УДК 811.161.1
DOI 10.52575/2712-7451-2022-41-3-590-607

Становление и формирование терминологии компьютерной лингвистики

Польщикова О.Н.
Белгородский государственный национальный исследовательский университет,
Россия, 308015, г. Белгород, ул. Победы, 85
E-mail: polshchikova@bsu.edu.ru

Аннотация. Важную роль в исследовании терминологии играет изучение ее формирования в диахроническом аспекте. Несмотря на стремительное развитие сферы компьютерной лингвистики, практически отсутствуют работы по анализу процессов эволюции соответствующей терминологии. Целью данного исследования является изучение процессов становления и формирования терминологии компьютерной лингвистики, выявление основных периодов ее развития. Исследование проводилось на основе метода диахронического анализа. Для изучения процессов становления и развития исследуемой терминологии использовался фактический материал в виде текстов по тематике компьютерной лингвистики, размещенных в монографиях, справочниках, учебной литературе, научных периодических журналах и сборниках, сетевых электронных ресурсах. В результате исследования установлено, что терминология компьютерной лингвистики находится в состоянии непрерывной динамики, выявлено три основных периода её формирования. Полученные результаты вносят вклад в изучение проблем научно-технической терминологии, обоснование тенденций её развития.

Ключевые слова: компьютерная лингвистика, терминология, периоды развития, диахронический анализ, автоматическая обработка естественного языка

Для цитирования: Польщикова О.Н. 2022. Становление и формирование терминологии компьютерной лингвистики. Вопросы журналистики, педагогики и языкознания, 41(3): 590–607. DOI: 10.52575/2712-7451-2022-41-3-590-607

Formation of Computational Linguistics Terminology

Olga N. Polshchykova
Belgorod National Research University,
85 Pobeda St, Belgorod 308015, Russia
E-mail: polshchikova@bsu.edu.ru

Abstract. An important role in the study of terminology is played by the study of its formation in the diachronic aspect. Despite the rapid development of the field of computational linguistics, there are practically no works on the analysis of the processes of evolution of the corresponding terminology. The purpose of this study is to study the processes of formation and formation of the terminology of computational linguistics, to identify the main periods of its development. The study was conducted on the basis of the method of diachronic analysis. To study the processes of formation and development of the terminology under study, factual material was used in the form of texts on the subject of computational linguistics, published in monographs, reference books, educational literature, scientific periodicals and collections, online electronic resources. As a result of the study, three main periods of the formation of the terminology of computational linguistics were identified. The first period is characterized by the formation of the terminology under study under the influence of linguistic terms. In the second period, the terminology of computational linguistics was formed under the influence of mathematical statistics terms. In the third period, the replenishment of the studied terminology is carried out by neural network terms. The results of the study showed that the terminology of computational

linguistics is in a state of continuous dynamics. The obtained results contribute to the study of the problems of scientific and technical terminology, substantiation of trends in its development.

Keywords: computational linguistics, terminology, periods of development, diachronic analysis, automatic natural language processing

For citation: Polshchykova O.N. 2022. Formation of Computational Linguistics Terminology. Issues in Journalism, Education, Linguistics, 41(3): 590–607 (in Russian). DOI: 10.52575/2712-7451-2022-41-3-590-607

Введение

Важную роль в исследовании терминологии играет изучение ее развития в диахроническом аспекте. Процессы становления и формирования терминологии компьютерной лингвистики неразрывно связаны с развитием научно-прикладной сферы, ориентированной на создание и применение компьютерных моделей естественного языка для решения широкого круга задач. С середины прошлого столетия до наших дней эти процессы сопровождают появление и апробацию теоретических результатов, издание научных трудов, а также разработку и внедрение программно-аппаратных средств в области систем машинного перевода, автоматической обработки естественно-языковых данных, информационно-поисковых систем, лингвистического обеспечения компьютерных систем и баз данных, электронных словарей.

Актуальность исследования обусловлена необходимостью проведения развернутого анализа истории формирования терминологии компьютерной лингвистики, выявления основных тенденций и периодизации её развития, а также недостаточным освещением данных вопросов в научной литературе.

Цель статьи – проанализировать процессы становления и формирования терминологии компьютерной лингвистики, обосновать, изучить и охарактеризовать основные периоды ее развития.

История становления и развития терминологии компьютерной лингвистики изложена на основе метода диахронического анализа. Для систематизации проработанных терминов в работе применен описательный метод исследования. В ходе проведения исследования использовались многочисленные источники фактического материала в виде текстов по тематике компьютерной лингвистики, размещенных в монографиях, справочниках, учебной литературе, научных периодических журналах и сборниках, сетевых электронных ресурсах.

Становление терминологии компьютерной лингвистки на этапе развития систем, основанных на правилах

Теоретические идеи создания универсального словаря на основе числового кодирования предложены еще в XVII веке Рене Декартом и Готфридом Лейбницем, а практические разработки для автоматизации перевода и обработки текстовой информации появились в тридцатых годах XX века до появления термина *компьютерная лингвистика* [Богданова, 2021]. Техническое устройство с амбициозным названием «Механический мозг», которое, в сущности, представляло собой машинный двуязычный словарь, в 1933 году запатентовал ученый из Франции Жорж Арцруни [История машинного …, 2019; Раренко, 2021]. В этом же году советский ученый Петр Троянский получил патент на «Машину для подбора и печатания слов при переводе с одного языка на другой или на несколько других одновременно» [Митренина, 2017]. Вышеуказанные изобретения обладали весьма скромными функциональными возможностями и не имели широкого распространения на практике.

Становление компьютерной лингвистики и соответствующей терминологии немыслимо без появления и совершенствования разработок в области компьютерной техники.

Первое техническое устройство, которое стало прообразом современных компьютеров, сконструировал в 1938 году инженер из Германии Конрад Цузе. Для обеспечения работы изобретения его создатель разработал первый в истории высокоуровневый язык программирования *Планкалкюль* (нем. *Plankalkül* — «запланированные вычисления») [Митренина, 2017]. Такая аппаратно-программная система последствии стала именоваться *электронно-вычислительной машиной* (*ЭВМ*).

Совершенствование ЭВМ в части увеличения объемов машинной памяти, улучшения вычислительных и других характеристик позволило сформировать техническую базу для создания разработок в сфере компьютерной лингвистики. Исследователи из университета города Джорджтаун совместно с компанией IBM разработали первую в мире систему, позволяющую с помощью ЭВМ переводить несложные типовые тексты с русского языка на английский. В 1954 году в Нью-Йорке состоялась «публичная демонстрация перевода с помощью вычислительной техники», которая осталась в истории как «Джорджтаунский эксперимент» [Дроздова, 2015, с. 156]. Для перевода экспериментальная программа использовала всего лишь 250 слов. Программное обеспечение, реализующее автоматизацию переводческих функций, стало называться *системами машинного перевода*.

В течение многих десятилетий до настоящего времени системы машинного перевода постоянно совершенствуются. Уже через год после Джорджтаунского эксперимента разработчиками Института точной механики и вычислительной техники Академии наук СССР создана система машинного перевода, в словаре которой содержалось 2300 слов [Митренина, 2017]. Представленные выше примеры и другие подобные разработки относятся к первому поколению компьютерных переводчиков. Их функционирование основано на том, что «каждому слову или речевому обороту в исходном тексте подбирается эквивалент на выходном языке, найденный в словаре» [Дроздова, 2015, с. 156]. Такая технология стала именоваться термином *пословный перевод*.

В этом же периоде (в пятидесятые годы прошлого столетия) было положено начало исследованиям и разработкам в области, именуемой *машинным переводом на основе правил* (англ. *Rule-based Machine Translation, RBMT*). При этом «предполагалось, что машина при переводе должна использовать те же методы, что и человек, опираясь на грамматики и словари» [Митренина, 2017, с. 8].

Проблемам алгоритмической обработки текстов уделялось значительное внимание в отечественной теории и практике. Велись разработки в этой области учеными Всероссийского института научной и технической информации Академии наук СССР. Результаты исследований с использованием соответствующей терминологии публиковались в академическом сборнике «Проблемы кибернетики». В конце 1950-х годов Л.И. Гутенмахер подготовил монографию «Информационно-поисковые системы», в которой освещались проблемы зарождающего в то время направления компьютерной лингвистики, именуемого термином *информационный поиск* [Блехман, 2012].

В следующем периоде, начиная с середины 1960-х годов, создаются системы машинного перевода второго поколения, которые вместо пословного перевода выполняют построение синтаксической структуры каждого предложения с использованием правил грамматики. После получения такой структуры «выполняется подстановка слов из словаря, т.е. синтез предложения на выходном языке» [Дроздова, 2015, с. 156]. Возможность создания подобных систем машинного перевода появилась благодаря появлению и развитию направления, именуемого термином *компьютерная морфология*. Применение компьютерной морфологии позволяет осуществлять анализ и синтез слов программными средствами. Выполнение морфологического анализа с помощью методов автоматической обработки естественно-языковых данных предполагает определение леммы, т.е. «постановку слова и словосочетания в каноническую форму» [Прикладная… 2016, с. 14]. Для обозначения этой процедуры стал использоваться термин *нормализация*. Создавались средства компьютерной морфологии и для решения обратной задачи, именуемой *порождением*

словоформы. Сущность этой задачи заключалась в том, чтобы поставить лемму в нужную грамматическую форму. Перед выполнением морфологического анализа компьютерными средствами текст разбивался на предложения, и в каждом выделялись отдельные единицы (слова, числа, формулы, знаки препинания и другие элементы), получившие название *токены*. Процесс такого разбиения стал именоваться термином *токенизация* [Прикладная… 2016].

При создании систем машинного перевода, информационного поиска и решении других задач, связанных с компьютерной обработкой текстовых данных возникла необходимость использования данных о структуре предложений, представленных на естественном языке. В целях получения этой информации начинает развиваться направление, называемое *автоматическим синтаксическим анализом*. В широком смысле автоматический анализ структуры любых текстовых данных, т.е. синтаксический анализ, стал именоваться термином *парсинг* (англ. *parsing* – «разбор»). В узком понимании этот термин «означает процедуру машинного анализа структуры текста на естественном языке, в том числе – структуры предложения» [Прикладная… 2016, с. 35].

Применение методов морфологического и синтаксического анализа позволило улучшить качество решения задач компьютерной лингвистики, «однако оставались трудности, связанные с семантикой» [Дроздова, 2015, с. 156]. По сути, на этапе становления компьютерной лингвистики многие исследователи пришли к выводу о необходимости создания некоторой искусственной системы, «мыслящей машины», «удовлетворительной модели мышления», т.е. такого интеллектуального устройства, что «человек, общающийся с ним заочно, не сумеет точно установить, с кем он имеет дело, – с другим человеком или автоматом» [Мельчук, 1999, с. 13]. Этот известный тезис получил наименование *теста Тьюринга*. Ещё в середине прошлого века Алан Тьюринг предсказывал возможность программирования работы машин так, что «шансы среднего человека установить присутствие машины через пять минут после того, как он начнет задавать вопросы, не поднимались бы выше 70%» [Тьюринг, 1960, с. 20].

Вдохновленный идеями создания единого формального математического языка, изложенными в 1930-х годах в трактате «Элементы математики» (автор Никола Бурбаки – коллективный псевдоним группы французских ученых), советский лингвист И.А. Мельчук в начале 1960-х годов начал трудиться над созданием концепции *семантического анализа* на основе формального описания смысла текста. Талантливому исследователю в сотрудничестве с коллегами А.К. Жолковским, Ю.Д. Апресяном, А.В. Гладким и Л.Н. Иорданской удалось разработать теорию, получившую название «*Смысл ⇔ Текст*». В этой теории естественный язык рассматривается как «особого рода преобразователь, выполняющий переработку заданных смыслов в соответствующие им тексты и заданных текстов в соответствующие им смыслы» [Мельчук, 1999, с. 10]. Цель теории состояла в обеспечении перехода от «текста на рассматриваемом языке к формальному описанию смысла этого текста, т.е. к его семантическому представлению (смысловой записи); разные, но интуитивно равнозначные (синонимичные тексты) должны получать одинаковые или хотя бы эквивалентные семантические представления, а текст, имеющий более одного смысла (омонимичный текст), – несколько разных семантических представлений» [Мельчук, 1999, с. 5]. Однако, по мнению автора, в полной мере обозначенная цель оказалась недостижимой, потому что в теории «имеют место огромные лакуны», «остались неразработанными многие важнейшие понятия» [Мельчук, 1999, с. 5]. Согласно данной теории, преобразование смысла в текст осуществляется на основе многоуровневой модели, в которой значительную роль играет семантический компонент. Центральную часть семантического компонента составляет специальный словарь – изобретение И.А. Мельчука, именуемое термином *толково-комбинаторный словарь*. Кроме того, автор рассматриваемой теории ввел понятие, обозначенное термином *лексическая функция* – «функция, аргумен-

тами которой являются некоторые слова или словосочетания данного языка, а значениями – множества слов и словосочетаний этого же языка» [Батура, 2016, с. 74].

Подход «Смысл ⇔ Текст» стал развитием учения Ноама Хомского под названием *генеративная лингвистика* [Chomsky, 1964]. В этом учении введено понятие *порождающей грамматики*, «цель которой – задавать (перечислять) все грамматически правильные (или все осмысленные) фразы привлекаемых языков» [Мельчук, 1999, с. 21].

Похожие на предложенные И.А. Мельчуком идеи создания многоуровневой модели алгоритмического перехода от текста к его смысловой записи и обратно были положены в основу исследований, проводимых в 1960-х – 1970-х годах рядом научных коллективов из Франции, США, Чехословакии, Великобритании, ФРГ [Мельчук, 1999, с. 16]. Опираясь на положения теории «Смысл ⇔ Текст», отечественный исследовательский коллектив под руководством Ю.Д. Апресяна в 1980-х годах начал создавать «компьютерную систему, реализующую формальную лингвистическую модель и способную работать с естественным языком во всем его объеме» [Апресян, 1992, с. 3]. Для наименования этой разработки использовался термин *лингвистический процессор*. Функционально лингвистический процессор проектировался в виде многоуровневого преобразователя. Каждый уровень (морфологический, синтаксический и семантический) обслуживался «соответствующим компонентом модели – массивом правил и определенным словарем или словарями» [Апресян, 1992, с. 8]. Текстовое предложение на каждом уровне представлялось с помощью определенной структуры (формального образа). Наиболее сложным оказался формальный образ, именуемый *семантической структурой*, под которой понималось «дерево зависимостей, в узлах которого стоят либо предметные имена, либо слова универсального семантического языка (например, имена таблиц, в которых сосредоточены сведения о данной предметной области, атрибуты таблиц, операторные символы), а дуги соответствуют универсальным отношениям семантического подчинения, таким, как аргументное, атрибутивное, конъюнкция, дизъюнкция, равенство, неравенство, больше, меньше, принадлежит, не принадлежит и т.п.» [Апресян, 1992, с. 6].

Сфера применения лингвистического процессора предполагалась в нескольких прикладных системах. Одна из них именовалась термином *система общения с компьютером*. Её назначение виделось в том, чтобы «организовать максимально дружественный интерфейс пользователя с компьютером» [Апресян, 1992, с. 9]. Данная система фактически стала прообразом современных *чат-ботов* и *вопросно-ответных систем*. Другой прикладной областью применения лингвистического процессора явились *системы машинного перевода научно-технических текстов*, позволявшие выполнять деловую документацию с иностранных языков на русских и наоборот. Кроме того, предполагалось, что лингвистический процессор будет полезен для решения задач, связанных с автоматическим пополнением баз данных непосредственно по текстам, а также задач в области, получившей в 1980-х годах название *планирования текста* (англ. *text planning*) [Апресян, 1992, с. 9]. При выполнении планирования текста реализуются активные, т.е. «синтезирующие или текстопорождающие» возможности лингвистического процессора, который «сначала преобразует концептуальную структуру в семантическую структуру будущего текста на естественном языке, а затем через ряд промежуточных этапов, превращает её в последовательность предложений, образующих связный рассказ на заданную пользователем тему» [Апресян, 1992, с. 9 – 10]. В перспективе допускалось использование лингвистического процессора в партнерских системах овладения иностранным языком, получивших название *помощник учителя*, а также в разного рода разработках, именуемых *компьютерными словарями*.

Лингвистический процессор задумывался как «фундаментальная разработка в области моделирования понимания и производства текстов» на естественном языке. Результатом работ над его созданием стала система «двунаправленного» перевода *ЭТАП* [Апресян, 1992, с. 13]. Это название образовано из словосочетания «электротехнический автомати-

ческий перевод». Автором термина стала Татьяна Коровина, программист московского института «Информэлектро» [Митренина, 2017]. Известны различные версии этой системы. Наиболее усовершенствованная версия ЭТАП-3 – это «полифункциональная система обработки текста на естественном языке», которая является прикладной разработкой лаборатории компьютерной лингвистики Института проблем передачи информации имени А.А. Харкевича Российской академии наук [Богуславский и др., 2000; Многоцелевой лингвистический… 2022]. Её создатели стремились с помощью компьютерных средств реализовать на практике направление, именуемое *лингвистическим моделированием естественного языка*. Кроме системы машинного перевода на основе системы ЭТАП-3 построен *универсальный сетевой язык* (англ. *Universal Networking Language, UNL*) – семантический формализм, который предназначен для выражения содержащейся в текстах важнейшей информации [Богуславский и др., 2020], разработан синтаксически размеченный корпус русских текстов *СинТагРус* [Синтаксически размеченный… 2003-2022], а также выполнена разметка синтаксического корпуса проекта «Национальный корпус русского языка» [Национальный… 2003–2022]. Фактически, все эти результаты получены отечественными группами лингвистов, математиков и программистов на основе теории «Смысл ⇔ Текст» И.А. Мельчука и лингвистического обеспечения Ю.Д. Апресяна.

Следует признать, что масштабность и глубина исследований, выполненных в процессе разработки теории «Смысл ⇔ Текст», позволили создать на её базе ряд прикладных систем, но не принесли значительного прорыва в понимании компьютером семантики текстов. Анализируя функциональность своих моделей, И.А. Мельчук отмечал исключительную важность их соотнесения «с психической, физиологической, нейрологической реальностью говорения и понимания». В то же время «значительная сложность соответствующей проблематики при её относительно слабой теоретической разработанности» обусловили отказ «от моделирования «настоящих» процессов функционирования языка» [Мельчук, 1999, с. 15]. С использованием терминологии Н. Хомского, справедливо утверждалось, теория «Смысл ⇔ Текст» дает возможность моделировать «*linguistic competence* говорящих (их языковые познания), а не *linguistic performance*, т.е. не действительные процессы применения говорящими и слушающими своих познаний в актах речевого общения» [Мельчук, 1999, с. 15].

К началу 1990-х годов появились различные интерактивные системы машинного перевода, функционирование которых осуществляется с участием человека. Системы, предполагающие привлечение человека для редактирования переведенного компьютером текста, получили название *систем машинного перевода с постредактированием*. Программные средства, требующие участия человека в подготовке текста для последующего компьютерного перевода, стали именоваться *системами машинного перевода с предредактированием*. *Системами частично автоматизированного перевода* стали называться системы, в которых осуществляется взаимодействие человека и компьютера в процессе перевода. Наконец, системы, работа которых основана на сочетании различных режимов участия человека в процессе перевода, получили название *смешанных систем машинного перевода* [Дроздова, 2015].

В начале 2000-х годов использовались средства компьютерной лингвистики, функционирующие на основе правил. К тому времени различали два подтипа систем перевода, основанных на правилах. К первому подтипу относились средства, обозначенные термином *трансферные системы*, второй подтип составили средства, именуемые *интерлингвистическими системами*. Функционирование трансферной системы в наибольшей мере соответствует логической схеме работы человека-переводчика. Такая система «анализирует исходное предложение, переводит его слова, выясняет их роли, а затем с помощью грамматики собирает новое предложение на конечном языке» [Прикладная… 2016, с. 160]. Системы машинного перевода, получившие название *интерлингвистических систем*, основывались на использовании языка-посредника, который именовался *интерлингвой*. При

этом предполагалось, что в качестве языка-посредника следовало понимать «не язык в привычном для нас смысле, а некоторое формальное представление смысла человеческой речи и мыслей» [Прикладная… 2016, с. 160]. Сейчас такие системы практически не используются, т.к. «создать полноценный универсальный вспомогательный язык ученым пока не удалось» [Прикладная… 2016, с. 160].

Реализации идеи метаязыка, «на котором желательно представлять содержание любого текста», посвящены работы Н.Н. Леонтьевой, которая исследовала особую роль *семантического компонента* «как наиболее содержательного участка компьютерного понимания» [Леонтьева, 2006, с. 3]. В своих исследованиях она стремилась пойти «по пути создания языка, пригодного для решения наиболее сложных информационных задач, к которым относятся накопление и классификация информации, ее преобразование и сжатие, фактографический поиск, документальный поиск, автоматический перевод с одного языка на другой, в частности на тот же самый, но с заданными критериями потерь» [Леонтьева, 2006, с. 105]. Такой многофункциональный метаязык именовался *информационным языком-посредником*. В конце 1990-х годов Н.Н. Леонтьева руководила созданием экспериментальной системы *ПОЛИТЕКСТ*, которая проектировалась не только в целях автоматического анализа политических текстов, но и как «единая многоканальная система понимания текстов и автоматического извлечения информации из текстов» [Леонтьева, 2006, с. 14]. Особое внимание обращалось на то, что текст «может быть понят с разной степенью подробности, с разными оценками, с разными фокусами внимания, с точки зрения разных предметных областей» [Леонтьева, 2006, с. 14]. Такая идея о том, что «один и тот же текст допускает разные результаты понимания в зависимости от разных условий» получила название *мягкого понимания текста* [Леонтьева, 2006, с. 4].

На основе результатов проекта ПОЛИТЕКСТ в 2001 году была создана пробная версия системы *ДИАЛИНГ*, предназначавшейся для машинного перевода с русского языка на английский. В его проектировании принял участие А.В. Сокирко, который затем вместе с рядом других выпускников факультета лингвистики Российского государственного гуманитарного университета вошел в состав рабочей группы *АОТ*, занимавшейся разработкой программного обеспечения в области под названием *автоматическая обработка текста* [Автоматическая обработка… 2003]. В рамках реализации одноименного проекта были разработаны модули для проведения графематического, морфологического, синтаксического и семантического анализа, а также *синтаксический анализатор именных групп* [Сокирко, Толдова, 2005; Прикладная… 2016].

Формирование терминологии компьютерной лингвистики на этапе применения статистических методов

Одной из лучших систем машинного перевода в XX веке, по мнению экспертов, была американская разработка под названием *Systran* [Systran translate], англо-русская версия которой создана ещё в 1973 году. В начале 1990-х годов она приобрела популярность и активно использовалась в браузерах «Yahoo!» и «Google». Однако в 2004 году компания «Google» отказалась от этой системы и приступила к разработке собственного принципиально нового машинного переводчика. Было принято «решение разработать систему перевода на основе статистики без использования грамматики и словарей» [Митренина, 2017, с. 9]. Функционирование системы предполагало использование объемного текстового корпуса, содержащего наборы предложений на исходном языке и соответствующие им переведенные предложения. Для получения результата компьютер «анализирует, какие фрагменты предложения (например, биграммы и триграммы) часто встречаются вместе в оригинале и в переводе, а затем, получив новое предложение, строит для него (только на основе статистики, без использования лингвистических знаний) наиболее вероятное предложение-перевод» [Митренина, 2017, с. 9]. Эта технология получила название *статистического машинного перевода* (англ. *Statistical Machine Translation, SMT*).

Статистические средства начали применять и в системах, основанных на правилах. В целях улучшения качества машинного перевода RBMT-переводчики стали дополняться элементами, именуемые термином *базы памяти переводов*. В них заранее сохранялись текстовые фрагменты и их выполненные человеком переводы, которые могли «вставляться в текст автоматически или по указанию пользователя» [Прикладная и …, 2016, с. 162]. Такая технология получила название *гибридного машинного перевода* (англ. *Hybrid Machine Translation, HMT*). В её основе лежали методы машинного перевода по правилам и статистические методы, при которых оказывались «задействованными двуязычные базы часто встречающихся предложений» [Раренко, 2021, с. 75]. Такая технология нашла применение, например, в разработках российской компании *PROMT* (аббревиатура «*PROject Machine Translation*») [История… 2019].

Потенциал статистических методов получил применение в компьютерной морфологии. В частности, при анализе слов выполнялась процедура, именуемая *автоматической частеречной разметкой* (англ. *part of speech tagging, POS-tagging*). В процессе данной разметки каждое слово в предложении получало соответствующую части речи метку, именуемую термином *тег* (англ. *tag* – «бирка»). При этом обращалось внимание на контекст, в котором употреблялось каждое слово, т.е. на вероятность следования определенного тега после той или иной части речи, например, учитывалась существующая статистическая закономерность, состоящая в том, что «в русском языке после предлога будет идти, скорее всего, существительное, а не глагол» [Прикладная и …, 2016, с. 28].

В начале 2000-х годов статистические данные применялись и для выполнения синтаксического анализа. При этом рассматривались сначала наиболее вероятные варианты синтаксических конструкций. В таких методах, получивших название *синтаксического парсинга*, использовались размеченные текстовые коллекции для оценивания того, «как часто в реальных текстах реализуется то или иное синтаксическое правило» [Прикладная… 2016, с. 47].

Методы морфологического и синтаксического анализа, основанные на обработке статистических данных, нашли применение в компьютерных системах информационного поиска. В основе поиска начали использовать базу данных, содержащую «сведения о словах и их позициях в документах» (Интернет-страницах), которую стали именовать термином *индекс* [Батура, 2016, с. 46]. В целях повышения скорости поиска и уменьшения Интернет-трафика в процессе передачи информации по сети было предложено загрузку часто запрашиваемых документов осуществлять не с сервера-источника, а сохранять их на промежуточных серверах или компьютере пользователя. Эта процедура получила называние *кэширование интернет-страниц*. Для удобства получения результатов необходимо, что наиболее подходящие запросу документы располагались как можно ближе к началу поисковой выдачи. Такой порядок отображения результатов поиска обеспечивался применением процедуры, именуемой термином *ранжирование*. Для обозначения соответствия найденного документа «смысловому содержанию информационного запроса» стал применяться термин *релевантность* [Батура, 2016, с. 48]. Использование статистических алгоритмов поиска и отбора, в частности, учет «частот лемм вместо частот слов» дало возможность получать «больший вес для релевантных документов» и помещать их в массив отобранных результатов [Батура, 2016, с. 50]. При этом применение морфологического и синтаксического анализа позволило увеличить показатели, именуемые терминами *полнота поиска* («доля найденных документов в числе всех релевантных документов в коллекции») и *точность поиска* («доля релевантных документов в числе всех найденных») [Прикладная… 2016, с. 198].

В качестве основных компонентов поисковых систем стали использоваться программные средства, именуемые *базовым поиском* и *метапоиском*. Базовый поиск представляет собой программу, служащую для обеспечения поиска по определенной части индекса. Под метапоиском стала пониматься программа, «обеспечивающая прием и разбор

(например, лингвистический) поисковых запросов; выбор базовых поисков и передачу им запросов; кэширование страниц с результатами поиска; агрегацию и ранжирование найденных документов» [Батура, 2016, с. 46-47]. Опыт разработки и применения поисковых систем со временем показал важность применения семантического анализа текстов. Действительно, пользователей интересует, прежде всего, «нахождение тех документов, в которых описываются конкретные свойства запрашиваемого предмета или явления, то есть так называемый *поиск по смыслу*», а не *поиск по ключевым словам* [Батура, 2016, с. 51].

Многообещающие возможности статистических методов способствовали переходу крупнейших компаний к разработке SMT-продуктов. Статистические модели нашли применение в 2007 году в системе *Google Translate,* и в 2011 году в программном обеспечении *Яндекс-переводчик* [Тихонов, 2017, с. 227]. Наиболее перспективные инструменты в те годы анализировали «статистические отношения между словами вместо того, чтобы использовать глубокую систему логических правил [Хобсон и др., 2020, с. 41]. К тому времени в средствах компьютерной лингвистики начали широко применяться так называемые *триграммные скрытые марковские модели*, позволяющие «предсказывать вероятность следующих элементов цепи, анализируя не всю цепочку, а только один или несколько последних ее элементов» [Прикладная и …, 2016, с. 116]. Согласно этим моделям для обработки естественно-языковой информации излишне «использовать большой предшествующий контекст», поэтому «стали брать контекст из двух предыдущих слов» [Митренина, 2019, с. 403-404]. Так, статистические исследования выдающего русского математика А.А. Маркова, выполненные ещё в начале XX века, дали возможность, спустя почти сто лет, улучшить морфологическую разметку текстов, а также повысить качество систем, функционирующих в области, именуемой *автоматическим распознаванием речи*.

Решение задач компьютерной лингвистики тесно связано с направлением, именуемым термином *автоматическая обработка естественного языка* (англ. *Natural Language Processing, NLP*). Последовательность выполняемых этапов обработки естественного языка называют *конвейером* (англ. *pipeline*). Первоначальным этапом конвейера NLP является токенизация, для выполнения которой используются программные средства, именуемые *токенизаторами*. С помощью этих средств разбиваются «неструктурированные данные, текст на естественном языке, на фрагменты информации, которые можно считать отдельными элементами» [Хобсон и др., 2020, с. 71]. Набор всех допустимых токенов в текстовом корпусе именуется термином *словарь* [Хобсон и др., 2020, с. 71]. При составлении словаря могут приниматься во внимание как отдельные слова, так и наборы стоящих рядом слов, например, *биграммы* или *триграммы* [Большакова и др., 2017, с. 26]. С целью «уменьшения сложности вычислений при извлечении информации из текста» в конвейере NLP исключаются так называемые *стоп-слова*, т.е. «распространенные слова на любом языке, которые встречаются очень часто, но несут в себе гораздо меньше содержательной информации о смысле фразы» [Хобсон и др., 2020, с. 91]. Следующим этапом выступает выполнение процедуры, именуемой *стеммингом*. Она состоит в определении «основы слова путем отбрасывания окончаний из известного набора (возможно, псевдоосновы за счёт отбрасывания псевдоокончаний)» [Большакова и др., 2017, с. 44]. Для более тонкой нормализации слова «до его семантического корня – леммы» выполняется этап, обозначенный термином *лемматизация* [Хобсон и др., 2020, с. 100]. Компьютерная программа, именуемая *лемматизатором*, «использует базу знаний синонимов и окончаний слов, чтобы объединять в один токен только близкие по смыслу слова» [Хобсон и др., 2020, с. 100]. Исключение стоп-слов, выполнение стемминга и лемматизации позволяет уменьшить размерность языковой модели, которая в результате «становится более общей, но также и менее точной» [Хобсон и др., 2020, с. 100].

В целях числового определения смысла отдельных слов, текстовых фрагментов или целых документов конвейер NLP предусматривает выполнение важнейшего этапа, имену-

емого *векторизацией*. В процессе автоматической обработки естественного языка используется та или иная признаковая модель, на основе которой из текста извлекаются наборы числовых данных в виде структуры, именуемой терминами *вектор* или *эмбеддинг* (англ. *embedding* – «вложение»). Простейший метод получения эмбеддингов использует бинарную модель векторов, соответствующую *унитарному кодированию* (англ. *one-hot encoding* – «горячее, однократное кодирование»). После подсчета токенов каждый документ «может быть представлен в виде вектора, последовательности целых чисел для каждого слова или токена в этом документе» [Хобсон и др., 2020, с. 54]. Упрощенным признаковым представлением текста является модель, именуемая термином *мешок слов* (англ. *bag of words, BOW*). Получаемая BOW-структура имеет вид «матрицы, каждая строка в которой соответствует отдельному документу или тексту, а каждый столбец – определенному слову», каждая ячейка которой «на пересечении строки и столбца содержит количество вхождений слова в соответствующий документ» [Хмельков, 2015].

Чтобы научить компьютер лучше понимать смысл текста, исследователи предпринимали различные попытки усложнить векторизацию вычислениями дополнительной информации, позволяющей глубже постигать семантику естественно-языковых данных. В качестве параметра оценки важности находящихся в тексте слов было предложено использовать величину, соответствующую частоте появления того или иного слова в документе, именуемую *частотностью*. Так в средствах NLP стали учитываться факты, показывающие, «насколько часто каждое слово языка (термин) встречается в корпусе и насколько важно его появление в конкретном тексте» [Крылов, 2019]. Для обозначения числовой структуры, отражающей «частоту вхождения слов, а не порядок их расположения» [Хобсон и др., 2020, с. 77], появился термин *вектор частотности слов*. Этот вектор, длина которого «равна длине словаря (числу отслеживаемых уникальных токенов)», начали использовать «для представления всего документа или предложения» [Хобсон и др., 2020, с. 77]. Распространенным показателем, отражающим степень важности слов, содержащихся в естественно-языковом документе, стала величина, обозначаемая аббревиатурой *TF-IDF* (англ. *Term Frequency Inverse Document Frequency* – «частотность терма умножить на обратную частотность документа») [Хобсон и др., 2020, с. 112].

Дальнейшие исследования в сфере NLP позволили разработать метод для определения смысла, содержащегося в словосочетаниях. Этот метод получил название *латентно-семантический анализ* (англ. Latent Semantic Analysis, LSA). Применение LSA дало возможность представлять «в виде векторов смысл не только слов, но и целых документов» [Хобсон и др., 2020, с. 141]. С помощью латентно-семантического анализа можно выполнять анализ таблицы векторов TF-IDF и группировать слова по темам. Использование векторов тем позволило «сравнивать значения слов, документов, высказываний и корпусов», кроме того, «находить документы, соответствующие запросу, а не просто хорошо подходящие к статистике слов» [Хобсон и др., 2020, с. 195], т.е. решать прикладные задачи, именуемые термином *семантический поиск*.

Период пополнения терминологии компьютерной лингвистики нейросетевыми терминами

Применение статистических методов для выполнения семантического анализа давало некоторые успехи в решении проблемы «неоднозначности естественного языка – того факта, что слова и фразы часто имеют несколько значений и интерпретаций» [Хобсон и др., 2020, с. 41]. Однако по мнению экспертов и простых пользователей качество статистического машинного перевода оставляло желать лучшего, ведь методы, основанные на статистике, «в первую очередь направлены на преодоление комбинаторных сложностей при обработке огромных массивов текстов и очень слабо учитывают внутренние структуры языка (по существу все происходит без учета лингвистики). Отсюда – существенные

затруднения в тех вопросах, где требуется понимание смысла текста» [Тихонов, 2017, с. 229]. Кроме того, при использовании вероятностной триграммной модели языка «машина использует всего один признак: частоту появления слова после двух других слов» [Митренина, 2019, с. 405]. В этих случаях не используются возможности «задавать машине другие признаки и решать другие задачи, используя математические методы» [Митренина, 2019, с. 405]. Мощнейшим аппаратом, позволяющим обрабатывать сложные данные, оказались искусственные нейронные сети, которые стали применяться «как попытка смоделировать на компьютере работу человеческого мозга. Именно они дают лучшие результаты в тех случаях, когда признаков так много, что непонятно, какие из них влияют на результат» [Митренина, 2019, с. 405]. Интерес к нейросетевым технологиям растет с каждым годом, они очень востребованы во многих предметных областях [Polshchykov et. al., 2017; Polshchykov et. al., 2019; Polshchykov et. al., 2020; Агузумцян и др., 2021; Alghazali et. al., 2021; Mahdi et. al., 2021; Velikanova, 2021], поэтому вполне оправдано их применение и в сфере компьютерной лингвистики.

Инновационный нейросетевой способ получения векторных представлений слов предложил чешский исследователь Томаш Миколов. Его идея состояла в предсказании слова на основе имеющейся информации о соседних словах с помощью довольно простой нейронной сети, обученной на большом корпусе немаркированных текстов. В 2013 году было разработано соответствующее программное обеспечение с «говорящим» названием *Word2vec*, реализующее преобразование слов в числовые векторы сравнительно небольшой размерности [Mikolov et. al., 2013]. Эксперименты показали, что ответы на вопросы средствами Word2vec давались многократно точнее, чем LSA. Специалисты в сфере NLP отмечали, что «Миколов и его соавторы добились точности ответов всего в 40 %. Но тогда, в 2013 году, это значительно превосходило любой другой подход к семантическим умозаключениям» [Хобсон и др., 2020, с. 239]. Вскоре после появления технология Word2vec была усовершенствована: в целях повышения скорости обучения нейросети и экономии компьютерных ресурсов разработан метод, основанный на «оптимизации *глобальных векторов* (англ. *global vectors*) совместной встречаемости слов» [Хобсон и др., 2020, с. 255]. Он получил соответствующее название *GloVe*.

Направление исследований и разработок, ориентированное на применение искусственных нейронных сетей для решения задач машинного перевода, получило название *нейросетевой* или *нейронный машинный перевод* (англ. *Neutral Machine Translation, NMT*). В 2016 году компания Google предоставила пользователям первую нейросетевую систему машинного перевода, работа которой показала «значительное улучшение качества переведенных текстов, и это направление, как и другие способы компьютерной обработки языка с помощью нейронных сетей, сейчас развивается наиболее активно» [Хобсон и др., 2020, с. 9].

Чтобы лучше уловить смысл текста, необходимо в процессе анализа иметь возможность учитывать полученную ранее информацию. Функция запоминания прошлых слов в текстовой последовательности реализована в NLP-системах, которые основаны на применении аппарата, называемого *рекуррентными нейронными сетями* (англ. *Recurrent Neural Networks, RNN*). Рекуррентная нейросеть обладает обратными связями и принимает на вход последовательность естественно-языковых данных (слов), при этом имеет значение порядок подачи этих данных. Однако веса нейронов в рекуррентной нейросети затухают слишком быстро по мере просмотра предложения, поэтому при использовании RNN возникают сложности в обнаружении взаимосвязей между достаточно далеко отстоящими друг от друга словами. Для «запоминания прошлого в масштабах всего предложения» [Хобсон и др., 2020, с. 330] разработаны *нейронные сети с долгой краткосрочной памятью* (англ. *Long Short-Term Memory, LSTM*). Нейросети LSTM, которые появились как разновидность RNN, демонстрировали настолько хорошие результаты, что «заменили рекуррентные нейронные сети практически во всех приложениях NLP» [Хобсон и др., 2020,

с. 330]. Применение LSTM-моделей позволило «выявлять очевидные для человека (и обрабатываемые на подсознательном уровне) языковые закономерности, которые позволяют не только точнее классифицировать примеры данных, но и генерировать на их основе совершенно новый текст» [Хобсон и др., 2020, с. 331].

Стремление создавать краткие изложения объемных статей, находить наиболее релевантную информацию в больших документах привело к появлению в сфере NLP очередного изобретения, именуемого *механизмом внимания* (англ. *attention mechanism*) [Vaswani et. al., 2017]. С его использованием в 2017 году создана архитектура нейронных сетей, получившая название *трансформер*. В нейросети-трансформере механизм внимания реализован в виде дополнительного слоя нейронов, который «делает возможной прямую связь между выходным и входным сигналами за счет выбора релевантных фрагментов входного сигнала» [Хобсон и др., 2020, с. 395]. В 2018 году появилась усовершенствованная версия трансформера, именуемая англоязычной аббревиатурой *BERT* (*Bidirectional Encoder Representations from Transformers*) [Devlin et. al., 2018]. Технология BERT для автоматической обработки естественно-языковых данных предполагает использование трансформеров, предварительно обученных на огромных текстовых массивах. Такие системы стали называться *предобученными нейронными сетями* (англ. *pre-trained neural networks*). Для решения конкретной NLP-задачи следует завершить настройку нейросети с помощью соответствующих специфических текстовых наборов. Развитие современных систем компьютерной лингвистики привело к разработке в 2020 году крупнейшей модели обработки естественно-языковых данных, именуемой *GPT-3* (*Generative Pre-trained Transformer 3*) [Brown et. al., 2020]. Конструктивно сложнейший NLP-трансформер третьего поколения многократно превосходит предыдущие версии по числу используемых для векторизации параметров. Обучение GPT-3 производилось на специально спроектированном суперкомпьютере с использованием десятков терабайт сжатых текстовых данных. Благодаря этому новый трансформер приобрел уникальный функционал, например, если ввести начало текста, то программа сгенерирует его наиболее вероятное продолжение. Алгоритмы GPT-3, работающие по принципу, именуемому термином *автодополнение*, способны создавать не только тексты различных историй, пресс-релизов, технической документации, но и писать песни, а также генерировать программный код [Нейросеть GPT-3…, 2020].

Современные нейросетевые средства компьютерной лингвистики находят применение для решения широкого спектра прикладных задач. Хотя нейронные сети изначально проектировались, «чтобы научить машину оценивать количественно входные данные», но «их поле деятельности выросло с тех пор от классификации и регрессии (анализа тем, тональностей) до возможности настоящей генерации нового текста на основе входных данных, ранее незнакомых моделей: перевода новых фраз на другой язык, генерации ответов на новые вопросы (чат-боты) и даже генерации нового текста в стиле конкретного автора» [Хобсон и др., 2020, с. 270]. К задачам, решаемым средствами NLP относится, например, распознавание и извлечение данных о каких-либо людях, организациях, географических локациях, называемых термином *именованные сущности*. При этом важное значение имеет корректное отнесение к именованным сущностям определенных именных групп или местоимений. Эта функция называется *автоматическим разрешением анафоры и кореферентности*. Стремительно развивается прикладное направление компьютерной лингвистики, которое ориентировано на обеспечение живого естественно-языкового диалога человека с виртуальным ассистентом, именуемым *диалоговой системой* или *чат-ботом*. С этим направлением тесно связано решение задачи, именуемой *пониманием естественного языка* (англ. *Natural Language Understanding, NLU*). Повышение уровня «интеллектуальности» диалоговых систем обеспечивается на основе их интеграции со средствами, именуемыми *вопросно-ответными системами*, цель которых – «не убедить пользователя в

своей «человечности», а предоставить максимально точный ответ на вопрос, заданный на естественном языке» [Прикладная… 2016, с. 244]. Чтобы NLP-система была максимально полезной человеку в предоставлении нужной информации, она наделяется способностью к выполнению функции краткого изложения содержания текста, называемой *автоматическим реферированием*. Довольно востребованными сегодня являются системы компьютерной лингвистики, дающие возможность оценить эмоциональное восприятие людьми той или иной информации, т.е. выполняющие функцию, называемую термином *анализ тональности*, или *сентимент-анализ*.

Несмотря на достигнутые значительные успехи в сфере автоматической обработки естественного языка, «компьютеры пока неспособны решать большинство практических задач NLP, таких как разговор и понимание прочитанной информации, так же точно и качественно, как это делают люди» [Хобсон и др., 2020, с. 41]. Многие специалисты считают, что «нейросеть умело бросает пыль в глаза, выдавая текст, похожий на человеческий, но даже таким примерам не хватает глубины проработки: это больше похоже на копирование и вставку готовой информации, нежели на осмысленный подход» [Нейросеть… 2020]. Однако средства NLP продолжают непрерывно совершенствоваться, а вместе с ними развивается и терминология компьютерной лингвистики.

Заключение

Выполненный диахронический анализ позволил описать процесс становления и формирования терминологии компьютерной лингвистики, выявить периоды ее развития, обосновать эволюцию системы понятий изучаемой предметной сферы. Проведенное исследование дало возможность выделить три основных периода формирования терминологии компьютерной лингвистики.

Первый период (1950–1990 гг.) соответствует этапу развития средств компьютерной лингвистики, основанных на правилах. Он характеризуется формированием исследуемой терминологии под влиянием лингвистических терминов. В этом периоде появляются термины *компьютерная морфология*, *интерлингвистический перевод*, *автоматический синтаксический анализ*, *порождающая грамматика*, *лингвистический процессор* и др.

Второй период (с 2000-х годов до начала 2010-х годов) соответствует этапу развития средств компьютерной лингвистики, основанных на статистических методах. Исследуемая терминология формировалась под влиянием терминов математической статистики. Во втором периоде вошли в употребление термины *статистический машинный перевод*, *вектор частотности слов*, *точность* и *полнота поиска*, *обратная частотность документа* и др.

В третьем периоде (с начала 2010-х годов до настоящего времени) в связи с доминированием средств компьютерной лингвистики, основанных на искусственных нейронных сетях, пополнение исследуемой терминологии осуществляется нейросетевыми терминами. Примерами терминов, сформированных в этом периоде, являются *нейросетевой машинный перевод*, *LSTM-модель*, *предобученный NLP-трансформер*, *BERT* и др.

Результаты исследования показали, что терминология компьютерной лингвистики находится в состоянии непрерывной динамики вследствие стремительного развития соответствующих научных методов и прикладных технологий, появления новых понятий, обусловленного процессами усовершенствования аппаратных и программных средств в сфере обработки естественно-языковых данных.

Полученные результаты вносят вклад в изучение проблем научно-технической терминологии, обоснование тенденций её развития.

Список источников

Автоматическая обработка текста. 2003. URL: http://www.aot.ru/history.html (дата обращения: 10.02.2022).

Богданова О. 2021. От Декарта до Google Translate. Удивительная история машинного перевода. Teletype, 18 апреля 2021 года. URL: https://teletype.in/@iambocca/machine-translation (дата обращения: 10.02.2022).

История машинного перевода: от гипотез Лейбница и Декарта – до мобильных приложений и облачных сервисов. ПРОМТ, 21 марта 2019 года. URL: https://www.promt.ru/press/blog/istoriya-mashinnogo-perevoda-ot-gipotez-leybnitsa-i-dekarta-do-mobilnykh-prilozheniy-i-oblachnykh-se/ (дата обращения: 10.02.2022).

Крылов В. 2019. Что такое эмбеддинги и как они помогают искусственному интеллекту понять мир людей. Наука и жизнь, 17 апреля 2019 года. URL: https://www.nkj.ru/open/36052/ (дата обращения: 10.02.2022).

Многоцелевой лингвистический процессор ЭТАП-3. 2022. Российская Академия наук. Институт проблем передачи информации им. А.А. Харкевича. URL: http://iitp.ru/ru/science/works/452.htm (дата обращения: 10.02.2022).

Национальный корпус русского языка. 2003-2022. URL: https://ruscorpora.ru/new/index.html (дата обращения: 10.02.2022).

Нейросеть GPT-3 от OpenAI пишет стихи, музыку и код. Почему она пока далека от настоящего ИИ, но способна поменять мир. Компьютерная лингвистика, анализ текстов, корпусная лингвистика, 8 августа 2020 года. URL: https://ai-news.ru/2020/08/nejroset_gpt_3_ot_openai_pishet_stihi_muzyku_i_kod_pochemu_ona_poka_dalek.html (дата обращения: 10.02.2022).

Синтаксически размеченный корпус русского языка: информация для пользователей. 2003-2022. Национальный корпус русского языка. URL: https://ruscorpora.ru/new/instruction-syntax.html (дата обращения: 10.02.2022).

Хмельков И. 2015. Мешок слов и сентимент-анализ на R. Хабр, 7 апреля 2015 года. URL: https://habr.com/ru/post/255143/ (дата обращения: 10.02.2022).

Systran translate. URL: https://www.systran.net/en/translate/ (accessed: February 10, 2022).

Список литературы

Агузумцян Р.В., Великанова (Герасимова) А.С., Польщиков К.А., Игитян Е.В., Лихошерстов Р.В. 2021. О применении интеллектуальных технологий обработки естественного языка и средств виртуальной реальности для поддержки принятия решений при подборе исполнителей проектов. *Экономика. Информатика*, 48(2): 392–404. DOI: 10.52575/2687-0932-2021-48-2-392-404

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. 1992. Лингвистический процессор для сложных информационных систем. Под ред. Л.П. Крысина. М., Наука, 256 с.

Батура Т.В. 2016. Математическая лингвистика и автоматическая обработка текстов на естественном языке. Новосибирск, РИЦ Новосибирский национальный исследовательский государственный университет, 166 с.

Богуславский И.М., Иомдин Л.Л., Крейдлин Л.Г., Фрид Н.Е., Сагалова И.Л., Сизов В.Г. 2000. Модуль универсального сетевого языка в составе системы ЭТАП-3[1]. В кн.: Сборник 2000. URL: https://www.dialog-21.ru/digest/2000/articles/boguslavsk_i_m/ (дата обращения: 10.02.2022).

Блехман М.С. 2012. Краткая историческая справка о зарождении и успешном развитии компьютерной лингвистики в СССР. *Петербургская библиотечная школа*, 2(39): 4–6. URL: http://www.rasl.ru/e_editions/pbsh_2012-2-39.pdf (дата обращения: 10.02.2022).

Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. 2017. Автоматическая обработка текстов на естественном языке и анализ данных. М., Изд-во НИУ ВШЭ, 269 с.

Дроздова К.А. 2015. Машинный перевод: история, классификация, методы. В кн.: Филологические науки в России и за рубежом. Материалы III международной научной конференции, Санкт-Петербург, 20–23 июля 2015 г. СПб., Свое издательство: 139–141. URL: https://moluch.ru/conf/phil/archive/138/8497/ (дата обращения: 08.04.2022).

Мельчук И.А. 1999. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М., Школа «Языки русской культуры», 346 с.

Митренина О.В. 2017. Назад, в 47-й: к 70-летию машинного перевода как научного направления. Вестник Новосибирского государственного университета. *Лингвистика и межкультурная коммуникация*, 15 (3): 5–12. DOI: 10.25205/1818-7935-2017-15-3-5-12

Митренина О.В. 2019. Нейронные сети и компьютерная обработка языка. *Journal of Applied Linguistics and Lexicography*, 1 (2): 399–408.

Леонтьева Н.Н. 2006. Автоматическое понимание текстов: системы, модели, ресурсы. М., Академия, 304 с.

Прикладная и компьютерная лингвистика. 2016. Под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. М., Ленанд, 320 с.

Раренко М.Б. 2021. Машинный перевод: от перевода «по правилам» к нейронному переводу. *Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия: Языкознание. Реферативный журнал*, 3: 70–79. DOI: 10.31249/ling/2021.03.05

Сокирко А.В., Толдова С.Ю. 2005. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). В кн.: Интернет-математика 2005: автоматическая обработка веб-данных. М.: 80-94. URL: http://hdl.handle.net/10995/1391 (дата обращения: 10.02.2022).

Тихонов А.С. 2017. Компьютерная лингвистика и межпредметные связи в преподавании математических и лингвистических дисциплин. В кн.: Математика, информатика, компьютерные науки, моделирование, образование. Сборник научных трудов научно-практической конференции МИКМО-2017 и Таврической научной конференции студентов и молодых специалистов по математике и информатике, Симферополь, 10–14 апреля 2017 г. Под ред. В.А. Лукьяненко. Симферополь, ИП Корниенко: 222–231.

Тьюринг А.М. 1960. Может ли машина мыслить? (С приложением статьи Дж. фон Неймана Общая и логическая теория автоматов.) Пер. с англ. Ю.В. Данилова. Под ред. С.А. Яновской. М., Государственное издательство физико-математической литературы, 67 с. URL: http://www.etheroneph.com/files/can_the_machine_think.pdf (дата обращения: 10.02.2022). (Turing A.M. 1950. Computing Machinery and Intelligence ((Neumann J. 1951. The General and Logical Theory of Automata. In: Cerebral Mechanisms In Behavior. The Hixon Symposium. Ed. L.A. Jeffress. New York—London: 2070–2098.) Mind, New Series, Vol. 59, No. 236: 433-460)

Хобсон Л., Ханнес Х., Коул Х. 2020. Обработка естественного языка в действии. Пер. с нем. И. Пальти, Сергей Черникова. СПб., Питер, 576 с. (Hobson L., Hannes M. H., Cole H. 2019. Natural Language Processing in Action Understanding, analyzing, and generating text with Python. Manning Publications, 544 p.)

Alghazali S.M.M., Polshchykov K., Hailan A.M., Svoykina L. 2021. Development of Intelligent Tools for Detecting Resource-intensive Database Queries. *International Journal of Advanced Computer Science and Applications*, 12 (7): 32–36.

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Voss A.H., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. 2020. Language Models are Few-Shot Learners. DOI: https://doi.org/10.48550/arXiv.2005.14165

Chomsky N. 1964. The logical basis of linguistic theory. In: Chomsky N., Lunt H. Proceedings of the Ninth International Congress of Linguistics, Cambridge, Mass., August 27-31, 1962. The Hague, Publ. Mouton and Co: 914–1008.

Devlin J. Chang M.-W., Lee K., Toutanova K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: arXiv:1810.04805 [cs.CL] (accessed: February 10, 2022). DOI: https://doi.org/10.48550/arXiv.1810.04805

Mahdi T.N., Jameel J.Q., Polshchykov K.A., Lazarev S.A., Polshchykov I.K., Kiselev V. 2021. Clusters partition algorithm for a self-organizing map for detecting resource-intensive database inquiries in a geo-ecological monitoring system. *Periodicals of Engineering and Natural Sciences*, 9 (4): 1138–1145. DOI: http://dx.doi.org/10.21533/pen.v10i1.2584

Mikolov T., Corrado G., Chen K., Dean J. 2013. Efficient Estimation of Word Representations in Vector Space. Available at: arXiv:1301.3781v3 [cs.CL] (accessed: February 10, 2022). DOI: https://doi.org/10.48550/arXiv.1301.3781

Polshchykov K.A., Lazarev S.A., Konstantinov I.S., Polshchykova O.N., Svoikina L.F., Igityan E.V., Balakshin M.S. 2020. Assessing the Efficiency of Robot Communication. *Russian Engineering Research*, 40 (11): 936–938. DOI: 10.3103/S1068798X20110155

Polshchykov K., Lazarev S., Polshchykova O., Igityan E. 2019. The Algorithm for Decision-Making Supporting on the Selection of Processing Means for Big Arrays of Natural Language Data. *Lobachevskii Journal of Mathematics*, 40 (11): 1831–1836. DOI: 10.1134/S1995080219110222

Polshchykov K.O., Lazarev S.A., Zdorovtsov A.D. 2017. Neuro-Fuzzy Control of Data Sending in a Mobile Ad Hoc Network. *Journal of Fundamental and Applied Sciences*, 9 (2S): 1494–1501. DOI: 10.4314/jfas.v9i2s.856

Vaswani A., Shazeer N., Parmar N. Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. 2017. Attention Is All You Need. Available at: https://arxiv.org/pdf/1706.03762.pdf (accessed: February 10, 2022). DOI: https://doi.org/10.48550/arXiv.1706.03762

Velikanova A.S., Polshchykov K.A., Likhosherstov R.V., Polshchykova A.K. 2021. The use of virtual reality and fuzzy neural network tools to identify the focus on achieving project results. In: Artificial Intelligence and Digital Technologies in Technical Systems II-2021. *Journal of Physics: Conference Series*, 2060 (1): 012017. DOI: 10.1088/1742-6596/2060/1/012017

References

Aguzumtsyan R.V., Velikanova (Gerasimova) A.S., Pol'shchikov K.A., Igityan E.V., Likhosherstov R.V. 2021. Application of intellectual technologies of natural language processing and virtual reality means to support decision-making when selecting project executors. *Economics. Information Technologies*, 48(2): 392–404. DOI: 10.52575/2687-0932-2021-48-2-392-404

Apresyan Yu.D., Boguslavskiy I.M., Iomdin L.L. 1992. Lingvisticheskiy protsessor dlya slozhnykh informatsionnykh system [A Linguistic processor for complex information systems]. Ed. L.P. Krysin. M., Publ. Nauka, 256 p.

Batura T.V. 2016. Matematicheskaja lingvistika i avtomaticheskaja obrabotka tekstov na estestvennom jazyke [Mathematical Linguistics and Automatic Processing of Natural Language Texts]. Novosibirsk, Publ. RIC Novosibirskij nacional'nyj issledovatel'skij gosudarstvennyj universitet, 166 p.

Boguslavskij I.M., Iomdin L.L., Krejdlin L.G., Frid N.E., Sagalova I.L., Sizov V.G. 2000. Modul' universal'nogo setevogo jazyka v sostave sistemy JeTAP-3[1] [Module of the universal network language as part of the ETAP-3 system[1]]. In: Sbornik 2000. Available at: https://www.dialog-21.ru/digest/2000/articles/boguslavsk_i_m/ (accessed: February 10, 2022).

Blehman M.S. 2012. Kratkaja istoricheskaja spravka o zarozhdenii i uspeshnom razvitii komp'juternoj lingvistiki v SSSR [Brief historical background on the origin and successful development of computational linguistics in the USSR]. Peterburgskaja bibliotechnaja shkola, 2 (39): 4–6.

Bol'shakova E.I., Voroncov K.V., Efremova N.Je., Klyshinskij Je.S., Lukashevich N.V., Sapin A.S. 2017. Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i analiz dannyh [Automatic natural language processing and data analysis]. M., Publ. Izd-vo NIU VShJe, 269 p.

Drozdova K.A. 2015. Mashinnyy perevod: istoriya, klassifikatsiya, metody [Machine translation: history, classification, methods]. In: Filologicheskie nauki v Rossii i za rubezhom [Philological sciences in Russia and abroad]. Proceedings of the III International Scientific Conference, St. Petersburg, July 20–23, 2015. SPb., Publ. Svoe izdatel'stvo: 139–141. Available at: https://moluch.ru/conf/phil/archive/138/8497/ (accessed: April 8, 2022).

Mel'chuk I.A. 1999. Opyt teorii lingvisticheskikh modeley «Smysl ⇔ Tekst» [Experience of the theory of linguistic models "Meaning ⇔ Text"]. M., Publ. Shkola «Yazyki russkoy kul'tury», 346 p.

Mitrenina O.V. 2017. Back to 1947: on the seventieth anniversary of machine translation as a scientific project. Vestnik Novosibirsk State University. Series: Linguistics and Intercultural Communication, 15(3): 5–12 (in Russian). DOI: 10.25205/1818-7935-2017-15-3-5-12

Mitrenina O.V. 2019. Artificial neural networks and natural language processing. *Journal of Applied Linguistics and Lexicography*, 1 (2): 399–408. (in Russian).

Leont'eva N.N. 2006. Avtomaticheskoe ponimanie tekstov: sistemy, modeli, resursy [Automatic text comprehension: systems, models, resources]. M., Publ. Akademija, 304 p.

Prikladnaya i komp'yuternaya lingvistika [Applied and Computational Linguistics]. 2016. Eds. I.S. Nikolaev, O.V. Mitrenina, T.M. Lando. M., Publ. Lenand, 320 p.

Rarenko M.B. 2021. Mashinnyy perevod: ot perevoda «po pravilam» k neyronnomu perevodu [Machine translation: from translation "by the rules" to neural translation]. Social Sciences and Humanities. Domestic and Foreign Literature. Series 6: Linguistics, 3: 70–79. DOI: 10.31249/ling/2021.03.05

Sokirko A.V., Toldova S.Yu. 2005. Comparing a stochastic tagger based on Hidden Markov Model with a rule-based tagger for Russian. In: Internet-matematika 2005: avtomaticheskaya obrabotka veb-dannykh [Internet Mathematics 2005: Automatic Web Data Processing]. M.: 80-94. Available at: http://hdl.handle.net/10995/1391 (accessed: February 10, 2022).

Tikhonov A.S. 2017. Computational linguistics and interdisciplinary relationships in mathematics and linguistics education. In: Mathematics Informatics Computer Science Modeling Education. Collection of scientific papers of the scientific-practical conference MIKMO-2017 and the Taurian scientific conference of students and young specialists in mathematics and computer science, Simferopol, April 10–14, 2017. Ed. V.A. Luk'yanenko. Simferopol', Publ. IP Kornienko: 222–231 (in Russian).

T'yuring A.M. 1960. Mozhet li mashina myslit' [Can a machine think]? (S prilozheniem stat'i Dzh. fon Neimana. Obshchaya i logicheskaya teoriya avtomatov [General and logical theory of automata].) Per. from English. Yu.V. Danilov. Ed. S.A. Yanovskaya. M., Publ. Gosudarstvennoe izdatel'stvo fiziko-matematicheskoi literatury, 67 p. Available at: http://www.etheroneph.com/files/can_the_machine_think.pdf (accessed: February 10, 2022). (Turing A.M. 1950. Computing Machinery and Intelligence ((Neumann J. 1951. The General and Logical Theory of Automata. In: Cerebral Mechanisms In Behavior. The Hixon Symposium. Ed. L.A. Jeffress. New York—London: 2070–2098.) Mind, New Series, Vol. 59, No. 236: 433-460)

Khobson L., Khannes Kh., Koul Kh. 2020. Obrabotka estestvennogo yazyka v deystvii [Natural language processing in action]. Per. s nem. I. Pal'ti, Sergey Chernikov. SPb., Publ. Piter, 576 p. (Hobson L., Hannes M. H., Cole H. 2019. Natural Language Processing in Action Under-standing, analyzing, and generating text with Python. Manning Publications, 544 p.)

Alghazali S.M.M., Polshchykov K., Hailan A.M., Svoykina L. 2021. Development of Intelligent Tools for Detecting Resource-intensive Database Queries. International Journal of Advanced Computer Science and Applications, 12 (7): 32–36.

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Voss A.H., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. 2020. Language Models are Few-Shot Learners. DOI: https://doi.org/10.48550/arXiv.2005.14165

Chomsky N. 1964. The logical basis of linguistic theory. In: Chomsky N., Lunt H. Proceedings of the Ninth International Congress of Linguistics, Cambridge, Mass., August 27-31, 1962. The Hague, Publ. Mouton and Co: 914–1008.

Devlin J. Chang M.-W., Lee K., Toutanova K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: arXiv:1810.04805 [cs.CL] (accessed: February 10, 2022). DOI: https://doi.org/10.48550/arXiv.1810.04805

Mahdi T.N., Jameel J.Q., Polshchykov K.A., Lazarev S.A., Polshchykov I.K., Kiselev V. 2021. Clusters partition algorithm for a self-organizing map for detecting resource-intensive database inquiries in a geo-ecological monitoring system. *Periodicals of Engineering and Natural Sciences*, 9 (4): 1138–1145. DOI: http://dx.doi.org/10.21533/pen.v10i1.2584

Mikolov T., Corrado G., Chen K., Dean J. 2013. Efficient Estimation of Word Representations in Vector Space. Available at: arXiv:1301.3781v3 [cs.CL] (accessed: February 10, 2022). DOI: https://doi.org/10.48550/arXiv.1301.3781

Polshchykov K.A., Lazarev S.A., Konstantinov I.S., Polshchykova O.N., Svoikina L.F., Igityan E.V., Balakshin M.S. 2020. Assessing the Efficiency of Robot Communication. *Russian Engineering Research*, 40 (11): 936–938. DOI: 10.3103/S1068798X20110155

Polshchykov K., Lazarev S., Polshchykova O., Igityan E. 2019. The Algorithm for Decision-Making Supporting on the Selection of Processing Means for Big Arrays of Natural Language Data. *Lobachevskii Journal of Mathematics*, 40(11): 1831–1836. DOI: 10.1134/S1995080219110222

Polshchykov K.O., Lazarev S.A., Zdorovtsov A.D. 2017. Neuro-Fuzzy Control of Data Sending in a Mobile Ad Hoc Network. *Journal of Fundamental and Applied Sciences*, 9(2S): 1494–1501. DOI: 10.4314/jfas.v9i2s.856

Vaswani A., Shazeer N., Parmar N. Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. 2017. Attention Is All You Need. Available at: https://arxiv.org/pdf/1706.03762.pdf (accessed: February 10, 2022). DOI: https://doi.org/10.48550/arXiv.1706.03762

Velikanova A.S., Polshchykov K.A., Likhosherstov R.V., Polshchykova A.K. 2021. The use of virtual reality and fuzzy neural network tools to identify the focus on achieving project results. In: Artificial Intelligence and Digital Technologies in Technical Systems II-2021. *Journal of Physics: Conference Series,* 2060(1): 012017. DOI: 10.1088/1742-6596/2060/1/012017

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.
Conflict of interest: no potential conflict of interest related to this article was reported.

Поступила в редакцию 13.04.2022.
Поступила после рецензирования 14.07.2022
Принята к публикации 10.09.2022

Received April 13, 2022.
Revised July 14, 2022
Accepted September 10, 2022

ИНФОРМАЦИЯ ОБ АВТОРЕ

Польщикова Ольга Николаевна, кандидат филологических наук, доцент кафедры русского языка, профессионально-речевой и межкультурной коммуникации, Белгородский государственный национальный исследовательский университета, г. Белгород, Россия

INFORMATION ABOUT THE AUTHOR

Olga N. Polshchykova, PhD in Philology, Associate Professor of the Department of Russian Language, Professional Speech and Intercultural Communication, Belgorod National Research University, Belgorod, Russia