УДК 004.5: 004.9: 001.891
https://doi.org/10.20913/2618-7575-2021-4-81-92

АНАЛИЗ ПУБЛИКАЦИЙ ПО ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА, ИНДЕКСИРОВАННЫХ В РИНЦ

ANALYSIS OF PUBLICATIONS ON NATURAL LANGUAGE PROCESSING INDEXED IN RSCI

© ***Садовская Лариса Леонидовна***
младший научный сотрудник, зав. отделом справочно-информационного обслуживания, Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук (ГПНТБ СО РАН), Новосибирск, Россия, sadovskaya@gpntbsib.ru

Обработка естественного языка (ОЕЯ), определяемая как общее направление искусственного интеллекта и математической лингвистики, является важным инструментом для понимания и обработки гигантского объема неструктурированных данных. Представляя собой теоретическую и практическую основу для решения большого количества задач, ОЕЯ активно исследуется учеными всего мира, поскольку является одной из самых популярных областей науки о данных и применяется во многих сферах деятельности человека.
Цель проведения библиометрического анализа: определить основные научные центры, тенденции исследований, рейтинг ведущих российских ученых в области ОЕЯ, а также показать картину распределения публикаций по тематическим полям. Результаты анализа позволят определить динамику развития области ОЕЯ в отечественной науке и предоставят ученым и специалистам, работающим в обозначенной научной области, актуальную информацию о различных аспектах рассматриваемого направления исследований.
В качестве источника данных для поиска научных публикаций в области ОЕЯ использовалась база данных (БД) «Российский индекс научного цитирования» (РИНЦ) – основная наукометрическая база в нашей стране.
Работа будет интересна исследователям в области ОЕЯ, поскольку содержит актуальные данные о динамике развития и структуре документально-информационного потока в области ОЕЯ в России, его распределения по отраслям знаний. Кроме того, проведенный библиометрический анализ документов позволяет получить информацию об авторах, наиболее продуктивно работающих в исследуемой области и их аффилиациях, а также о наиболее цитируемых статьях – это обзор 30 самых цитируемых публикаций российских ученых и топ-10 научных организаций Российской Федерации по количеству публикаций с отражением основных проблем и научных достижений в области ОЕЯ.

Ключевые слова: *обработка естественного языка, ОЕЯ, библиометрический анализ, РИНЦ*

Sadovskaya Larisa Leonidovna
Junior Researcher, Head of the Department of Reference and Information Service, State Public Scientific Technological Library of the Siberian Branch of the Russian Academy of Sciences (SPSTL SB RAS), Novosibirsk, Russia, sadovskaya@gpntbsib.ru

Natural Language Processing (NLP), defined, as the general direction of artificial intelligence and mathematical linguistics, is an important tool for understanding and processing a gigantic amount of unstructured data.
Representing theoretical and practical basis for solving a large number of problems, NLP is being actively studied by scientists around the world, since it is one of the most popular areas of data science and is used in many fields of human activity.
The purpose of this bibliometric analysis is to identify the main research centers, research trends, the rating of leading Russian scientists in the field of NLP, as well as to show the picture of thematic fields' distribution of publications. The results of the analysis will allow us to determine the dynamics of NLP field development in the domestic science and will provide scientists and specialists working in the field of NLP with up-to-date information on various aspects of the considered research area.
Russian Science Citation Index (RSCI) – the main scientometric database in our country was used as the data source for scientific publications in the field of NLP search. The RSCI platform contains information on current research areas, allows you to assess the productivity of scientists, scientific communities, etc.
The presented work will be interesting to researchers in the field of NLP, since it contains actual data on the development dynamics and structure of the documentary and information flow in the field of NLP in Russia, its distribution by branches of knowledge. Besides, the conducted bibliometric analysis of documents allows us to gain information about the authors most productively working in the studied area and their affiliations, as well as about the most cited articles – this is the review of the 30 most cited publications of Russian scientists and rating of TOP-10 scientific organizations of the Russian Federation by the number of publications reflecting the main problems and scientific achievements in the field of NLP.

Keywords: *natural language processing, NLP, bibliometric analysis, RSCI*

Введение

XXI в. характеризуется увеличением объема данных в геометрической прогрессии, представленных в большей степени языковыми данными. Сфера автоматизированной обработки данных естественных языков (ЕЯ) переживает бурный рост, благодаря улучшению доступа к языковым данным и колоссальному увеличению вычислительных мощностей.

Чаще всего обработку естественного языка (ОЕЯ) определяют как общее направление искусственного интеллекта (ИИ) и математической лингвистики, изучающее проблемы компьютерного анализа и синтеза ЕЯ [1]. Сегодня технологии ОЕЯ имеют широкий спектр применения во многих научных и прикладных дисциплинах для решения разных типов задач, становясь одной из основных технологий практического применения ИИ.

Развитие научных направлений приводит к соответствующему росту объема публикаций. Об этом свидетельствуют, например, результаты исследования в области ОЕЯ (англ. Natural Language Processing) с 1959 по 2019 г., по данным информационно-аналитической системы Scopus. БД Scopus за 2000–2019 гг. содержит около 57 тыс. публикаций по ОЕЯ, что составляет 92,6 % всех включенных в БД работ по рассматриваемой области и позволяет судить о значительном росте интереса к области исследования ОЕЯ в мире за последние два десятилетия [2]. В статье представлено развитие направления ОЕЯ в нашей стране.

Исследование состоит из двух частей: библиометрического анализа информационного массива публикаций в области ОЕЯ по БД РИНЦ и обзора 30 наиболее цитируемых трудов. Глубина временного охвата документов, используемых в исследовании: с 1980 (в этом году появилась первая статья по ОЕЯ в РИНЦ) по 2019 г. (рецензирование и загрузка публикаций в РИНЦ за 2019 г. завершена, поэтому представление информации за этот год можно считать корректным).

Результаты библиометрического анализа позволят определить тенденции развития интересующего нас направления исследований в российском сегменте науки.

Материалы, методы и результаты проведенного библиометрического исследования

Методы библиометрического анализа позволяют объективно представить состояние, тенденции развития тематик и проблем, выявить наиболее динамично развивающиеся или затухающие направления научных исследований в мире, определить вклад отдельных ученых, коллективов в науку и т. д. [3].

По сформированному поисковому запросу «Обработка естественного языка (OR Natural Language Processing)» в РИНЦ была проведена выборка документов по названию, аннотации, ключевым словам, временной охват – без ограничений. Из полученного массива были исключены документы авторов, не аффилированных в российских организациях. В результате получена выборка из 2302 публикаций, из них: 56 % – журнальные статьи, 35 % – статьи в сборниках трудов конференций, 6 % – диссертации, около 3 % – другие типы публикаций (книга, сборник статей, глава в книге).

Анализ количества публикаций по ОЕЯ в РИНЦ по годам представляет картину стабильной положительной динамики роста за последнее десятилетие (рис. 1).

По тематическим направлениям все публикации в области ОЕЯ в РИНЦ распределились следующим образом: «Кибернетика» и «Языкознание» – по 23 %; «Автоматика. Вычислительная техника» – 18 %; «Информатика» – 14 %; «Науковедение; общие и комплексные проблемы естественных, точных, технических и прикладных наук и др.» – 14 %; «Математика» – 8 %.

По научному направлению ОЕЯ в РИНЦ отражены публикации 4004 авторов из более 200 организаций, имеющих аффилиацию в Российской Федерации (рис. 2). Наибольшее количество публикаций у сотрудников Национального исследовательского университета «Высшая школа экономики» (НИУ «Высшая школа экономики») – 88, за ними следуют авторы Санкт-Петербургского государственного университета (СПбГУ) – 77 работ и Федерального исследовательского центра «Информатика и управление» РАН (ФИЦ «Информатика и управление РАН») – 68.

Наиболее публикуемыми авторами в рассматриваемой научной области являются: Е. Б. Козеренко (21 работа), Д. М. Коробкин (20), Т. В. Батура (18), А. В. Глазкова (18), А. Г. Сбоев (17), А. В. Пруцков (16), С. А. Фоменков (16), Д. А. Усталов (15), Р. Б. Рыбка (14), М. М. Шарнин (13).

Обзор топ-30 цитируемых трудов в области ОЕЯ в БД РИНЦ

Анализ полученной выборки показывает, что 894 публикации российских авторов имеют количество цитирований в РИНЦ от 81 до 1. 30 самых цитируемых работ российских авторов, попавших в выборку, были опубликованы в 2002–2017 гг. По содержанию работы разделяются на следующие группы:
– лингвистика, отражающая разные языковые аспекты ОЕЯ-исследований;

– технологии, алгоритмы, инструменты, используемые в системах ОЕЯ;
– методы, подходы, применяемые в исследованиях ОЕЯ;
– проблемы и достижения в области ОЕЯ.

Работы по лингвистике, отражающие языковые аспекты исследований ОЕЯ

Как уже отмечалось выше, имеется большое количество актуальных прикладных задач, для решения которых наиболее эффективны методы компьютерной лингвистики (КЛ) и автоматической обработки текстов на ЕЯ. КЛ показывает хорошие результаты в различных приложениях по автоматической ОЕЯ, часто качество результатов достигает экспертного уровня. Прогресс в области КЛ связан с более точным учетом лингвистических особенностей текстов на различных этапах обработки и применением лингвистических моделей. Поэтому не удивительно, что большой популярностью у исследователей и разработчиков программ

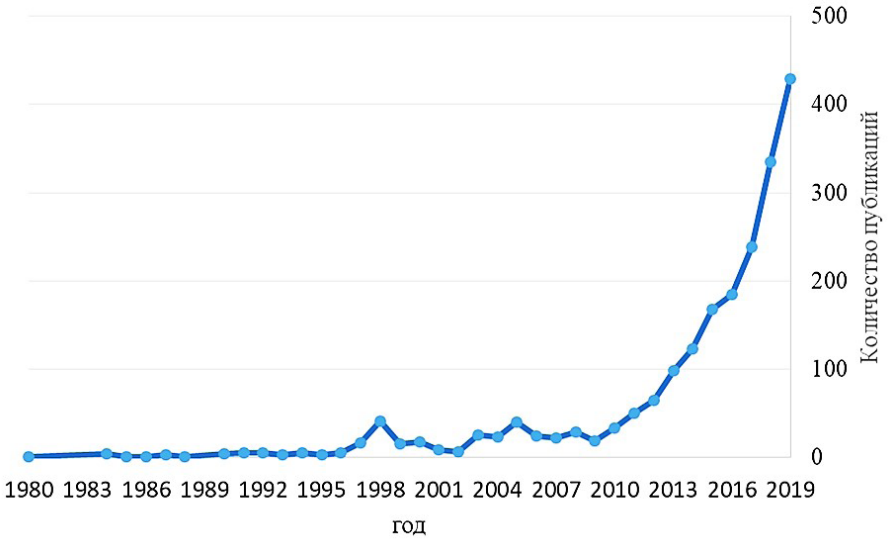

Рис. 1. Распределение публикаций 1980–2019 гг. по ОЕЯ в РИНЦ

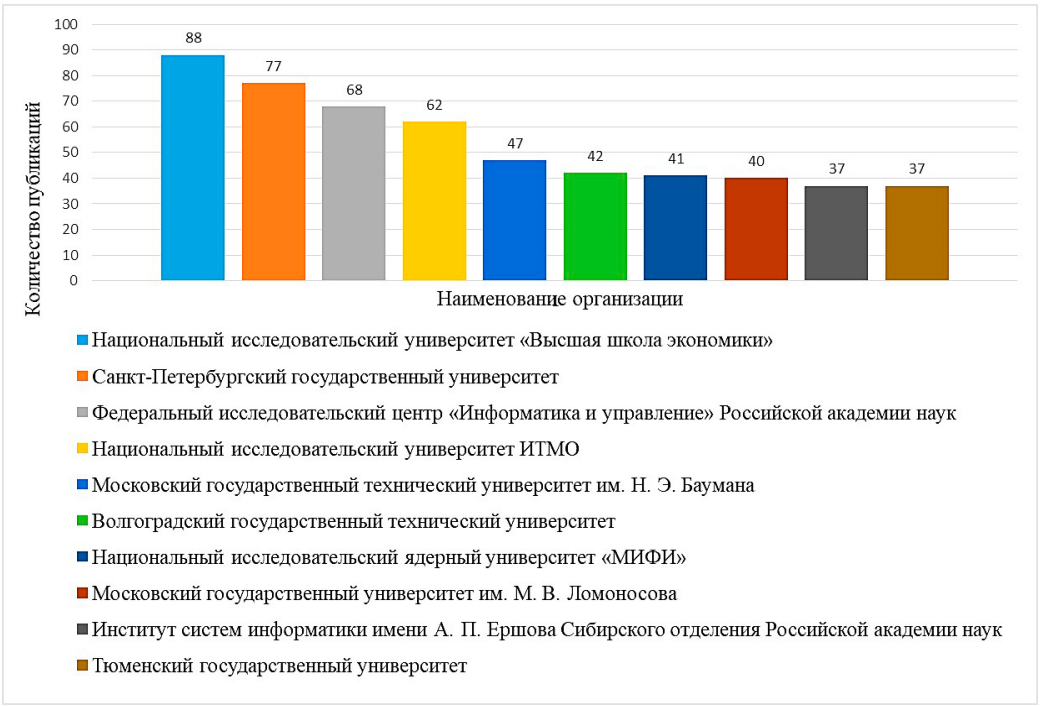

Рис. 2. Топ-10 организаций РФ по количеству публикаций в области ОЕЯ в РИНЦ

в области ОЕЯ пользуется научная литература по методам, технологиям, ресурсам КЛ. К их числу можно отнести коллективную монографию «Прикладная и компьютерная лингвистика» [4], (52) *[1], которая поможет углубить знания по современным лингвистическим технологиям, перспективам применения знаний о ЕЯ для решения прикладных задач.

Определение термина «компьютерная лингвистика», основных понятий предметной области, классификация лингвистического программного обеспечения, соотношение терминологии КЛ дается в работе «Предметная область компьютерной лингвистики» [5], (12) *. КЛ определена как междисциплинарная наука, развитие которой детерминируется математическими, техническими, лингвистическими основами, математике и языкознанию отведена методологическая роль.

Отметим полезность «Портала знаний по основным разделам КЛ», появившегося в 2008 г. Этот ресурс содержит информацию об используемых моделях и методах, о выполняемых в этой области проектах. Портал обеспечивает содержательный доступ к информационным ресурсам, представляющим реальные прикладные системы, технологии и программные продукты для ОЕЯ, лингвистические ресурсы и базы данных [6], (39) *.

Для специалистов в области ОЕЯ также представляет интерес проведенный анализ применения таких методов прикладной лингвистики, как автоматизированное извлечение информации, обучение на основе данных, текстовый поиск в крупномасштабных корпусах, которые применяются в преподавании иностранного языка. Любопытным представляется рассмотрение потенциала обучения на базе корпусного языкового материала.

Сегодня в компьютеризированном языковом обучении для создания «умных» методик интенсивно ведутся разработки технологий с применением методов ОЕЯ. В этих методиках используются моделирование реакций обучающегося и лингвистический анализ для адекватной оценки языкового материала, свободно сформулированных ответов вплоть до эссе. Таким образом, обучение языку при помощи обозначенных технологий отходит от деконтекстуализованных способов подачи материала и, кроме использования в процессе преподавания иностранного языка, может применяться также и для оценки методических материалов [7], (50) *.

[1] * Здесь и далее: количество цитирований публикации по данным РИНЦ.

Методы, подходы, используемые в системах ОЕЯ

В целях разработки универсального метода (УМ) генерации и определения форм слов выявлены преимущества и недостатки существующих отечественных и зарубежных методов морфологической обработки текстов на ЕЯ. Морфологический анализ и синтез, ориентированный на один ЕЯ, не подходит для построения УМ. Словари таких методов специализированы под особенности только одного языка и не могут быть изменены для других языков. В УМ имеется возможность корректировать подходы к морфологическому анализу.

На основе анализа рассмотренных подходов сформулированы требования к УМ, в частности: обработка словоформ языков различных групп и семейств; универсальность структуры словарей, не требующей конвертации для решения задач определения или генерации словоформ; модель формообразования, на основе которой построен метод, должна описывать любые виды образования форм всей парадигмы слова [8], (25) *. Предложенная модель формообразования ЕЯ позволяет представлять словоформы с данным грамматическим значением в виде последовательности конечного числа преобразований над основой. Предложенный метод удовлетворяет всем критериям универсальности [9], (25) *.

Специалистам известно, что автоматическое извлечение текста формы события является важным шагом в приобретении знаний и заполнении базы знаний. Ручная работа при разработке системы извлечения необходима либо для аннотации корпуса, либо для создания словарей и шаблонов для системы, основанной на знаниях. В последнее время исследования были сосредоточены на адаптации существующей системы (для извлечения из английских текстов) к новым областям. Извлечение событий на других языках не изучалось из-за отсутствия ресурсов и алгоритмов, необходимых для ОЕЯ. В 2016 г. [10], (21) * удалось определить такой набор лингвистических ресурсов, необходимых для разработки системы извлечения событий на основе знаний на русском языке – словарей, которые являются базовыми блоками для семантических шаблонов, – и представить набор методов для создания таких словарей как на русском, так и на других языках.

Создание новых методов семантического анализа текстов на ЕЯ актуально для решения многих задач ОЕЯ. При этом семантический анализ считается самым сложным этапом ОЕЯ, поскольку процесс человеческого мышления, как и ЕЯ, трудно поддается формализации. Сегодня существует много методов представления смысла высказываний, но пока ни один из них не может являться универсальным. В работе Т. В. Батура [11], (20) * предпринята

попытка систематизировать известные достижения в области машинно-ориентированного семантического анализа.

Значительная роль в ОЕЯ отводится методам конструирования моделей предметной области, основанной на извлечении объектов и процессов ОЕЯ, а также подходам к созданию базы знаний на основе механизма расширенных семантических сетей с описанием принципов построения и расширения терминологических тезаурусов.

Ассоциативные связи между терминами, понятиями и другими элементами ЕЯ играют важную роль в решении широкого класса прикладных задач ОЕЯ, среди которых интеллектуальная обработка текстов на ЕЯ с формированием баз знаний и организация различных видов поиска, в том числе семантического. Методы автоматизированного выявления ассоциативных связей в текстах и построения ассоциативных портретов предметных областей (АППО) ориентированы на решение перечисленных задач [12], (20)*.

Кроме того, разработана модель, в которой при обработке текстов ЕЯ, описывающих некоторую предметную область, строятся наборы фрагментов семантической сети, образующие базу знаний заданной предметной области. Последующая обработка поступающих документов позволяет не только распознать уже выделенные объекты, но и определить их свойства, связи с другими объектами предметной области. Большой объем выделенных и выверенных знаний позволяет свести к минимуму количество ошибок при построении структуры данных для определенной предметной области. Кроме этого, анализ построенных знаний дает возможность выделить из них закономерности их построения и создавать метазнания [13], (20)*.

Обобщает и систематизирует разработанные ранее подходы к обработке ЕЯ-текстов работа «Ассоциативно-онтологический подход к обработке текстов на естественном языке» [14], (17)*. Наиболее распространен лингвистический подход, использующий анализатор текстов, основанный на синтаксисе. Предложенный метод расширяет методы лингвистической статистики и логико-статистические методы для извлечения знаний и построения ассоциативной онтологии заданной предметной области. На основе описанного подхода могут быть построены поисковые и справочные системы с использованием ассоциативно-онтологического поиска информации. Подход дает возможность реализовать подсистему тематической локализации области поиска (по областям знания и сферам деятельности), а система может решать задачу поиска документов и фактов даже при недостатке начальных данных.

Поиск релевантных документов, их содержательный анализ, индексирование и классификация на основе предметного словаря и онтологии, сведение ресурсов, относящихся к одной области знаний, в единое информационное пространство – сложный процесс. Для его организации необходимо решить задачу автоматизированного извлечения онтологической информации о предметной области из текстов ЕЯ. Для этого используется объем уже выполненных работ по описанию ключевых понятий: онтологическая информация извлекается из энциклопедических словарей по различным областям знания. Далее текст словаря переводится в структурированный вид и из него извлекается онтологическая информация. Для решения задачи автоматизированного построения онтологий предложен язык лингвистических шаблонов, позволяющий гибко и компактно задавать правила выделения лингвистических структур из текста на ЕЯ. Кроме того, язык лингвистических шаблонов можно применять и для решения других задач: для классификации предложений или выделения иных специфических лингвистических структур. Наконец, сам язык лингвистических шаблонов можно расширять, увеличивать его выразительную силу в целях использования для углубленного анализа результатов ОЕЯ [15], (20)*.

Работы по применению стилометрических методов при решении задач ОЕЯ, таких как атрибуция и проверка авторства, профилирование авторства и классификация текста по жанрам и настроениям, систематизированы в обзоре авторов К. Лагутиной и др. [16], (13)*. Стилометрия – это раздел КЛ, изучающий количественную оценку языковых особенностей в текстах на ЕЯ, выбор стилометрических характеристик текста является наиболее важным этапом исследования из-за сложности и многомерности текста. Исследователи в области стилометрии в первую очередь оперируют признаками, отражающими количественные показатели довольно низкоуровневых текстовых функций. Однако стиль автора часто выражается в аспектах, которые довольно трудно найти. Проблемы, препятствующие использованию специфики авторского стиля в задачах атрибуции, верификации и других, являются недостатком информации о соотношении стилометрических признаков друг с другом и признаков с доменами. Эти задачи могут быть решены путем изучения стилистических особенностей в идиостиле разных авторов, в разных жанрах и создания корпусов текстов с разметкой авторского стиля, которые могут быть использованы для более глубокого исследования.

Анализ тональности текста (Sentiment Analysis) – класс математических методов ОЕЯ для выявления и изучения эмоциональной составляющей текста. Задача анализа тональности является частной задачей классификации текстов и извлечения

информации, которая лежит в области КЛ. Сложность задач КЛ связана с тем, что ЕЯ – сложная многоуровневая система знаков, возникшая для обмена информацией между людьми и постоянно изменяющаяся. Другая сложность разработки методов КЛ связана с многообразием ЕЯ, существенными отличиями их лексики, морфологии, синтаксиса, вариативностью выражения смысла.

Наибольший научный интерес с точки зрения улучшения точности анализа представляет статистический подход, а с точки зрения улучшения качества – аспектный. Лингвистический подход по своей сути не обладает никакими интеллектуальными особенностями, поскольку формализует уже накопленные лингвистические знания, однако используемые в данном подходе правила могут успешно применяться и в других подходах для повышения точности классификации [17], (11)*.

Что касается исследований методов сентимент-анализа русскоязычных текстов, то отмечено: наилучшими являются способы, в которых учитываются связи между словами, структура языка [18], (11)*. В одном из методов сентимент-анализа предложено использовать два типа свойств: узловые и граничные. Узловые обозначают узлы дерева, а граничные – связи между словами. Вместо использования свойств границ, соединяющих зависимый узел и его родителя, предлагается использовать свойства гипергранцы, которые соединяют всех потомков корневого узла дерева, то есть слово может быть потомком нескольких узлов дерева. После этого для подсчета конечного результата для корневого узла выделяются все возможные поддеревья. Так как их может быть довольно много, то часть из них исключается статистическими методами. Программно реализован метод, основанный на словарях, и метод с использованием библиотеки сентимент-анализа Стэнфорда и сервиса переводов «Яндекса».

Технологии, алгоритмы, инструменты, применяемые в системах ОЕЯ

Начнем с программно-информационного комплекса, который успешно применяется при построении прагматически-ориентированных систем и информационных технологий ОЕЯ, представленного авторами Д. Ш. Сулеймановым и А. Р. Гатиатуллиным. В их монографии [19], (17)* дана структурно-функциональная модель аффиксальных морфем и программно-информационный инструментарий, созданный на ее основе. Модель морфем является открытой, что позволяет вносить в нее новые характеристики или аспекты, при необходимости модифицируя и структуру модели. Особенно важной и в корне отличающей структурно-функциональную модель от электронных словарей является возможность ввода логических запросов и решения лингвистических задач в режиме вычисления на основе определенных данных и условий. На базе модели лексики разработан и реализован морфологический анализатор, используемый в настоящее время в электронных словарях и прикладных программах по обработке ЕЯ-текстов. Наряду с теоретическими выкладками и построениями в области лингвистического моделирования в работе представлено описание программно-информационного комплекса, готового для практического применения при проведении статистических исследований и при работе со словарями.

В результате внедрения компьютерных систем во все сферы человеческой жизни все больше проявляется проблема перехода от визуальных и командных интерфейсов к естественно-языковым. В статье рассмотрены методы КЛ и ОЕЯ. Представлено полное описание всех стадий ОЕЯ, таких как морфологический, синтаксический и семантический анализ. Рассмотрен ограниченный язык как подмножество ЕЯ, на котором текст хорошо воспринимается носителем ЕЯ без дополнительных усилий. Подобное решение помогает избежать разночтений на лингвистическом уровне. Разработанный прототип естественно-языкового пользовательского интерфейса производит преобразования пользовательского запроса на естественном языке в SQL-запрос к БД. Интерфейс взаимодействует с БД, содержащей информацию о существующих программных библиотеках и фреймворках. Таким образом, использование методов ОЕЯ позволяет разработать ЕЯ пользовательский интерфейс для взаимодействия с диалоговой системой [20], (9) *. Что касается описания особенностей алгоритмов и программ ОЕЯ, применяемых на морфологическом, лексическом, синтаксическом и дискурсивном уровнях языковой системы, то они достаточно полно представлены в работе В. А. Яцко [21], (18)*.

Миварные технологии накопления и обработки информации позволяют обрабатывать тексты на сверхбольших объемах двудольных графов в масштабе реального времени, то есть мивары позволили выйти на качественно новый уровень исследований в компьютерных науках и в области искусственного интеллекта, поэтому их роль сложно преувеличить. Вероятно, по этой причине наибольшее количество обращений имеют статьи 2016 г. [22], (81)* и 2013 г. [23], (66)*, посвященные последним достижениям в использовании миварных технологий, которые являются наследием советской математической школы, сегодня результаты миварных технологий России превосходят мировой уровень. Российскими исследователями предложено использовать миварные технологии в области разработки автономных интеллектуальных роботов нового поколения, в частности, по таким направлениям,

как понимание текстов и ЕЯ. В работах [22; 23] представлены успехи в использовании миварных технологий в математическом моделировании понимания ЕЯ, изображений и человеческой речи. Авторы внедрили системный подход, разработали более сложные инструменты моделирования, накопления и логической обработки данных. Достижения миварных технологий в накоплении и обработке информации обеспечат переход на новый качественный уровень автоматизированной обработки текстов, в том числе на ЕЯ на основе логической обработки больших массивов данных и учета контекста.

В автоматических системах обработки текстов многие исследователи подчеркивают важность построения электронных тезаурусов и перспективу их использования. Решая задачи в области автоматической ОЕЯ, специалисты обнаружили тот факт, что тезаурус является удобной моделью предметной области. Тезаурус может использоваться и как информационно-поисковый ресурс, и как источник или эталон терминологии. Связи между словами являются материалом для построения лексико-семантических сетей для извлечения знаний, для определения семантической близости слов.

В 2014 г. появилась общедоступная версия тезауруса русского языка RuThes – RuThes-Lite, которая имела сходство с WordNet (семантическая сеть для английского языка, разработанная в Принстонском университете), но ее структура и отношения были основаны на психолингвистических экспериментах и не предназначались для выполнения задач ОЕЯ. Существенно новым в предложенной модели стал набор отношений лингвистической онтологии, специально подобранный для описания предметной области. Двумя годами позже был проведен процесс преобразования тезауруса русского языка версии RuThes-Lite 2.0 в RuWordNet. Достижением стали: разделение данных на части речи – ориентированные структуры с перекрестными ссылками между ними и обеспечение набора отношений, аналогичных WordNet-подобным источникам. Сравнение веб-страничных представлений RuThes 2 и Ru-WordNet показало, что RuThes выглядит как онтология, описывающая понятия и их отношения, а RuWordNet – как сеть слов [24], (34)*; [25], (31)*.

Наиболее распространенным способом представления информации являются текстовые документы, часто относящиеся к определенной предметной области. В рамках проекта Texterra была создана технология, позволяющая решать широкий класс задач, связанных с обработкой текстовых данных на ЕЯ. В зависимости от решаемой задачи Texterra может быть использована как библиотека алгоритмов, расширяемый фреймворк или масштабируемый облачный сервис. В отличие от большинства существующих систем ОЕЯ, Texterra предоставляет возможность перехода от работы с отдельными словами и терминами к работе с их значениями. Это позволяет увеличить точность решения многих прикладных задач. Важным преимуществом технологии Texterra являются низкие затраты на внедрение и поддержание системы за счет автоматизации процесса построения и обновления базы знаний. Предложенный подход позволяет не только применять разработанные методы к заранее определенной предметной области, но и быстро адаптировать технологию к новым задачам и языкам. В работе представлены данные о деталях реализации проекта, вариантах использования и результатах экспериментальных исследований разработанных инструментов [26], (16)*.

В рамках единого инструментального комплекса, ориентированного на организацию баз знаний и на их использование для решения интеллектуальных задач в области ОЕЯ, предложены и реализованы семантические методики по извлечению структур знаний из текстов на ЕЯ. Предметные и лингвистические знания представляются на единой основе, что позволяет свести неоднородные задачи к преобразованию структур знаний. Это упрощает создание программ, обеспечивающих анализ высокой степени глубины и сложности [27], (20)*.

Компьютерные системы сегодня внедрены в самые различные сферы деятельности человека, ускорить и упростить работу с которыми позволяют системы общения на ЕЯ. Задачей морфологического этапа является классификация слов по частям речи в соответствии с их грамматическими характеристиками и получение форм слов с заданными характеристиками. Существует несколько подходов к решению задач генерации и определения форм слов в системах общения на ЕЯ, однако они обладают некоторыми недостатками. Разработанная система позволяет устранить недостатки. В основе этой системы лежит принцип: определение и генерацию форм слов можно представить как конечную последовательность преобразований – операций замены и добавления подстрок. Особенность этой системы состоит в том, что алгоритм не зависит от конкретного типа преобразований. Это происходит за счет выноса процедуры преобразования за рамки алгоритма. Такой подход позволяет реализовать практически любые преобразования форм слов, специфичные для отдельных языков [28], (17)*.

Приведенные примеры получения форм слов как цепочки преобразований дают возможность осуществления практически любых преобразований – по этому показателю можно считать разработанную систему и предлагаемые алгоритмы универсальными. Становится также возможным использование разработанной системы в автоматизированных обучающих системах для генерации заданий при обучении морфологии ЕЯ.

Для решения некоторых прикладных задач ОЕЯ необходим синтаксический и семантический парсер – инструмент извлечения структурированных данных из текста на ЕЯ. Парсер ABBYY (система, использующая различные методы синтаксического разбора) занимается разработкой технологии анализа ЕЯ Compreno (технология синтаксического и семантического анализа текста, опирающаяся на универсальную для всех языков иерархию понятий), конечной целью работы этой системы является достижение синтаксической и семантической неоднозначности по решению множества задач, связанных с ОЕЯ [29], (49)*.

В качестве синтаксического анализатора может использоваться также парсер MaltParser (инструмент для работы с деревьями зависимостей), обученный на национальном наборе русского языка. Результатом работы всей системы является база знаний, в которую попадают все извлеченные знания и их отношения, а затем используется для обработки поступающих запросов. В зависимости от типа документа используется соответствующий метод анализа, который имеет наилучшие показатели точности и полноты при анализе текстовых документов данного типа [30], (16)*.

Еще один из разработанных вариантов парсера успешно справляется с разбором простых предложений, но вместе с тем переход его в рабочее состояние потребовал решения некоторых проблем. Первый класс проблем связан с особенностями синтаксической организации русского предложения: возникают сложности, вызванные относительно свободным порядком слов, омонимией на уровне слов и словоформ, синтаксической неоднозначностью.

Второй класс проблем определяется спецификой функционирования и видом выходных данных морфологического анализатора Pymorphy2. Третья категория проблем вызвана особенностями устройства категориальной грамматики в NLTK (Natural Language Toolkit – инструментарий для ОЕЯ).

Грамматика составляющих хорошо зарекомендовала себя в автоматической ОЕЯ с более строгим порядком слов, таких как английский или немецкий. Еще одна сложность связана с тем, что контекстно-свободные грамматики задаются на основе формальных морфологических параметров и не учитывают лексическое значение слова. Наконец, категориальная грамматика в NLTK считает свой разбор верным только тогда, когда может построить дерево составляющих полностью. По итогам проведенного исследования верно утверждение, что идея создания синтаксического анализатора на основе категориальной грамматики является состоятельной, но работа над расширением списка правил и устранением ошибок анализа должна продолжаться [31], (10)*.

Проблемы и достижения области ОЕЯ: отражение в публикациях топ-30

Темой большого количества теоретических и практических исследований сегодня является область построения автоматизированных человеко-машинных систем, которые реализуют комплекс функций по извлечению, обработке знаний, содержащихся в текстах на ЕЯ. Одна из публикаций явилась результатом критического анализа практических достижений в области ОЕЯ по состоянию на 2009 г. В работе сделана попытка определения актуальных направлений развития и собственных экспериментальных исследований ученого в выбранном направлении, сформулированы актуальные направления прикладных исследований [32], (22)*.

Краткий обзор методов ОЕЯ и их использование в моделях поиска представлены в публикации «Обработка текстов естественного языка в моделях поисковых систем» [33], (12)*. В ней рассмотрены основные модели поиска информации, существующие системы семантического поиска. Большинство информационно-поисковых систем (ИПС) являются системами с предварительной обработкой (индексированием) всех имеющихся в системе документов, исключения составляют метапоисковые системы. Отмечены основные проблемы, возникающие при обработке текстов на ЕЯ. Вероятностная модель ИПС характеризуется низкой вычислительной масштабируемостью, необходимостью постоянного обучения системы. Наиболее распространенными являются алгебраические теоретико-множественные модели, так как их практическая эффективность обычно выше. Предлагаемые в последнее время новые реализации проектов ИП являются гибридными моделями и обладают свойствами моделей разных классов.

Выделено одно из перспективных направлений развития ИПС – построение моделей семантического поиска, потенциал у таких систем действительно большой, однако в настоящее время реализованы далеко не все возможные семантические технологии. По сути, сейчас они только помогают выделить ключевые слова из фраз, построенных на ЕЯ, и подобрать дополнительные словоформы для составления корректного поискового запроса. Данное направление методов поиска требует развития.

Исследователи В. Ю. Максимов, Э. С. Клышинский и Н. В. Антонов отмечают, что стремительное развитие искусственного интеллекта в последние годы заставило по-новому взглянуть на проблемы ОЕЯ [34], (10)*. Обзор широкого круга проблем, связанных с пониманием в контексте использования систем ИИ, показывает, что наиболее сложной является проблема машинного понимания ЕЯ. В то же время по многим направлениям развития ОЕЯ отмечен прогресс и ИИ стали открываться такие области, которые ранее однозначно исклю-

чали использование машин. Для дальнейшего расширения области использования ИИ, развития области ОЕЯ ключевым моментом является разрешение именно проблемы машинного понимания ЕЯ.

Заключение

Подход к доступу и обработке информации, основанный на ОЕЯ, будет важным направлением исследований еще в течение долгого времени [35].

На основе данных библиометрического анализа можно сделать вывод, что ОЕЯ является динамично развивающимся направлением и в нашей стране: количество публикаций за последние 10 лет (2010–2019) выросло более чем в 15 раз.

Наибольшее количество самых цитируемых работ относится к тематическим рубрикам:
«Автоматика. Вычислительная техника» – 40 %,
«Кибернетика» и «Информатика» – по 24 %,
«Языкознание» – 12 %.

В результате анализа 30 наиболее цитируемых публикаций российских авторов в области ОЕЯ выявлено, что наибольшее количество обращений имеют работы, посвященные
достижениям в использовании миварных технологий ([22], (81)*; [23], (66)*);
вопросам лингвистики, отражающим разные языковые аспекты ОЕЯ-исследований ([7], (50)*; [6], (39)*);
инструментам извлечения структурированных данных из текста на ЕЯ ([29], (49)*);
моделям усовершенствованных тезаурусов ([24], (34)*; [25], (31)*).

Целый ряд российских авторов имеет значимые исследовательские труды по ОЕЯ и смежным разделам научных дисциплин. Среди них известные ученые:
М. М. Шарнин, опубликовавший около 100 работ;
А. М. Галиева – более 70;
Дж. Ш. Сулейманов – более 60;
А. Ф. Галимянов, И. П. Кузнецов, А. А. Хорошилов – порядка 50;
Г. Г. Белоногов, Л. А. Гращенко и З. Д. Усманов – более 30;
М. М. Аюпов, С. В. Клименко – более 20 публикаций и т. д.

Методы и подходы ОЕЯ актуальны для решения задач и в новых направлениях. Например, разработанный в 2018 г. прообраз системы электронного контроля e-Assessment (электронная оценка), предусматривающей автоматическую проверку ответов обучаемого на ЕЯ, при реализации проверки в качестве основной использует модель, описанную в докторской диссертации Д. Ш. Сулейманова [36]. Эта модель является основой для создания систем и информационных технологий обработки ЕЯ-текстов.

Таким образом, работы российских ученых вносят важный вклад в исследования современных ключевых технологий, в том числе по динамично развивающемуся направлению ОЕЯ.

Список источников

1. Большакова Е. И., Клышинский Э. С., Ланде Д. Э., Носков А. А., Пескова О. В., Ягунова Е. В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. Москва : Параграф, 1990. 160 с.
2. Садовская Л. Л., Гуськов А. Е., Косяков Д. В., Мухамедиев Р. И. Обработка текстов на естественном языке: обзор публикаций // Искусственный интеллект и принятие решений. 2021. № 3. С. 66–86.
3. Писляков В. В. Основные методы оценки научного знания по показателям цитирования // Социологический журнал. 2007. № 1. С. 128–140.
4. Митренина О. В., Николаев И. С., Ландо Т. М. Прикладная и компьютерная лингвистика. Москва : URSS, 2016. 320 с.
5. Яцко В. А. Предметная область компьютерной лингвистики // Вестник Иркутского государственного лингвистического университета. 2014. № 2. С. 24–35.
6. Боровикова О. И., Загорулько Ю. А., Загорулько Г. Б., Кононенко И. С., Соколова Е. Г. Разработка портала знаний по компьютерной лингвистике // КИИ-2008 : Одиннадцатая Нац. конф. по искусств. интеллекту с междунар. участием. Москва, 2008. С. 380–388.
7. Нагель О. В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Язык и культура. 2008. № 4. С. 53–54.
8. Пруцков А. В., Розанов А. К. Методы морфологической обработки текстов // Прикаспийский журнал: управление и высокие технологии. 2014. № 3. С. 119–133.
9. Пруцков А. В. Генерация и определения форм слов естественных языков на основе их последовательных преобразований // Вестник Рязанского государственного радиотехнического университета. 2009. № 27. С. 51–58.
10. Solovyev V., Ivanov V. Knowledge-driven event extraction in Russian: corpus-based linguistic resources // Computational Intelligence and Neuroscience. 2016. Vol. 2016. Art. 4183760. DOI: 10.1155/2016/4183760.
11. Батура Т. В. Семантический анализ и способы представления смысла текста в компьютерной лингвистике // Программные продукты и системы. 2016. № 4. С. 45–57. DOI: 10.15827/0236-235X.116.045-057.
12. Шарнин М. М., Сомин Н. В., Кузнецов И. П., Морозова Ю. И., Галина И. В., Козеренко Е. Б.

Статистические механизмы формирования ассоциативных портретов предметных областей на основе естественно-языковых текстов больших объемов для систем извлечения знаний // Информатика и ее применения. 2013. Т. 7, № 2. С. 92–99.

13. Золотарев О. В., Козеренко Е. Б., Шарнин М. М. Принципы построения моделей бизнес-процессов предметной области на основе обработки текстов естественного языка // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2014. № 4. С. 82–88.

14. Кулешов С. В., Зайцева А. А., Марков В. С. Ассоциативно-онтологический подход к обработке текстов на естественном языке // Интеллектуальные технологии на транспорте. 2015. № 4. С. 40–45.

15. Власов Д. Ю., Пальчунов Д. Е., Степанов П. А. Автоматизация извлечения отношений между понятиями из текстов естественного языка // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2010. Т. 8, № 3. С. 23–33.

16. Lagutina K., Lagutina N., Boychuk E., Vorontsova I., Shlakhtina O., Balyarva O., Paramonov I., Demidov H. A survey on stylometric text features // 25th Conference of Open Innovations Association (FRUCT) (November 5-8, 2019). Helsinki, 2019. P. 184–195. DOI: 10.23919/FRUCT48121.2019.8981504.

17. Горбушин Д. А., Гринченков Д. В., Мохов В. А., Хау Н. Ф. Системный анализ подходов к решению задачи идентификации тональности текста // Известия высших учебных заведений. Северо-Кавказский регион. Технические науки. 2016. № 2. С. 36–41.

18. Сарбасова А. Н. Исследование методов сентимент-анализа русскоязычных текстов // Молодой ученый. 2015. № 8. С. 143–146.

19. Сулейманов Д. Ш., Гатиатуллин А. Р. Структурно-функциональная компьютерная модель татарских морфем. Казань : Фэн, 2003. 220 с.

20. Посевкин Р. В., Бессмертный И. А. Естественно-языковой пользовательский интерфейс диалоговой системы // Программные продукты и системы. 2016. № 3. С. 5–9.

21. Яцко В. А. Алгоритмы и программы автоматической обработки текста // Вестник Иркутского государственного лингвистического университета. 2012. № 1. С. 150–160.

22. Варламов О. О., Лазарев В. М., Чувиков Д. А., Пунам Дж. О перспективах создания автономных интеллектуальных роботов на основе миварных технологий // Радиопромышленность. 2016. № 4. С. 96–105. DOI: 10.21778/2413-9599-2016-4-96-105.

23. Varlamov O. O., Adamova L. E., Eliseev D. V., Mayboroda Yu. I., Antonov P. D., Sergushin G. S., Chibirova M. O. Mivar technologies in mathematical modeling of natural language, images and human speech understanding // International Journal of Advanced Studies. 2013. Т. 3, № 3. P. 17–23. DOI: 10.12731/2227-930X-2013-3-3.

24. Loukachevitch N. V. Lashevich G., Gerasimova A. A., Ivanov V. V., Dobrov B. V. Creating Russian WordNet by conversion // Computational linguistics and intellectual technologies : pap. of annu. conf. “Dialogue 2016” (Moscow, June 1-4, 2016). Moscow, 2016. P. 405–415.

25. Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I. RuThes–Lite, a publicly available version of thesaurus of Russian language RuThes // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегод. Междунар. конф. «Диалог 2014». Москва, 2014. С. 340–350.

26. Турдаков Д., Астраханцев Н., Недумов Я., Сысоев А., Андрианов И., Майоров В., Федоренко Д., Коршунов А., Кузнецов С. Texterra: инфраструктура для анализа текстов // Труды Института системного программирования РАН. 2014. Т. 26, № 1. С. 421–438. DOI: https://doi.org/10.15514/ISPRAS-2014-26(1)-18.

27. Kuznetsov I. P., Kozerenko E. B., Charnine M. M. The System for extracting semantic information from natural language text // MLMTA’03 : proc. of Intern. conf. on machine learning, models, technologies a. applications (June 32-26, 2003). Las Vegas, 2003. P. 75–80.

28. Пруцков А. В. Морфологический анализ и синтез текстов посредством преобразований форм слов // Вестник Рязанской государственной радиотехнической академии. 2004. № 15. С. 70–75.

29. Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. H., Zuev K. A. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегод. Междунар. конф. «Диалог 2012». Москва, 2012. С. 91–103.

30. Козлов П. Ю. Методы автоматизированного анализа коротких неструктурированных текстовых документов // Программные продукты и системы. 2017. № 1. С. 100–105.

31. Москвина А. Д., Орлова Д. Н., Митрофанова О. А. Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK // Интернет и современное общество : тр. объедин. науч. конф. IMS-2016. Санкт-Петербург, 2016. С. 44–54.

32. Ермаков А. Е. Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии. 2009. №. 7. С. 50–55.

33. Диковицкий В. В., Шишаев М. Г. Обработка текстов естественного языка в моделях поисковых систем // Труды Кольского научного центра РАН. 2010. № 3. С. 29–34.

34. Максимов В. Ю., Клышинский Э. С., Антонов Н. В. Проблема понимания в системах искусственного интеллекта // Новые информационные технологии в автоматизированных системах. 2016. № 19. С. 43–60.

35. Liddy E. D. Natural language processing // Encyclopedia of library and information science. New York, 2003. P. 2126–2136.

36. Сулейманов Д. Ш. Системы и информационные технологии обработки естественно–языковых текстов на основе прагматически-ориентированных лингвистических моделей : автореф. дис. ... д-ра техн. наук : 05.13.14. Казань, 2000. 43 с.

References

1. Bolshakova E. I., Klyshinsky E. S., Lande D. E., Noskov A. A., Peskova O. V., Yagunova E. V. *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp’yuternaya lingvistika* [Automatic text processing in natural language and computational linguistics]. Moscow, Paragraph, 1990. 160 p. (In Russ.).

2. Sadovskaya L. L., Guskov A. E., Kosyakov D. V., Mukhamediev R. I. Text processing in natural language: a review of publications. *Iskusstvennyi intellekt i prinyatie reshenii,* 2021, 3: 66–86. (In Russ.).

3. Pislyakov V. V. Basic methods of evaluation of scientific knowledge by citation indicators. *Sotsiologicheskii zhurnal*, 2007, 1: 128–140. (In Russ.).

4. Mitrenina O .V., Nikolaev I. S., Lando T. M. *Prikladnaya i komp'yuternaya lingvistika* [Applied and computational linguistics]. Moscow, URSS, 2016. 320 p. (In Russ.).

5. Yatsko V. A. The subject area of computational linguistics. *Vestnik Irkutskogo gosudarstvennogo lingvisticheskogo universiteta,* 2014, 2: 24–35. (In Russ.).

6. Borovikova O. I., Zagorulko, Yu. A., Zagorulko G. B., Kononenko I. S., Sokolova E. G. Development of a knowledge portal on computational linguistics. *KII-2008: Odinnadtsataya Nats. konf. po iskusstv. intellektu s mezhdunar. uchastiem.* Moscow, 2008: 380–388. (In Russ.).

7. Nagel O. V. Corpus linguistics and its use in computerized language teaching. *Yazyk i kul'tura*, 2008, 4: 53–54. (In Russ.).

8. Prutskov A. V., Rozanov A. K. Methods of morphological text processing. *Prikaspiiskii zhurnal: upravlenie i vysokie tekhnologii,* 2014, 3: 119–133. (In Russ.).

9. Prutskov A. V. Generation and definition of word forms of natural languages based on their successive transformations. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo uiversiteta*, 2009, 27: 51–58. (In Russ.).

10. Solovyev V., Ivanov V. Knowledge–driven event extraction in Russian: corpus-based linguistic resources. *Computational Intelligence and Neuroscience,* 2016, 2016: 4183760. DOI: 10.1155/2016/4183760.

11. Batura T. V. Semantic analysis and ways of representing the meaning of a text in computational linguistics. *Programmnye produkty i sistemy*, 2016, 4: 45–57. DOI: 10.15827/0236-235X.116.045-057. (In Russ.).

12. Sharnin M. M., Somin N. V., Kuznetsov I. P., Morozova Yu. I., Galina I. V., Kozerenko E. B. Statistical mechanisms for the formation of associative portraits of subject areas based on natural language texts of large volumes for knowledge extraction systems. *Informatika i ee primeneniya,* 2013, 7(2): 92–99. (In Russ.).

13. Zolotarev O. V., Kozerenko E. B., Sharnin M. M. Principles of building models of business processes in the subject area based on natural language text processing. *Vestnik Rossiiskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz i upravleniye*, 2014, 4: 82–88. (In Russ.).

14. Kuleshov S. V., Zaitseva A. A., Markov V. S. Associative–ontological approach to text processing in natural language. *Intellectual Technologies on Transport,* 2015, 4: 40–45. (In Russ.).

15. Vlasov D. Yu., Palchunov D. E., Stepanov P. A. Automation of extraction of relations between concepts from natural language texts. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii*, 2010, 8(3): 23–33. (In Russ.).

16. Lagutina K. Lagutina N., Boychuk E., Vorontsova I., Shlakhtina O., Balyarva O., Paramonov I., Demidov H. A survey on stylometric text features. *5th Conference of Open Innovations Association (FRUCT) (November 5-8, 2019).* Helsinki, 2019: 184–195. DOI: 10.23919/FRUCT48121.2019.8981504.

17. Gorbushin D. A., Grinchenkov D. V., Mokhov V. A., Khau N. F. System analysis of approaches to solving the problem of identifying the tonality of the text. *Izvestiya vysshikh uchebnykh zavedenii. Severo–Kavkazskii region. Tekhnicheskie nauki,* 2016, 2: 36–41. (In Russ.).

18. Sarbasova A. N. Investigation of methods of sentimental analysis of Russian–language texts. *Molodoy uchenyy,* 2015, 8: 143–146. (In Russ.).

19. Suleimanov D. Sh., Gatiatullin A. R. *Strukturno-funktsional'naya komp'yuternaya model' tatarskikh morfem* [Structural and functional computer model of Tatar morphemes]. Kazan', Fen, 2003. 220 p. (In Russ.).

20. Posevkin R. V., Bessmertnyi I. A. Natural language user interface of a dialog system. *Programmnye produkty i sistemy*, 2016, 3: 5–9. (In Russ.).

21. Yatsko V. A. Algorithms and programs of automatic text processing. *Vestnik Irkutskogo gosudarstvennogo lingvisticheskogo univers*iteta, 2012, 1: 150–160. (In Russ.).

22. Varlamov O. O., Lazarev V. M., Chuvikov D. A., Punam D. On the prospects of creating autonomous intelligent robots based on mivar technologies. *Radiopromyshlennost',* 2016, 4: 96–105. DOI: 10.21778/2413-9599-2016-4-96-105. (In Russ.).

23. Varlamov O. O., Adamova L. E.., Eliseev D. V., Mayboroda Yu. I., Antonov P. D., Sergushin G. S., Chibirova M. O. Mivar technologies in mathematical modeling of natural language, images and human speech understanding. *International Journal of Advanced Studies*, 2013, 3(3): 17–23. DOI: 10.12731/2227-930X-2013-3-3.

24. Loukachevitch N. V., Lashevich G., Gerasimova A. A., Ivanov V. V., Dobrov B. V. Creating Russian WordNet by conversion. *Computational linguistics and intellectual technologies: pap. of annu. conf. "Dialogue 2016" (Moscow, June 1-4, 2016).* Moscow, 2016: 405–415.

25. Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I. RuThes–Lite, a publicly available version of thesaurus of Russian language RuThes. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po materialam ezhegod. Mezhdunar. konf. «Dialog 2014».* Moscow, 2014: 340–349.

26. Turdakov D. Astrakhantsev N., Nedumov Y., Sysoev A., Andrianov I., Mayorov V., Fedorenko D., Korshunov A., Kuznetsov S. Texterra: infrastructure for text analysis. *Trudy Instituta sistemnogo programmirovaniya RAN,* 2014, 26(1): 421–438. (In Russ.).

27. Kuznetsov I. P., Kozerenko E. B., Charnine M. M. The system for extracting semantic information from natural language text. *MLMTA'03 : proc. of Intern. conf. on machine learning, models, technologies a. applications (June 32–26, 2003).* Las Vegas, 2003: 75–80.

28. Prutskov A. V. Morphological analysis and synthesis of texts by means of transformations of word forms. *Vestnik Ryazanskoi gosudarstvennoi radiotekhnicheskoi akademii,* 2004, 15: 70–75. (In Russ.).

29. Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. H., Zuev K. A. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po materialam ezhegod. Mezhdunar. konf. «Dialog 2012». Moscow,* 2012: 91–103.

30. Kozlov P. Yu. Methods of automated analysis of short unstructured text documents. *Programmnye produkty i sistemy,* 2017, 1: 100–105. (In Russ.).

31. Moskvina A. D., Orlova D. N., Mitrofanova O. A. Development of a parser core for the Russian language

based on NLTK libraries. *Internet i sovremennoe obshchestvo: tr. ob"edin. nauch. konf. IMS-2016.* Saint Petersburg, 2016: 44–54. (In Russ.).

32. Ermakov A. E. Extraction of knowledge from the text and their processing: state and prospects. *Informatsionnye tekhnologii*, 2009, 7: 50–55. (In Russ.).

33. Dikovitsky V. V., Shishaev M. G. Processing texts in natural language models search engines. *Trudy Kol'skogo nauchnogo tsentra RAN,* 2010, 3: 29–34. (In Russ.).

34. Maximov V. Yu., Klyshinski E. S., Antonov N. V. The problem of understanding artificial intelligence systems. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh,* 2016, 19: 43–60. (In Russ.).

35. Liddy E. D. Natural language processing. *Encyclopedia of library and information science*. New York, 2003: 2126–2136.

36. Suleimanov D. Sh. *Sistemy i informatsionnye tekhnologii obrabotki estestvenno-yazykovykh tekstov na osnove pragmaticheski-orientirovannykh lingvisticheskikh modelei: avtoref. dis. ... d-ra tekhn. nauk* [Systems and information technologies for processing natural language texts based on pragmatically oriented linguistic models: diss. abstr.]. Kazan', 2000. 43 p. (In Russ.).

Статья поступила в редакцию	07.11.2021	Received	07.11.2021
Получена после доработки	20.12.2021	Revised	20.12.2021
Принята для публикации	11.01.2022	Accepted	11.01.2022