значительно возрастает.

Методология и методика исследования. Методологически настоящая работа построена на лингвистическом эксперименте, который был проведен среди 67 студентов 2-4 курса Политехнического колледжа НовГУ, обучающихся по специальности «Информационные системы и программирование», с целью выявить их подходы к написанию текстовых запросов для генеративных моделей. В рамках эксперимента учащимся был задан вопрос: «Как бы вы сформулировали запрос для Stable Diffusion, чтобы получить портрет молодого программиста, работающего на ноутбуке за столом в современном офисе с панорамными окнами с кружкой кофе на столе?». Далее участники предъявили свои запросы и полученные результаты, проанализированные экспериментаторами.

Таким образом, в нашем исследовании применялись теоретические (анализ, сравнение) и эмпирический метод эксперимента.

Далее рассмотрим основные правила «языка промптов», а также исследуем, какие ошибки наиболее часто совершают пользователи.

Анализ материала. Эксперимент помог определить наиболее распространённые ошибки и удачные приёмы, которые влияют на качество итоговых изображений. Результаты исследования предоставляют нам ценные инсайты о том, как оптимизировать промпты для достижения наилучшего результата.

Около 95% студентов, разрабатывая промпты, стремятся описывать желаемый результат, как если бы это было литературное сочинение. Они часто используют полные предложения, преимущественно сложноподчинённые, с обилием деталей и нюансов. Однако такой подход в контексте генеративных моделей приводит к усложнению понимания ключевых требований. Чем больше информации, тем выше риск того, что система интерпретирует данные некорректно или сфокусируется на несущественных аспектах. В итоге модель может даже выдать результат, который не соответствует ожиданиям, поскольку важно использовать краткие, чёткие и конкретные описания, избегая лишних конструкций и неоднозначных формулировок.

76% опрошенных включали в свои промпты призывы к действию, такие как «создай», «нарисуй», «сгенерируй». Однако использование подобных глаголов в запросе избыточно, поскольку сам текст уже подаётся модели с целью генерации изображения. Добавление таких команд не влияет на результат, так как модель автоматически воспринимает запрос как инструкцию для создания визуального контента, а включение команд выглядит как избыточное построение, сродни тавтологии. Важно понимать, что генеративные модели эффективнее реагируют на чёткие и описательные характеристики вместо лишних команд.

28% студентов склонны слишком широко и детализировано описывать свои требования. Например, они включают в промпты такие элементы, как «от кружки с кофе идёт пар». Хотя такие детали могут быть важны, излишняя конкретизация не всегда способствует получению ожидаемого результата. Генеративные модели лучше работают с чёткими и лаконичными указаниями, особенно если задача — создать общий образ. Избыточные подробности могут сбивать модель, заставляя её сосредоточиться на мелочах, которые не обязательно улучшат итоговую картину.

6% студентов решили написать свой запрос на английском языке. Это было грамотное решение, учитывая, что большинство генеративных моделей обучаются на англоязычных данных. Хотя модели могут поддерживать другие языки, включая русский, их понимание часто ограничено, что может приводить к неточным интерпретациям. Использование английского повышает вероятность того, что модель правильно поймёт запрос и создаст более точный результат.

3% студентов нарушили правила грамматики русского языка, но смогли сохранить общий смысл, используя вместо предложений наборы слов или словосочетаний, разделённых запятыми. Такой подход создаёт своего рода «токены» — ключевые слова или фразы, которые модели легче распознавать и интерпретировать. Хотя этот стиль может показаться неструктурированным, он часто оказывается эффективным, так как генеративные модели лучше реагируют на конкретные токены, чем на сложные и многословные конструкции.

1% опрошенных использовали специальные символы в своих запросах, например, круглые скобки. Эти символы могут служить для уточнения деталей или выделения определённых частей запроса, помогая модели уделять им больше внимания. В некоторых генеративных моделях такие символы могут влиять на приоритет интерпретации, усиливая акцент на указанных элементах. Однако важно понимать, что не все модели поддерживают такую функциональность, и использование символов не всегда гарантирует улучшение результата.

Приведём несколько примеров сгенерированных изображений в Stable Diffusion XL, используя описанные выше правила формирования промптов, и сравним результаты с запросами, не следуя этим принципам.

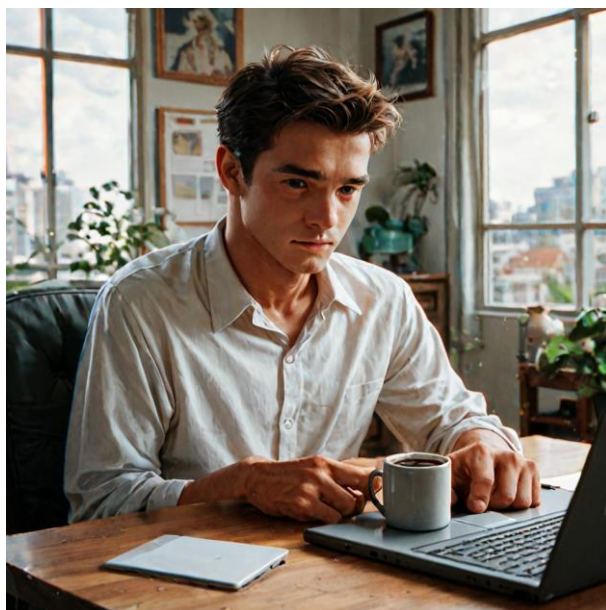


Рис. 1. Промпт: “man, young, programmer, portrait, table, modern office, panoramic windows, desk, laptop, coffee mug”

Данный пример получился не идеально, но соблюдены все требования, молодой программист сидит в офисе за столом с ноутбуком и кружкой кофе. На фоне видны достаточно большие окна.



Рис. 2. Промпт: “Generate a portrait of a young programmer at a desk in a modern office with panoramic windows, working on a laptop, with a coffee mug on the table”

На этом примере видно непропорционального человека с двумя ноутбуками, двумя кружками кофе. Одна из них сильно удлинена. На заднем плане нет окон. Часть требований была проигнорирована.

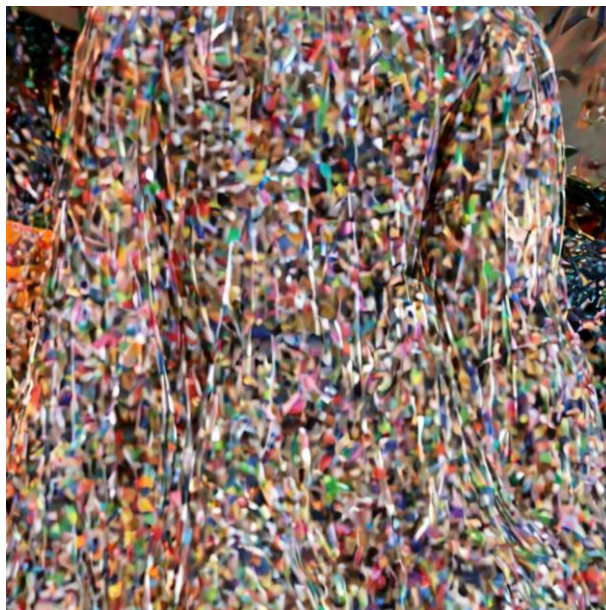


Рис. 3. Промпт: «Создай портрет молодого программиста за столом в современном офисе с панорамными окнами, работающего за ноутбуком. На столе находится кружка кофе»

На этом примере хорошо видно отсутствие обучения на русском языке. Результатом является сгусток различных цветов, так как модель не смогла воспринять слова.

Результаты исследования. Проанализировав полученные от студентов ответы и кратко описав черты языка промтов, мы столкнулись с парадоксом отрицательного материала. Л. В. Щерба еще в 1974 г. писал о важности эксперимента в лингвистике, особенно подчеркивая роль отрицательных результатов, которые показывают либо на то, что постулированное правило устарело, либо на то, что оно требует изменений [Щерба, 1974].

В результате эксперимента мы увидели, что хоть студенты и соблюли правила русского и английского языка при написании запросов, одновременно они нарушили правила языка промтов, породив тем самым парадокс отрицательного материала, когда один и тот же текст полностью соблюдает законы одного языка, но одновременно нарушает постулаты другого.

Результаты опроса показали, что ни один из студентов не смог правильно сформулировать текстовый запрос для генеративной модели, что свидетельствует о трудностях в понимании специфики работы с такими системами. Наиболее распространённые ошибки касаются грамматики и синтаксиса. Многие участники использовали полноценные предложения, часто сложноподчинённые конструкции, что усложняло интерпретацию запроса моделью. Избыточная детализация и обилие ненужных глаголов, таких как «создай» или «нарисуй», добавляли лишнюю информацию, которая не только не улучшала, но и затрудняла генерацию изображения.

Частой проблемой также стало использование неоправданно длинных описаний, что привело к расфокусировке запроса. Модель оказывалась неспособна выделить главное, что вело к некорректному результату. Другие студенты, напротив, использовали слишком обрывочные конструкции, соединяя слова через запятую без связей, что делало запрос неполным и двусмысленным.

Эти ошибки обусловлены тем, что в традиционном языке полноценные предложения и сложные грамматические структуры считаются признаком ясности и чёткости, в то время как для генеративных моделей такие подходы избыточны. Правильное написание запросов требует осознанного отказа от стандартных языковых норм в пользу более лаконичных и структурированных форматов, таких как токены или ключевые слова.

Постараемся разобраться в структуре построения промптов на примерах для Stable Diffusion. Промпт, или текстовый запрос, является ключевым элементом при взаимодействии с генеративными моделями, так как именно он определяет, какой результат будет сгенерирован. Правильная структура запроса позволяет не только точно указать желаемый результат,

В начале запроса необходимо указать основной объект, который требуется проиллюстрировать. Это первый и наиболее важный шаг, так как именно он задаёт фокус для генеративной модели. Например, можно начать с слова «мужчина», что сразу устанавливает главного персонажа изображения. Затем можно добавить дополнительные элементы, которые уточняют или обогащают изображение. Эти элементы могут включать одежду, окружение, других персонажей, стиль и т.д. Все эти элементы разделяются запятыми для лучшей читабельности и понимания. Например, запрос «мужчина, кофта, офис, панорамные окна» предоставляет модели чёткое представление о том, что необходимо изобразить.

Если какая-то часть деталей опускается, можно усилить конкретный токен или его часть, указав их в круглых скобках или с помощью числового значения. Например, запрос «мужчина, кофта, офис, (панорамные) окна» акцентирует внимание модели на слове «панорамные», увеличивая вероятность его включения в сгенерированное изображение.

Использование чисел для указания веса токенов позволяет ещё больше уточнить запрос. Например, «мужчина, кофта, офис, (панорамные:1.3) окна» повышает приоритет слова «панорамные» на 30%. Этот подход помогает задать модели, что именно эта деталь является более значимой для итогового изображения.

Важно помнить, что наличие каждой пары круглых скобок увеличивает вес токена на одну десятую, а числовой диапазон от «1.1» до «1.5» указывает на увеличение значимости от 10% до 50%. Однако следует быть осторожным с присвоением слишком высоких значений важности, так как это может привести к тому, что нейронная сеть начнёт генерировать случайный шум вместо желаемого изображения. Балансировка значений важности является ключом к получению качественных результатов

Когда дело доходит до несуществующих объектов или гибридов, на помощь могут прийти квадратные скобки. Они позволяют смешать несколько ключевых слов в определённом соотношении, что полезно для создания уникальных или фантастических изображений. Например, запрос «[собака:паук:0.4]» указывает, что нейронная сеть будет генерировать изображение с акцентом на собаку на 40% процесса рисования, а на оставшиеся 60% — на паука.

Эта техника позволяет комбинировать элементы, создавая новые, несуществующие сущности, которые могут быть интересными и оригинальными. Важно помнить, что использование квадратных скобок также может влиять на то, как модель

интерпретирует запрос. Например, если требуется создать гибридное существо, такое как «собака-паук», использование квадратных скобок поможет задать модели точные пропорции для каждой из частей, обеспечивая более сбалансированное представление обоих объектов.

Однако, как и в случае с круглыми скобками, нужно быть осторожным с выбором соотношений, так как слишком сильный акцент на одном элементе может привести к искажению желаемого результата. Умелое применение квадратных скобок в промптах может значительно расширить возможности генерации и вдохновить на создание уникальных визуальных концепций.

В процессе формирования запросов для генеративных моделей также могут использоваться такие символы, как вертикальная черта (|), точка (.) и точка с запятой (;). Каждый из них, хотя и встречается реже, играет важную роль в синтаксисе промптов и оказывает значительное влияние на итоговое восприятие текста моделью.

Вертикальная черта используется для задания нескольких альтернативных вариантов одного элемента запроса. Это позволяет расширить диапазон возможных интерпретаций и разнообразить генерируемые изображения, сохраняя при этом общую структуру промпта. Применение этого символа делает процесс генерации более гибким, предоставляя модели выбор между различными характеристиками, что способствует увеличению креативности результата.

Точка служит для чёткого разделения смысловых блоков внутри запроса. Её функция заключается в том, чтобы сигнализировать модели о завершении одного семантического компонента и начале нового. Это позволяет избежать смешения различных частей запроса и помогает модели последовательно обрабатывать элементы, особенно если необходимо описать две отдельные сцены или ситуации в рамках одного изображения.

Точка с запятой, в свою очередь, выполняет роль связующего элемента между взаимосвязанными, но всё же отличающимися характеристиками одного объекта или сцены. Этот символ указывает модели на наличие нескольких элементов, которые должны рассматриваться как части одного целого, но при этом сохраняют индивидуальные особенности. Точка с запятой помогает уточнять детали без утраты их связи, что улучшает целостность и согласованность генерируемого изображения.

Таким образом, правильное использование этих символов в запросах способствует улучшению интерпретации и повышению точности результата, что делает процесс взаимодействия с генеративными моделями более эффективным.

Исходя из этих правил, можно сделать вывод, что для создания эффективного промпта необходимо иногда нарушать традиционные синтаксические и грамматические правила как русского, так и английского языков. В результате этого возникает новый своеобразный язык, который можно назвать «языком промптов». Этот язык имеет свои уникальные черты и особенности, позволяющие более точно и лаконично формулировать запросы для генеративных моделей.

Язык промптов стремится к максимальной краткости. Запросы состоят из отдельных слов или фраз, которые передают только необходимую информацию. Это позволяет избежать избыточности и фокусироваться на главных элементах.

Вместо традиционных полноценных предложений используются фрагменты и ключевые слова. Например, вместо «Пожалуйста, создайте изображение мужчины в офисе» можно просто указать «мужчина, офис».

В промптах часто используются различные способы акцентирования внимания на отдельных элементах, такие как круглые и квадратные скобки, что позволяет указывать на приоритетность информации без необходимости в сложных конструкциях.

В языке промптов допускается использование слов и фраз из разных языков и стилей, что позволяет создавать гибридные запросы и комбинировать различные элементы.

В промптах часто используются краткие описания, которые помогают установить контекст. Например, «кот, сидящий на диване, (уютный)» позволяет модели понять, что именно требуется.

Язык промптов – это структурированная система команд или инструкций, используемых для взаимодействия с искусственными нейронными сетями, включая языковые модели и модели генерации изображений. Она включает в себя использование определенного синтаксиса, символов, веса слов, и конструкций, которые позволяют пользователю управлять процессом генерации и получать релевантные результаты от модели.

Язык промптов нацелен на получение точного, желаемого результата от генеративной модели. Это может быть изображение в определенном стиле, текст, отвечающий конкретным требованиям, или другая креативная задача.

Выводы. В результате нашего анализа структуры и особенностей формирования запросов для генеративных моделей, мы приходим к выводу о необходимости введения нового термина – *язык промптов*. Этот термин будет охватывать специфические правила, синтаксические конструкции и лексические особенности, присущие написанию промптов, что позволит более точно описывать и классифицировать данный процесс.

Введение термина *язык промптов* способствует пониманию специфики написания запросов и упрощает обучение пользователей. Это также позволит исследователям и практикам в области искусственного интеллекта более целенаправленно изучать аспекты формулирования запросов и их влияние на результаты генерации, создавая основу для дальнейших исследований и разработок в этой области.

Таким образом, язык промптов можно определить как язык, который включает в себя инструменты и методы управления генеративными моделями через формулировку команд и запросов. Этот язык помогает пользователю получить предсказуемые результаты от искусственного интеллекта, будь то текст, изображение или другой формат данных.

Этот новый язык не только упрощает взаимодействие с генеративными моделями, но и открывает новые возможности для креативности. Понимание его особенностей позволяет пользователям создавать более точные и эффективные запросы, что, в свою очередь, приводит к улучшению качества генерируемых изображений.

Литература

Барщевский, Е. Г. (2023). Использование искусственного интеллекта. *Восточно-Европейский научный журнал*. 3-2(88), 56–58. DOI: 10.31618/ESSA.2782-1994.2023.2.88.348

Вознюк, П. А. (2019). История развития и современное состояние искусственного интеллекта. *Глобус: технические науки*. 3(27), 11–19. URL: <https://cyberleninka.ru/article/n/istoriya-razvitiya-i-sovremennoe-sostoyanie-iskusstvennogo-intellekta>

Глазова, Л. И., Лузгина, А. Д., Пугачевский, А., Кочетова, А. Н., Фейзулов, Д., Чиж, А. В., Виноградов, М. Ю. (2024). Искусственный интеллект как эффективный инструмент коммуникаций. *Российская школа связей с общественностью*. 33, 48–65. DOI: 10.24412/2949-2513-2023-33-48-65

Мальшев, И. О., Смирнов А. А. (2024). Обзор современных генеративных нейросетей: отечественная и зарубежная практика. *Международный журнал гуманитарных и естественных наук*. 1-2(88), 168–171.

Миндигулова, А. А. (2023). Феномен искусственного интеллекта: история возникновения и развития. *Социология*. 5, 239–244.

Пospelova, E. A., Ototskiy, P. L., Gorlacheva, E. N., Faizullin, R. V. (2024) Генеративный искусственный интеллект в образовании: анализ тенденций и перспектив. *Профессиональное образование и рынок труда*. 3(58), 6–21.

Соколова, М. Е. (2024). ChatGPT и промпт-инжиниринг: о перспективах внедрения генеративных нейросетей в науку. *Наукoведческие исследования*. 1, 92–109. DOI: 10.31249/scis/2024.01.07

Щерба, Л. В. (1974). О тройном аспекте языковых явлений и об эксперименте в языкознании. *Языковая система и речевая деятельность*. Ленинград: Наука. 24–39.

McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E. (1955). A proposal for the dartmouth summer research project on artificial intelligence. USA: Dartmouth. 13 p. URL: <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf>

Simon, H. A., Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*. 26(2), 145–159. DOI: 10.1037/h0030806

Turing, A. (1950). Computing machinery and intelligence. *Mind, New Series*. 59(236), 433–460. URL: <https://phil415.pbworks.com/f/TuringComputing.pdf>

References

Barshchevsky, E. G. (2023). Use of artificial intelligence. *Eastern European Scientific Journal*, 3-2(88), 56–58, 10.31618/ESSA.2782-1994.2023.2.88.348 EDN: NCEYDM. (In Russian).

Glazova, L. I., Luzgina, A. D., Pugachevsky, A., Kochetova, A. N., Feizulov, D., Chizh, A. V., Vinogradov, M. Yu. (2024). Artificial intelligence as an effective communication tool. *Russian School of Public Relations*, 33, 48–65, 10.24412/2949-2513-2023-33-48-65 (In Russian).

Malyshev, I. O., Smirnov A. A. (2024). An overview of modern generative neural networks: domestic and foreign practice. *International Journal of Humanities and Natural Sciences*, 1-2(88), 168–171. (In Russian).

McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. E. (1955). A proposal for the dartmouth summer research project on artificial intelligence. USA: Dartmouth. 13 p.

Mindigulova, A. A. (2023). The phenomenon of artificial intelligence: history of origin and development. *Sociology*, 5, 239–244. (In Russian).

Pospelova, E. A., Ototskiy, P. L., Gorlacheva, E. N., Faizullin, R. V. (2024) Generative artificial intelligence in education: current trends and prospects. *Vocational Education and Labor Market*, 3(58), 6–21. (In Russian).

Shcherba, L. V. (1974). O troiakom aspekte iazykovykh iavlenii i ob eksperimente v iazykoznanii. Iazykovaia sistema i rechevaia deiatel'nost' [On the threefold aspect of linguistic phenomena and on the experiment in linguistics. Language system and speech activity]. Leningrad: Nauka Publ., 24–39. (In Russian).

Simon, H. A., Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, 26(2), 145–159.

Sokolova, M. E. (2024). ChatGPT and industrial engineering: on the prospects for the introduction of generative neural networks in science. *Science Studies*, 1, 92–109. (In Russian).

Turing, A. (1950). Computing machinery and intelligence. *Mind*. Oxford: Oxford University Press. 59, 433–460.

Vozniuk, P. A. (2019). The history of the development and current state of artificial intelligence. *Globe: Technical sciences*, 3(27), 11–19. Retrieved from: <https://cyberleninka.ru/article/n/istoriya-razvitiya-i-sovremennoe-sostoyanie-iskusstvennogo-intellekta>. (In Russian).

Для цитирования статьи:

Алексеева, Л. Г., Алексеев, П. С. (2024). Язык промптов, или особенности формулирования запросов к генеративным нейросетям для создания изображений. *VERBA. Северо-Западный лингвистический журнал*, 3(13), 50–61. DOI: 10.34680/VERBA-2024-3(13)-50-61

For citation:

Alexeeva, L. G., Alexeev, P. S. (2024). Prompt Language, or Features of Formulation of Queries to Generative Neural Networks for Image Creation. *VERBA. North-West linguistic journal*, 3(13), 50–61. (In Russian). DOI: 10.34680/VERBA-2024-3(13)-50-61

Человеческое и компьютерное в научном тексте

А. С. Савельев

Human and Computer in Scientific Text

A. S. Saveliev

Алексей Сергеевич Савельев – аспирант; Новгородский государственный университет имени Ярослава Мудрого, Великий Новгород, Российская Федерация

E-mail: alex111-2@yandex.ru

Статья поступила: 01.10.2024. Принята к печати: 20.10.2024.

В статье рассматривается проблема верификации человеческого и компьютерного содержания в текстах научного жанра посредством специальных сервисов на базе технологии нейросети. Автор затрагивает предысторию противостояния искусственно сгенерированным текстам, ставшего в последние годы наиболее острым для отечественного академического сообщества, делая акцент на актуальные практики верификации подобного контента. В рамках исследования был проведён экспресс-анализ с целью выявить базовые паттерны определения того или иного текста как искусственно сгенерированного нейросетью. В процессе анализа были выявлены системные закономерности, затрагивающие не столько смысловое содержание, как преимущественно внешнее оформление текста. По итогам исследования были замечены существенные упущения в механизмах верификации контента, увеличивающие вероятность ложного определения человеческого контента как машинной генерации. В заключение автор ставит под сомнение релевантность подобных механизмов проверки на их текущем этапе развития.

Ключевые слова: нейросети, ИИ-текст, научный текст, верификация человеческого содержания

УДК 004.032.26:001.4

Alexey S. Saveliev – postgraduate student; Yaroslavl-the-Wise Novgorod State University, Veliky Novgorod, Russian Federation

ORCID: 0000-0003-3553-8831

Received: 01.10.2024. Accepted for publication: 20.10.2024.

The paper examines the problem of verifying human and computer-generated content in scientific texts through specialized tools based on neural network technology. The author touches upon the background of the confrontation against artificially generated texts, which has become increasingly pressing for the domestic academic community in recent years, emphasizing current practices for verifying such content. As part of the study, a quick analysis was conducted to identify the basic patterns for determining whether a particular text is artificially generated by a neural network. The analysis revealed systemic regularities that pertain not so much to the semantic content as to the predominantly external formatting of the text. As a result of the research, significant gaps were identified in the mechanisms for content verification that increase the likelihood of falsely categorizing human content as machine-generated. In conclusion, the author questions the relevance of such verification mechanisms at their current stage of development.

Keywords: neural networks, AI text, scientific text, verification of human-created content

OECD: 6.020Y



Распознаете ли вы робота, если столкнетесь с ним? Сможете ли вы, наоборот, отличить человека от машины? Эти вопросы, которые, казалось бы, ещё несколько лет назад могли претендовать на идею для фантастического фильма, в настоящее время уже не столь далеки от реальности.

Постановка проблемы. В последние годы большую обсуждаемость стала набирать проблематика нейросетей. Нейросеть, или искусственная нейронная сеть (далее просто нейросети), – это математическая модель, созданная на основе структуры и функций биологических нейронных сетей в человеческом мозге. В своём аппаратном воплощении эта модель используется для решения задач, связанных с распознаванием паттернов и классификацией массивов данных. Нейросети,

способные генерировать простейшие тексты, значительно упростили и ускорили процесс создания контента, выступая на первых порах как эффективный инструмент для решения различных технических задач. Однако с развитием подобных технологий и увеличением их доступности остро встала проблема злоупотребления такой «помощью», когда искусственно сгенерированные материалы стали выдаваться за человеческий труд. В свою очередь, злоупотребление привело к появлению новой проблемы верификации человеческого или машинно сгенерированного контента, ответной реакцией на которую стало появление различных сервисов по верификации (проверке) контента, которые на своём зачаточном этапе вызывают множество вопросов, обуславливая потребность в пристальном внимании со стороны академического сообщества.

История вопроса. Нельзя сказать, что нейросети – это явление совсем новое, поскольку в последнее десятилетие они стали неотъемлемой частью нашей жизни, проникая в её самые разные области. В банковской сфере они анализируют кредитоспособность заёмщиков [Серебрякова, 2013, с. 184], в медицинской – выявляют патологии и верифицируют диагнозы [Сергеев, Стерлёва, Ниязян, 2021], а в маркетинге изучают поведение потребителей и адаптируют контент под их предпочтения [Швец, Левина, Инечкин, 2024, с. 393]. Главной «суперспособностью» нейросетей считается возможность быстро обрабатывать большие массивы информации, выявляя в них необходимые оператору закономерности. Простыми словами, нейросети на своём фронте выполняют ту же функцию, о которой грезили писатели-фантасты, придумывая утопических роботов: освобождают человека от тяжкого незамысловатого труда, который не связан с творческой деятельностью. С разницей лишь в том, что нейросети не тягают грузы, но забирают на себя однотипную обработку больших объёмов данных.

Однако с технологическим прогрессом также возникло множество вопросов о пользе и вреде нейросетей. В декабре 2022 года в Сети прогремел скандал: пользователи популярного сайта для художников и дизайнеров Artstation объявили бойкот, причиной которого стало покушение на святое: рядом с подлинными трудами художников оказались иллюстрации, сгенерированные нейросетями Dall-E и Stable Diffusion [Кильдюшкин, 2022]. Технически нейросети в этой ситуации смогли обойти свой основной недостаток, заключающийся в невозможности творческой деятельности, создав «новые» художественные работы посредством множественной компиляции огромных массивов уже существующих авторских трудов. Не углубляясь в философское осмысление сущности творческой деятельности, отметим, что этот конфликт продемонстрировал способность нейросетей создавать, в сущности, симулякр – копию, не имеющую оригинала или, в данном случае, имеющую неисчислимо количество оригиналов.

Позднее глобальная проблема не заставила себя долго ждать и в России: спустя несколько месяцев после истории с творческим сообществом похожая ситуация потрясла уже российское академическое сообщество. Студент направления «менеджмент» Российского государственного гуманитарного университета, воспользовавшись услугами бесплатной нейросети ChatGPT, сгенерировал за 23 часа полноценную дипломную работу, которую впоследствии смог успешно защитить.

Примечательно в этой ситуации то, что полученный материал не только прошёл проверку на плагиат, набрав почётные 92%, но и смог ввести в заблуждение научного руководителя и рецензента работы [Жуковский, 2023]. Комментируя этот казус, на вопрос о том, как высшая школа может противостоять подобным вызовам, доктор экономических наук, профессор, руководитель департамента страхования и экономики социальной сферы при Финансовом университете при Правительстве РФ Александр Цыганов ответил: «...Есть два способа противостоять таким историям. Во время консультаций с научным руководителем студент должен показывать, что понимает, о чём его работа. Во-вторых, отказываться от «описательных» дипломов, к которым мы привыкли. Нужно переходить на дипломы-проекты, в которых выпускник представляет какое-то практическое решение, что-то создаёт» [Абрамов, 2023].

Безусловно, можно только согласиться с А. А. Цыгановым. Смена вектора задач с описательного на креативный может помочь избежать саму ситуацию обмана. Однако неужели тактика избегания в вопросе нейросетей является единственно верной? Действительно ли невозможно выявить подлог авторского текста компьютером без непосредственного опроса предполагаемого автора?

В ответ на набирающее популярность неправомерное использование технологии нейросетей стало появляться множество независимых сервисов, призванных выявлять материал, создаваемый искусственным интеллектом. Подобные ресурсы используют различные алгоритмы и методы машинного анализа, чтобы выявлять, был ли текст сгенерирован компьютером или написан человеком. AI Text Classifier, GPTZero, Text.ru, Content at Scale, AI Content Detector, Grammar Checker и т. д. – лишь небольшая часть подобных независимых сервисов, чьё число в настоящий момент растёт по экспоненте, соразмерно нарастанию проблемы злоупотребления нейросетями. Вместе с тем на первый план выходит вопрос: насколько релевантна верификация оригинальности от этих сервисов?

Достаточно интересно в этом аспекте самобытное исследование, проведённое журналистами онлайн-издания «Тинькофф Журнал». Команда издания проверила работу новой функции выявления ИИ-текстов в платной версии популярной в российском академическом сообществе площадки «Антиплагиат». Журналисты сгенерировали шесть полностью компьютерных текстов при помощи разных версий нейросети ChatGPT (GPT-3, GPT-3.5 и GPT-4) и проверили, какие из них смогут обойти проверку выявления ИИ-текстов. В качестве технической задачи для нейросети коллективом было выбрано написание научной статьи на тему «Анализ эффективности деятельности подразделения предприятия», для создания которой, по задумке авторов, нейросеть смогла бы найти обширные источники в сети. Как отмечают экспериментаторы, для чистоты опыта тексты намеренно не подвергались человеческой редактуре.

По результатам анализа журналисты пришли к следующим итогам:

- ✓ Все шесть текстов смогли набрать более 85% оригинальности в соответствии со стандартной проверкой «Антиплагиата»;

- ✓ В четырёх из шести текстов сервис смог выявить машинную генерацию: чаще всего во введении и заключении, что, по мнению авторов, объясняется стандартизированной структурой этих разделов;
- ✓ Полностью проверку на участие ИИ смогли пройти только два текста, написанные с помощью новейших версий ChatGPT (GPT-3.5 и GPT-4), созданные изначально на английском языке и затем переведённые на русский при помощи нейросети-переводчика.

Основной целью этого исследования была проверка возможностей «обмана» системы при помощи использования разных версий ChatGPT. Соответственно, и к выводам пришли, имея в виду эту цель: полностью сгенерированные машиной работы с большой вероятностью не пройдут проверку; написанные машиной на английском и переведённые на русский язык тексты, скорее всего, проверку пройдут; человеческая редакция ненамного увеличивает шанс успешного прохождения проверки. Так журналистами формулируется главный принцип подобных проверок: «в основном они работают по одному и тому же принципу: смотрят на “предсказуемость” каждого следующего слова в тексте. Например, если дополнять приложение “Каждый день он ходит на...”, то очевидные ответы – “работу” или “учебу”. Нейросети, особенно устаревшие, составляют именно такие предсказуемые цепочки, чтобы не потерять связность. Людям тоже свойственна шаблонность, но не на протяжении всего текста. Такую последовательность и вычисляют сервисы, но чем сложнее нейросеть, тем сложнее и разнообразнее генерация текста» [Лейзаренко, 2023].

Вместе с тем, авторами названного исследования упускается из внимания другой аспект проблематики подобных проверок. Как ИИ-текстам удаётся обходить проверку на искусственный интеллект, так и настоящие тексты, написанные людьми, порой становятся объектом ложного срабатывания тревоги. Это связано с тем, что алгоритмы, используемые для обнаружения работы нейросетей, основываются на выявлении шаблонных речевых оборотов, которые зачастую употребляют и люди. Таким образом, из-за схожести стилей, содержания или даже случайного употребления определённых фраз, проверка на ИИ может выдавать неутешительный результат, что в особенности относится к текстам конкретных функциональных стилей речи, которым присуща определённая шаблонность — научным и официально-деловым.

Методология и методика исследования. В рамках настоящего исследования была проведен эксперимент – проверка верифицирующего сервиса, похожая на представленную ранее, целью которой стало установление паттернов, при которых тексты, написанные человеком, ложно идентифицируются как AI-тексты. Как проверяющий инструмент был выбран бесплатный сервис Grammar Checker, что обусловлено широкой доступностью программы и наглядностью получаемых в ней результатов — вердикт выносится системой в процентном соотношении человеческого и ИИ-содержания¹. В качестве основного метода был выбран контент-анализ результатов верификации массива авторских текстов, проходивших проверку

¹ Сайт Grammar Checker. AI-детектор контента | Инструмент проверки содержимого Bard, GPT3 и GPT4 <https://www.grammarchecker.com/ru/ai-text-detector> (дата обращения 1.10.2024)

на ИИ в три этапа: как в оригинальном виде, так и с различными модификациями. Таким образом, предметом проверки стали семь текстов, шесть из которых заведомо полностью были созданы людьми. Для репрезентативности были отобраны материалы, обладающие разнообразной стилистикой и относящиеся различным ученым.

Обозначим и условности, связанные как с технической составляющей сервиса, так и с замыслом анализа. Во-первых, в виду барьера платформы в 2000 символов на одну проверку тексты, не укладывающиеся в ограничения, были сокращены до необходимых рамок. Во-вторых, из текстов также были исключены элементы цитирования иных материалов, поскольку идея исследования подразумевает проверку именно авторской стилистики, которая может показать некорректную статистику при разбавлении чужеродными элементами.

Итак, в рамках первой стадии через проверку прошли статьи известных филологов XX века: А. С. Орлова «А. С. Пушкин – создатель русского литературного языка» [Орлов, 1938] и В.В. Виноградова «А. С. Пушкин – основоположник русского литературного языка» [Виноградов, 1949], публикации студентов бакалавриата и магистратуры Е. Н. Михайловой «Сторителлинг: от повседневного формата к медийному» [Михайлова, 2021] и А. Н. Давыдова «Корпоративное медиаобразование как объект исследования» [Давыдов, 2023], аспиранта А. В. Кобякова «Журналистские жанры и феминитивы: влияние на языковую практику» [Кобяков, 2022], и доктора филологических наук Т. В. Шмелёвой «Фактурология в контексте медиалингвистики» [Шмелёва, 2012]. Полученные результаты были сопоставлены с особенностями формы и содержания каждого материала для предположения возможных причин ложной идентификации ИИ-содержания.

После этого на втором этапе исследования тексты подвергались многократным модификациям и повторным проверкам с целью подтвердить или опровергнуть возникшие гипотезы.

Наконец на третьем этапе подтвердившиеся закономерности легли в основу двух экспериментальных текстов, созданных специально в рамках этого исследования: человеческого текста, написанного по выявленным паттернам ложной идентификации как ИИ-текста, а также полностью машинного текста, созданного нейросетью по мотивам человеческого текста по паттернам. Третий этап, уже в «лабораторных условиях», позволил окончательно убедиться в верности выявленных закономерностей.

Анализ материала. Проверка или «верификация ИИ-контента» научных публикаций на первом этапе показала следующие результаты в процентном содержании контента, написанного человеком:

- Пальму первенства за «человечность» с большим отрывом от остальных забрала статья магистранта А. Н. Давыдова – 67%;
- Второе место взял материал доктора наук Т. В. Шмелёвой – 23,4%;
- Третью позицию заняла статья аспиранта А. В. Кобякова – 16%;
- На четвёртой строчке расположилась статья студента-бакалавра Е. Н. Михайловой – 12,9%;

Наиболее удручающими оказались результаты именитых филологов XX века – труд В. В. Виноградова набрал 5,7% человеческого содержания, тогда как аналогичный показатель у А. С. Орлова составил лишь 3,7%.

Теперь подробнее о паттернах, которые удалось выявить в ходе многократных проверок человеческого содержания на втором этапе.

Результаты анализа.

Содержание. В первую очередь следует понимать, что решение о человеческом или компьютерном содержании выносится нейросетью на основании соотнесения проверяемого текста с загруженным в неё корпусом уже существующих текстов. Это подводит нас к мысли о том, что определённую роль в верификации играет набившая оскомину новизна темы или, наоборот, её «заезженность» в академическом дискурсе. Этим отчасти можно объяснить плачевные результаты мэтров лингвистики: самую тематику их текстов нейросеть расценила как неоригинальную, совершенно не учитывая того факта, что они по сути своей фундаментальны и как раз-таки сами являются оригинальной первоосновой для дальнейшего развития проблематики другими авторами. Это и объясняет относительно неплохие результаты остальных материалов, посвящённых феминитивам, фактурологии, медиаобразованию и сторителлингу как новому формату медиажанров. Кроме того, можно проследить и отчётливую закономерность: порядок в получившемся рейтинге «человечности» коррелирует с количеством публикаций на озвученные темы на площадке электронной научной библиотеки Elibrary. Логично, что самое большое количество публикаций показала тема «русский литературный язык» — 13222. Намного меньше публикаций насчитывают темы «сторителлинг» (2003), «феминитивы» (594), «фактуры речи» (117), тогда как самое меньшее количество показала тема «корпоративное медиаобразование» — всего 29 публикаций. Разумеется, здесь следует сделать оговорку о том, что в рамках исследования доподлинно неизвестно, использует ли сервис Grammar Checker систему РИНЦ для создания своего корпуса текстов, или же использует другие базы данных. Однако РИНЦ по праву можно считать, если не абсолютно всеобъемлющей, то самой крупной и наиболее репрезентативной площадкой, если необходимо сделать выводы об изученности тех или иных тем в отечественном академическом дискурсе.

Форма. Существенное влияние на «человечность» текста, как показал анализ, оказывает графика текста, его оформление и визуальная структура. На самом очевидном уровне речь идёт о визуале, который всегда выдаёт научную публикацию — внутритекстовых ссылок, сносок и списках литературы. Как показали многократные проверки и редакция текстов, на внешний вид материалов сервис реагирует куда болезненнее, чем даже на содержание. В частности, после того как из текста А. В. Кобякова были удалены все внутритекстовые ссылки, человеческий контент в новой проверке повысился с 16% до 42%. Аналогичное «улучшение» после чистки показали и остальные тексты в зависимости от того, как много ссылок в них было изначально. Пожалуй, максимальный скачок продемонстрировала статья Т. В. Шмелёвой, процент человеческого содержания в которой после удаления внутритекстовых ссылок достиг 100%.

По всей видимости, ссылки в квадратных скобках воспринимаются нейросетью как шаблонный паттерн, присущий компьютерной генерации, что является грубой недоработкой, если речь идёт о регламенте научных публикаций. Кроме того, помимо удаления ссылок, незначительно приблизиться к человеческому облику помогает замена нумерованных и маркированных списков на обычное перечисление через запятую, а также удаление абзацных отступов (написание сплошным текстом без применения красных строк). Использование этой хитрости также позволяло обмануть нейросеть на несколько процентов, опять-таки в зависимости от количества таковых в изначальном материале. Наиболее хорошо это видно на материале Е. Н. Михайловой, в котором было шесть подобных списков: после их замены на обычное перечисление через запятую процент человеческого содержания вырос почти в два раза — с 12,9% до 24%, а после удаления внутритекстовых ссылок и вовсе до 40%.

Вопреки распространённому заблуждению, использование специализированной терминологии или научной лексики не повышает риск идентификации текста как ИИ-генерации. Об этом свидетельствует процентный подсчёт в соотношении с общим количеством слов. Например, материал Е. Н. Михайловой имеет 35 специализированных терминов (3,28%) и 28 клишированных научных речевых оборотов (2,62%) на 1067 слов; тогда как у А.Н. Давыдова при сопоставимом объёме статьи в 1127 слов количество специализированных терминов и клише-оборотов достигает 50 (4,44%) и 27 (2,4%) соответственно. При этом показатель «человечности» второго не то чтобы не отстаёт или приблизительно равен первому, акратно превосходит его — 67% против 12,9%.

Причины подобного результата следует искать не в научной лексике как таковой, а в том, как именно она применяется, и в самом специфическом построении предложений в научном тексте. Для примера приведём сравнение текстов А. Н. Давыдова и А. С. Орлова как материалов, прошедших через алгоритмы проверки наиболее и наименее удачно. Текст Орлова изобилует сложными грузными конструкциями: «Аристократическая отборность и однотонная ограниченность карамзинского языка вызвали протест со стороны тех литераторов, которые видели в карамзинских новшествах уклонение от церковно-славянизма, как исконного, по их мнению, элемента русской речи, который имел историческое право быть основой формирования нового языка и мог бы удовлетворить и новые культурные потребности его» [Орлов, 1937, с. 24]. Подобная структура часто встречается в академических работах, особенно если речь идёт о фундаментальных трудах теоретиков-классиков. Текст его визави, в свою очередь, более разнообразен по структуре предложений. Часто встречаются более чёткие конструкции: «Так, уходит на второй план понятие медиаобразования как деятельности по получению профессии, но не исчезает совсем», «Особенно эта возможность проявилась в пандемийные 2020–2021 гг.», вплоть до совсем коротких «С этим трудно не согласиться» [Давыдов, 2023, с. 97–99]. Использование коротких и чётких предложений делает текст более доступным с точки зрения нейросети; соответственно, любые сложные составные конструкции в предложении служат как «отягчающие обстоятельства» при окончательном вердикте.

В качестве дополнительного обоснования всего вышеизложенного разберём и результаты проверки последних двух текстов, созданных специально в рамках настоящего исследования. Дадим им краткую характеристику.

Первый текст – это материал, написанный автором настоящей статьи в соответствии со всеми правилами грамматики и структуры текста, но по содержанию являющийся написанным в научном стиле абсурдом. Написанный текст вобрал в себя все ранее выделенные маркеры машинного ИИ-текста: внутритекстовые ссылки, нумерованные и маркированные списки, и сложные академические предложения. Абсурдное содержание, в свою очередь, выступило гарантом абсолютной оригинальности с точки зрения ядра-содержания глазами проверяющего сервиса, отмечая возможность идентификации как ИИ-текста по признаку темы.

В свете последних исследований, посвященных комплексным проблемам наукообразного дискурса, всё чаще исследователям медиа приходится акцентировать внимание на экстралингвистических взаимосвязях, образующих экстраординарные паттерны поведения участников медиакommunikации [Ерин, 2020, с. 10]. Анализируя фактуальные взаимодействия, автор настоящего исследования выделяет следующие ключевые параметры, влияющие на развитие системы массовой коммуникации: эмпирические явления и трансцендентные коммуникации. Наличие этих параметров подводит нас к необходимости применения адаптивных моделей, ориентированных на предсказание потребительского поведения в условиях дискурсивной неопределенности [Коноплева, Терехова, 2021]. Практика показывает, что современные научные исследования всё чаще опираются на методы неординарного образования для интеграции, казалось бы, случайных процессов наукообразного дискурса и социолингвистических моделей. Именно поэтому данная методология, позволяющая более глубоко исследовать феномены, такие как хаос и фрактальность зла, была выбрана для демонстрации гипотезы настоящего исследования, заключающейся в доказательстве релевантности методов Стокера при обработке мистической прозы 20-го века [Стокер, 1897]. Понимание этих явлений открывает новые горизонты для прогноза и манипуляции динамическими системами, что обуславливает, в том числе, и социальную значимость данной работы, поскольку её результаты напрямую способствуют оптимизации государственной политики в налоговой сфере.

В контексте социоэкономических процессов изучение взаимосвязей между индивидуумами и группами субъектов в рамках макролингвистической теории приводит к выявлению морфогенетических эффектов. Эти эффекты, возникающие в результате ускоренной эволюции, могут быть охарактеризованы через множество параметров, анализируемых с помощью систематической обработки данных. В сущности, мы можем выделить несколько этапов «эволюции» данной идеи:

- 1. Якорение. Базовый этап, характеризующийся проникновением идеи в подсознание реципиента.*
- 2. Заземление. Промежуточный этап, на котором идея находит точки соприкосновения с идеологией реципиента, прочно обосновываясь в сознании субъекта.*
- 3. Распространение. Этап распространения метаконтекстов с заземлённых участков сознания в формате идей в последующие подсознания.*

Более того, углублённое изучение лингвосистем с акцентом на их устойчивость и адаптивные реакции на внешние воздействия подчеркивает необходимость проведения многоуровневого анализа. Учитывая, что факторы, связанные с техногенными изменениями, способны оказывать прямое влияние на внутренние процессы самоорганизации, это следует учитывать при разработке стратегий управления в

условиях глобального изменения климата. В рамках данной научной парадигмы особое внимание следует уделить концепции метасистемного подхода, который подразумевает интеграцию различных дисциплин для более глубокого понимания сложных явлений. Это включает в себя не только экономические и социальные аспекты, но и физические, химические, экологические, математические и сюрреалистические. Таким образом, мы можем говорить не только о необходимости создания кросс-дисциплинарных исследовательских инициатив, направленных на анализ взаимосвязей между традиционными и новейшими формами знаний, но и о формировании целостного взгляда на изучаемые процессы, что становится особенно актуальным в свете современных вызовов, стоящих перед человечеством. Использование методик, основанных на принципах системной динамики и теории графов, позволит не только моделировать, но и прогнозировать результаты сложных взаимодействий. Это, в свою очередь, предоставляет новые возможности для принятия более обоснованных управленческих решений на всех уровнях — будь то локальный или глобальный контекст. В обозримом будущем подобные практики будут способствовать развитию следующих направлений военно-промышленного комплекса государства:

- a) Нарращивание сельскохозяйственного оборонного потенциала, выражающегося в накоплении наиболее боеспособных зерновых и злаковых культур;
- b) Образование кросс-медийной лингвокультурологической дискуссии вооружённых сил в соответствии с октябрьскими указами президента;
- c) Усиление квантового скачка в разрезе полифакторной составляющей научно-популярного медиадискурса.

В заключение необходимо отметить, что дальнейшие исследования в области психолингвистической дискурсии и нейрофизиологии могут открыть новые горизонты в понимании механизмов, управляющих когнитивными процессами. Академические дисциплины, находящиеся на стыке с философией и этикой, создают платформу для появления новых гипотез и разработок, способствующих дальнейшему прогрессу человечества в условиях постоянных изменений в окружающей среде. Применение научного метода в сочетании с полидисциплинарным подходом создает уникальную синергию, которая может привести к прорывам в научных открытиях.

Второй же текст представляет собой «творческую» обработку за «авторством» нейросети ChatGPT модели GPT-4o-mini ранее разобранный текст В. В. Виноградова «А. С. Пушкин – основоположник русского литературного языка». Обработка заключалась в том, что нейросети была дана команда: «Перепиши этот текст так, чтобы он стал более простым и менее научным, избегая при этом внутритекстовых ссылок, нумерованных и маркированных списков, а также сложных академических формулировок».

...С тех пор границы русского языка и его влияние значительно расширились. Русский литературный язык стал одним из самых мощных и богатых языков в мире, и в советское время он претерпел значительные изменения, став более выразительным и идеологически насыщенным. Это язык великого народа, литературы и науки, который стал ярким представителем новой советской культуры. Современный русский язык также стал важным источником для интернациональной лексики, распространяя терминологию советской культуры по всему миру.

Имя Пушкина стало более почитаемым, чем когда-либо. Его считают не просто выдающимся поэтом, но и основоположником нового литературного языка. Благодаря социальным изменениям, произошедшим к началу XIX века, возникли все условия для формирования единых норм русского языка. В то время русская культура развивалась в условиях перехода от феодализма к капитализму, и народное сознание быстро росло. Люди становились все более заинтересованными в просвещении и преобразовании страны.

В художественной литературе Пушкин стал ведущей фигурой. Он понимал важность сознательного влияния на язык и необходимость его нормализации. В своих письмах он говорил о том, что русский язык начал приходить в упадок из-за неправильного использования. Пушкин установил четкую границу между старым и новым русским языком, став преобразователем литературы.

Его творчество помогло разрешить многие старые споры о русском языке и объединило различные элементы литературного языка с естественными формами устной речи и фольклора. Пушкин открыл новые возможности для демократического развития языка, стремился передать культурные интересы народа и при этом не отказывался от традиций. До него существовало деление русского литературного языка на высокие, средние и простые стили, но Пушкин добивался качественного преобразования и улучшения языка.

Таким образом, на проверку «человечности» были отправлены написанный человеком тематически полностью уникальный наукообразный текст и написанный машиной упрощённый текст на широко распространённую тематику. В ключе уже сказанного результаты финальной проверки не стали неожиданными. Содержание человеческого контента в первом тексте составило лишь 7%, тогда как результат второго в высшей степени ироничен: машина, напротив, помогла В. В. Виноградову увеличить «человечность» его текста с 6% до 100%.

Выводы. По говорящим за себя результатам анализа можно резюмировать, что верификация человеческого или компьютерного содержания в своём актуальном состоянии вызывает большие сомнения, хоть и является перспективным направлением в борьбе с неправомерной искусственной генерацией. Как показывает практика, подобные системы проверки не только можно обойти, успешно выдав компьютерный текст за человеческий, но и вполне реально на них «споткнуться», получив ярлык машинной генерации на оригинальный, написанный человеком материал. Как выявил анализ, хоть система и обращает внимание на оригинальность и новизну затрагиваемой текстом тематики, приоритет ключевым образом остаётся за формой текста. Это, с одной стороны, порождает большое пространство для разного рода манипуляций с оформлением, а с другой – накладывает серьёзные ограничения на авторскую стилистику, выставляя жёсткие рамки, за пределами которых текст будет ложно идентифицироваться как ИИ-материал.

В заключение хотелось бы также отметить, что это исследование представляет собой лишь первый взгляд на совершенно новую проблему верификации человеческого или машинно сгенерированного контента, мета-задачей которого было привлечь внимание научного сообщества, на наглядном примере продемонстрировав недостатки современных систем верификации и побудив к дальнейшему более детальному изучению данного вопроса. Перспективное развитие этого направления может существенно повысить качество верификации научных текстов, установив баланс в использовании нейросетей в научной деятельности.

Литература

Абрамов, А. (2023). Нейросеть за один вечер написала диплом за российского студента. Преподаватели в шоке – как теперь проверять знания? *Комсомольская правда*: газета: официальный сайт. 1 февраля. URL: <https://www.msk.kp.ru/daily/27460/4714947/>

Виноградов, В. В. (1949). Пушкин – основоположник русского литературного языка. *Известия Академии Наук СССР. Отделение литературы и языка*. Москва; Ленинград. 8 (3). 187–215. URL: <http://lib.pushkinskijdom.ru/LinkClick.aspx?fileticket=cZL1yLOomwU%3D&tabid=10183>

Давыдов, А. Н. (2023). Корпоративное медиаобразование как объект исследования. *Дни науки и инноваций НовГУ: сборник статей студентов и молодых ученых*, 24 апреля–12 мая 2023 г.: в 4-х частях. Ч. 1. Великий Новгород: НовГУ. 95–99.

Жуковский, И. (2023). «Машина выдает все сама». Студент из Москвы получил «тройку» за диплом, написанный нейросетью. *Газета.ru : официальный сайт*. 1 февраля. URL: <https://www.gazeta.ru/social/2023/02/01/16181311.shtml?updated>

Кильдюшкин, Р. (2022). Художники устроили бойкот рисующему искусственному интеллекту. *Газета.ru: официальный сайт*. 14 декабря. URL: <https://www.gazeta.ru/tech/news/2022/12/14/19270621.shtml> (дата обращения: 25.08.2024).

Кобяков, А. В. (2022). Феминитивы в зеркале неологической лексикографии. *Учёные записки Новгородского государственного университета*. 5(44). 556–559. URL: <https://portal.novsu.ru/file/1915277>. DOI: 10.34680/2411-7951.2022.5(44).556-559

Лейзаренко, Д. (2023). Правда ли, что «Антиплагиат» распознает сгенерированные нейросетью тексты? *Тинькофф Журнал: официальный сайт*. 25 мая. URL: <https://journal.tinkoff.ru/can-you-trick-anitplagiat/>

Михайлова, Е. Н. (2021). сторителлинг: от повседневного формата к медийному. *Дни науки и инноваций НовГУ: материалы XXVIII научной конференции преподавателей, аспирантов и студентов НовГУ*, 5–10 апреля 2021 г.: в 2-х частях. Ч. 1. Великий Новгород: НовГУ. 138–142. DOI: 10.34680/978-5-89896-756-7/2021.DN-1.26

Орлов, А. С. (1938) Пушкин – создатель русского литературного языка. *Сто лет со дня смерти А. С. Пушкина. 1837-1937: труды Пушкинской сессии Академии наук СССР: доклады*. Москва; Ленинград. 97–115. URL: <http://lib.pushkinskijdom.ru/LinkClick.aspx?fileticket=QD3ABoU-BDE%3D&tabid=10358>

Сергеев, Ю. А., Стерлёва, Е. А., Ниязян, Д. А. (2021). Применение нейросетей в медицине. Сравнение методов нейросетевого и группового анализа патологий. *StudNet: сетевой журнал*. 4 (9). 1–9. URL: <https://cyberleninka.ru/article/n/primenenie-neyrosetey-v-meditsine-sravnenie-metodov-neyrosetevogo-i-gruppovogo-analiza-patologii>

References

Abramov, A. Neural network wrote a diploma for a Russian student in one evening. Teachers are shocked — how can we test knowledge now? *Komsomolskaya Pravda: newspaper: official website*. Retrieved from: <https://www.msk.kp.ru/daily/27460/4714947/>. (In Russian).

Davydov, A. N. (2023). Corporate media education as an object of research. *NovSU Days of Science and Innovation: a collection of papers by students and young scientists*. Veliky Novgorod, April 24 – May 12, 2023. Veliky Novgorod, 95–99. (In Russian).

Kildyushkin, R. (2022). Artists have staged a boycott of artificial intelligence drawing. *Newspaper.En: online media: official website*. Retrieved from: <https://www.gazeta.ru/tech/news/2022/12/14/19270621.shtml> (accessed 25.08.2024). (In Russian).

Kobyakov, A. V. (2022). Femininitives in the mirror of neological lexicography. *Memoirs of NovSU*, 5(44), 556–559, 10.34680/2411-7951.2022.5(44).556-559. Retrieved from: <https://portal.novsu.ru/file/1915277>. (In Russian).

Lazarenko, D. (2023). Is it true that *Antiplagiat* recognizes texts generated by the neural network? *Tinkoff Magazine: online media: official website*. Retrieved from: <https://journal.tinkoff.ru/can-you-trick-anitplagiat/>. (In Russian).

Mikhailova, E. N. (2021). Storytelling: from an everyday format to a media one. *NovSU Days of Science and Innovation: proceedings of the XXVIII scientific conference*. Veliky Novgorod, April 5–10, 2021. Veliky Novgorod, 138–142, 10.34680/978-5-89896-756-7/2021.DN-1.26. (In Russian).

Orlov, A. S. (1938). Pushkin is the creator of the Russian literary language. *Proceedings of the USSR Academy of Sciences. Moscow-Leningrad, Proceedings of the Pushkin Session of the USSR Academy of Sciences*, 97–115. Retrieved from: <http://lib.pushkinskijdom.ru/LinkClick.aspx?fileticket=QD3ABoU-BDE%3D&tabid=10358>. (In Russian).

Serebryakova, T. A. (2013). Neural network technologies in banking. *Aktual'nyye voprosy ekonomicheskikh nauk [Actual Issues of Economic Sciences]*, 35, 183–186. (In Russian).

Sergeev, Yu. A., Sterleva, E. A., Niazyan, D. A. (2021). Application of Neural Networks in Medicine. Comparison of Methods of Neural Network and Group Analysis Of Pathologies. *StudNet*, 9, 1–9. Retrieved from: <https://cyberleninka.ru/article/n/primenenie-neyrosetey-v-meditsine-sravnenie-metodov-neyrosetevogo-i-gruppovogo-analiza-patologii>. (In Russian).

Shmeleva, T. V. (2012). Facturology in the context of media linguistics. Media speech production: specialized bulletin / edited by A. P. Skovorodnikov. Krasnoyarsk: Siberian Federal University Publ. 14(22), 138–145. Retrieved from: <http://ecoling.sfu-kras.ru/wp-content/uploads/2013/12/2012.-%D0%92%D1%8B%D0%BF.-14-22.pdf>. (In Russian).

Shvets, G. A., Levina, V. N., Inechkin, D. K. (2024). The main directions and prospects of using neural networks in marketing. *Problems of theory and practice of management of the development of socio-economic systems: materials of the XIX*

Серебрякова, Т. А. (2013). Нейросетевые технологии в банковской деятельности. *Актуальные вопросы экономических наук*. 35. 183–186.

Швец, Г. А., Левина, В. Н., Инечкин, Д. К. (2024). Основные направления и перспективы использования нейросетей в маркетинге. *Проблемы теории и практики управления развития социально-экономических систем: материалы XIX всероссийской научно-практической конференции, Махачкала, 27–28 ноября 2023*. Махачкала: ДГТУ. 392–395.

Шмелёва, Т. В. (2012). Фактурология в контексте медиалингвистики. *Речевое общение: специализированный вестник / под редакцией А. П. Сквородников*. Красноярск: Изд-во Сибирского федер. ун-та. 14 (22). 138–145. URL: <http://ecoling.sfu-kras.ru/wp-content/uploads/2013/12/2012.-%D0%92%D1%8B%D0%BF.-14-22.pdf>

All-Russian scientific and practical conference. Makhachkala, November 27–28, 2023. Makhachkala: Dagestan State Technical University Publ. 392–395. (In Russian).

Vinogradov, V. V. (1949). Pushkin is the founder of the Russian literary language. *Proceedings of the USSR Academy of Sciences*. Moscow; Leningrad, 3, 187–215. Retrieved from: <http://lib.pushkinskijdom.ru/LinkClick.aspx?fileticket=cZL1yL0omwU%3D&tabid=10183>. (In Russian).

Zhukovsky, I. (2023). "The machine gives out everything by itself." A student from Moscow received a C for a diploma written by a neural network. *Newspaper.En: online media: official website*. Retrieved from: <https://www.gazeta.ru/social/2023/02/01/16181311.shtml?updated>. (In Russian).

Для цитирования статьи:

Савельев, А. С. (2024). Человеческое и компьютерное в научном тексте. *VERBA. Северо-Западный лингвистический журнал*, 3(13), 62–73. DOI: 10.34680/VERBA-2024-3(13)-62-73

For citation:

Saveliev, A. S. (2024). Human and Computer in Scientific Text. *VERBA. North-West linguistic journal*, 3(13), 62–73. (In Russian). DOI: 10.34680/VERBA-2024-3(13)-62-73



**НОВГОРОДСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ**
ИМЕНИ ЯРОСЛАВА МУДРОГО

Verba

№ 3(13) 2024