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

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

Статья поступила: 01.10.2024. Принята к печати: 20.10.2024.

Received: 01.10.2024. Accepted for publication: 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 нализ 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 (тезаурусном) 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.

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: 6.02OY

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 лингвистический 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 большего 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, Донина, 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 вручную 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-basedlinguistics, представляет исследования, предполагающие использование корпуса для подтверждения некоторых гипотез и идей. Второе направление (corpus-drivenlinguistics) 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 объем» [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 сведению 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* (НКРЯ[1]: 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 диструбутивных 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 shallknow a wordbythecompanyitkeeps» («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», «LatentSemantic 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 десятилетие 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 [Залесская, 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 прайммингу, 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, когдв 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 трубач схохмил 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.

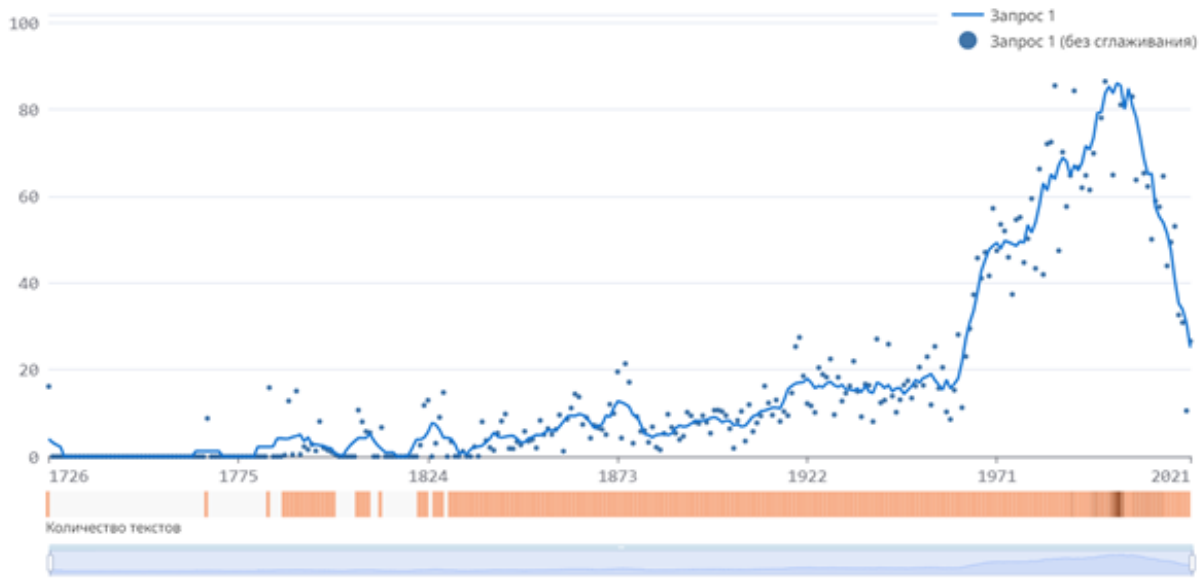

Рис. 1. Частотность употребления вводного слова *естественно* по временной шкале в Национальном корпусе русского языка (Основной корпус)

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 корпусы 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 корпусы, 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 корпусы 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 корпусы 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 удается 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.

Залесская, 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). C*orpus 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, 10.10170S0954394502143018.

Chebotyreva, 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-tekhnicheskaya 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 РуСентилекс. *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 памятник 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'sReadingsin Language Engineering and AppliedLinguistics: Proceedings, Saint Petersburg, November 27, 2017. Saint Petersburg: Creative Commons CCO. 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/llc/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., Донина, 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). *AutomatedSentiment Analysis ofFeminitivesinthe 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 паремиологическим 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: 10.10170S0954394502143018

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/llc/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, 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 yazykoznaniya*, 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 Feminitives 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. *ComputationalLinguistics, 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