ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 81’25:001.4

Ю.И. Бутенко, Ю.В. Строганов, А.М. Сапожков

Система извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов

Описано создание системы извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов на основе структурных моделей терминологических единиц с последующим применением ограничений на лексическую сочетаемость общеупотребительных слов и терминов. Проанализирована формальная структура терминологических единиц. Выделены структурные модели многокомпонентных терминов и изложен метод извлечения терминов из научно-технических текстов на английском и русском языках на основе структурных моделей многокомпонентных терминов. Описан алгоритм работы системы извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов. В качестве примера приведены результаты работы такой системы на текстах по космонавтике, а также проанализированы допущенные ошибки.

Ключевые слова: *научно-технические тексты, многокомпонентный термин, извлечение терминов, параллельные тексты, научно-технические тексты*

DOI: 10.36535/0548-0027-2022-09-3

ВВЕДЕНИЕ

Стремительное развитие и внедрение технологий искусственного интеллекта и технологий автоматической обработки текстовой информации способствуют развитию лингвистических баз данных как основы создания прикладных программных средств, проведения лингвистических исследований источников информации при решении ряда прикладных задач, где однозначная и упорядоченная терминология имеет особую значимость. Под терминологической базой данных принято понимать организованную в соответствии с определёнными правилами и поддерживаемую в памяти компьютера совокупность данных, характеризующую актуальное состояние некоторой предметной области и используемую для удовлетворения информационных потребностей пользователей [1]. Каждый термин дополнен информацией о его значении, эквивалентах в других языках, кратких формах, синонимах, сведениях об области применения. По целевому назначению терминологические базы данных разделяют на одноязычные, предназначенные для обеспечения информацией о стандартизованной и рекомендованной терминологии, и многоязычные, ориентированные на работы по переводу научно-технической литературы и документации [2, 3].

Создание терминологических баз данных представляет собой сложный и трудоемкий процесс, требующий значительного количества времени на их создание и обновление, что особенно важно для развивающихся терминологий таких предметных областей, как авиация, космонавтика, нанотехнологии, биоинженерия, информационные технологии и многих других. Одним из наиболее время-затратных процессов является ручной сбор иллюстративного материла – извлечение специальной терминологии из коллекций текстов, что требует наличия средств автоматического извлечения многокомпонентных терминов при обработке научно-технических текстов.

Существующие программные средства автоматического извлечения терминов основаны на лингвистических и статистических методах. Могут быть использованы и методы машинного обучения [4], сложность их реализации вызвана необходимостью наличия огромных массивов обучающих данных, которые могут отсутствовать для определенной предметной области. В основе лингвистических методов лежит использование грамматики лексико-синтаксических шаблонов, представляющих собой структурные модели лингвистических конструкций [5, 6]. Статистический подход заключается в нахождении n-грамм слов по заданным частотным характеристикам [7, 8]. Гибридный под-

ход для выделения терминологических сочетаний, объединяющий лингвистический и статистический методы, заключается в предварительном описании моделей, по которым могут быть построены термины для последующего поиска их в коллекции текстов [9].

Выравнивание терминологических единиц в параллельных текстах обычно осуществляется в два этапа: сначала выделяют терминологические единицы на каждом языке отдельно, затем один из одноязычных списков терминов-кандидатов интерпретируется как язык источника, и для каждого термина-кандидата на языке источнике предлагаются потенциально эквивалентные термины в списке терминов-кандидатов на языке перевода [10].

Цель настоящей статьи – описание системы извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов на основе структурных моделей терминологических словосочетаний, а также выявление лексическо-грамматических ограничений на сочетаемость общеупотребительной лексики с терминами.

ДЛИННА И СТРУКТУРА ТЕРМИНОЛОГИЧЕСКИХ ЕДИНИЦ

Участие языкового субстрата в структуре термина, существование термина в виде лексической единицы определенного естественного языка проявляется в том, что термин обладает определенной формальной структурой. Он может быть однокомпонентным и состоять из ключевого слова, или представлять собой терминологическую группу, в состав которой входит ключевое слово или ядро группы, одно или несколько левых определений и одно или несколько правых или предложных определений, которые уточняют или модифицируют смысл терминологической единицы [11].

В процессе автоматического извлечения терминов из научно-технических текстов наибольшую сложность представляют многокомпонентные термины – терминологические словосочетания, созданные лексическим и синтаксическим способами, т. е. словосочетания, составленные по определенным моделям. Способ образования терминов в виде цепочки слов часто используется на практике [12].

Наиболее продуктивным способом номинации является создание составных терминов, состоящих из двух, трех и более компонентов. Скорее всего, это связано с увеличением семантической дифференциации и мотивированности терминов, что обусловлено развитием каждой отдельной предметной области [13].

Структурные модели русскоязычных многокомпонентных терминов, отраженные в системе разметки многокомпонентных терминов, включают следующие модели: имя прилагательное + имя существительное; причастие + имя существительное; имя существительное + имя существительное в родительном падеже; имя существительное + имя существительное в творительном падеже; имя существительное + имя существительное в родительном падеже + имя существительное в творительном падеже; имя прилагательное + имя прилагательное + имя существительное; причастие + имя прилагательное + имя существительное, имя прилагательное + имя существительное + имя существительное в творительном падеже; причастие + имя существительное + имя существительное в творительном падеже; имя прилагательное + имя существительное + имя существительное в винительном падеже; причастие + имя существительное + имя существительное в винительном падеже; имя прилагательное + имя существительное + наречие; имя существительное + имя существительное в родительном падеже + имя существительное в родительном падеже + имя существительное в родительном падеже; имя прилагательное + существительное + имя прилагательное + имя существительное в творительном падеже; имя прилагательное + имя существительное + имя прилагательное + имя существительное в предложном падеже; имя прилагательное + имя существительное + имя прилагательное + "предлог" + имя существительное в творительном падеже; имя прилагательное + имя существительное + имя существительное в предложном падеже + имя существительное в родительном падеже; имя прилагательное + имя существительное + предлог + имя существительное в творительном падеже + имя существительное в родительном падеже; имя прилагательное + имя прилагательное + имя существительное + предлог + имя существительное в творительном падеже; имя прилагательное + имя существительное + имя существительное в творительном падеже + предлог + имя существительное в творительном падеже.

Структурные модели англоязычных многокомпонентных терминов, отраженные в системе разметки многокомпонентных терминов, включают следующие модели: noun + noun; adjective/participle + noun; noun + “of” + noun; noun + noun + noun; noun + noun + “of” + noun; noun + “of” + noun + noun; noun + adjective/participle + noun; noun + “of”/”with” + adjective/participle + noun; adjective/participle + noun + noun; adjective/participle + noun + “of”/”for” + noun; adjective/participle + adjective/participle + noun; adjective/participle + “and” + adjective/participle + noun; noun + noun + noun + noun; noun + “for” + noun + “of” + noun + noun; noun + adjective/participle + noun + noun; noun + noun + “with” + adjective/participle + noun; adjective/participle + noun + adjective/participle + noun; adjective/participle + noun + “of”/”for” + adjective/participle + noun; adjective/participle + noun + noun + noun; adjective/participle + noun + “of”/”for” + noun + “and” + noun; adjective/participle + noun + “of”/”for” + noun + noun; adjective/participle + adjective/participle + noun + noun; adjective/participle + adjective/participle + noun + “of” + noun; adjective/participle + adjective/participle + adjective/participle + noun; noun + noun + noun + noun + noun; noun + noun + noun + adjective/participle + noun; noun + noun + noun + “of”/”for” + noun + noun; noun + noun + noun + “of”/”for” + adjective/participle + noun; noun + noun + “with” + adjective/participle + adjective/participle + noun; adjective/participle + noun + “of”/”for” + noun + “and” + adjective/participle + noun; adjective/participle + noun + “of”/”for” + noun + “and” + noun + noun; noun + noun + “with” + adjective/participle + adjective/participle+ adjective/participle + noun; adjective/participle + noun + “of” + adjective/participle + noun + noun + noun.

МЕТОД ИЗВЛЕЧЕНИЯ МНОГОКОМПОНЕНТНЫХ ТЕРМИНОВ НА ОСНОВЕ СТРУКТУРНЫХ МОДЕЛЕЙ

Для демонстрации метода по извлечению терминов-кандидатов мы взяли небольшой отрывок из предметной области «Авиация и космонавтика»:

Transfers in the central Newtonian field are considered under the assumption that low thrust that is constant in magnitude is zeroed when spacecraft with solar batteries enters the Earth's shadow.

Отбор терминов-кандидатов состоит из нескольких этапов. На первом этапе выполняется морфологический разбор текста, где каждому слову приписываются его морфологические характеристики:

Transfers$^{n, v}$ in$^{prep, adv, v, n, adj}$ the$^{art, adv}$ central^{adj} Newtonian$^{adj, n}$ field$^{n, v}$ are$^{v, n}$ considered^{v} under$^{prep, adv, adj}$ the$^{art, adv}$ assumption^{n} that$^{conj, det, pron, adv, n}$ low$^{adj, n, adv, v}$ thrust$^{n, v}$ that$^{conj, det, pron, adv, n}$ is$^{v, n}$ constant$^{adj, n}$ in$^{prep, v, adv, n, adj}$ magnitude^{n} is$^{v, n}$ zeroed^{v} when$^{adv, conj, pron, n, interj}$ spacecraft^{n} with$^{prep, adv, n}$ solar$^{adj, n}$ batteries^{n} enters$^{n, v}$ the$^{art, adv}$ Earth's$^{n, v}$ shadow$^{n, v, adj}$.

На втором этапе по результатам морфологического анализа необходимо убрать часть слов, которые не входят в терминологическую систему (глаголы, знаки препинания и т. д.), чтобы выявить допустимые терминологические словосочетания:

Transfers$^{n, v}$ in$^{prep, adv, v, n, adj}$ ~~the$^{art, adv}$~~ central^{adj} Newtonian$^{adj, n}$ field$^{n, v}$ ~~are$^{v, n}$~~ ~~considered^{v}~~ under$^{prep, adv, adj}$ ~~the$^{art, adv}$~~ assumption^{n} ~~that$^{conj, det, pron, adv, n}$~~ low$^{adj, n, adv, v}$ thrust$^{n, v}$ ~~that$^{conj, det, pron, adv, n}$~~ ~~is$^{v, n}$~~ constant$^{adj, n}$ in$^{prep, v, adv, n, adj}$ magnitude^{n} ~~is$^{v, n}$~~ ~~zeroed^{v}~~ ~~when$^{adv, conj, pron, n, interj}$~~ spacecraft^{n} with$^{prep, adv, n}$ solar$^{adj, n}$ batteries^{n} ~~enters$^{n, v}$~~ ~~the$^{art, adv}$~~ Earth's$^{n, v}$ shadow$^{n, v, adj}$.

На третьем этапе сравниваются цепочки слов допустимых терминологических словосочетаний с терминологическими моделями – извлекаются последовательности слов, которые соответствуют морфосинтаксическим шаблонам однословных терминов и терминологических словосочетаний. Таким образом в рассматриваемом примере получен следующий список терминов-кандидатов:

1. *Transfers in (the) central Newtonian field (N+in+ADJ+ADJ+N)*
2. *Central Newtonian field (ADJ+ADJ+N)*
3. *Low thrust (ADJ+N)*
4. *Constant in magnitude (N+in+N)*
5. *Spacecraft with solar batteries (N+with+ADJ+N)*
6. *Solar batteries (ADJ+N)*
7. *Earth's shadow (N+N)*

На этом этапе выполняется проверка ряда лингвистических условий, например: терминологическое словосочетание как в английском, так и русском языке не может начинаться с предлога или состоять из одного элемента, кроме существительного. В первом случае предлог исключается из словосочетания, а во втором слово удаляется из списка терминов-кандидатов. В английском языке также опускаются артикли.

На четвёртом этапе оцениваются выбранные слова и словосочетания. Сначала каждое словосочетание проверяется по терминологическому словарю. Если такое словосочетание отсутствует в словаре, то оставшиеся словосочетания проходят проверку на стоп-слова. Под стоп-словами в разрабатываемой нами системе понимаются слова, которые сочетаются с терминами таким образом, что рассматриваемая синтаксическая конструкция совпадает со структурной моделью многокомпонентного термина, например: ADJ+N+N: *artificial information technology* (термин) и *new information technology* (общеупотребительное слово + термин). В таком случае необходимо сформировать список общеупотребительных слов с оценочной или внепредметной семантикой (*new, modern, developed, analyzed и т.п.*), чтобы не учитывать их при выборе терминов.

Для русского языка многокомпонентные термины извлекаются аналогичным образом [14].

На последнем этапе термину приписываются специальные теги: выделяется ядерный элемент, предметная область и ряд других лексических и грамматических характеристик. Система извлечения многокомпонентных терминов из параллельных текстов позволяет идентифицировать термины-кандидаты, которые не зарегистрированы в существующих переводных терминологических словарях.

СИСТЕМА ИЗВЛЕЧЕНИЯ АНГЛО- И РУССКОЯЗЫЧНЫХ МНОГОКОМПОНЕНТНЫХ ТЕРМИНОВ ИЗ ПАРАЛЛЕЛЬНЫХ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

Для извлечения многокомпонентных терминов из текстов на русском и английском языках, нами разработано приложение на языке Python с использованием библиотек tkinter, nltk, pymorphy2, os, sys. Программа предоставляет пользователю интерфейс для ввода текстов, подлежащих анализу. Из полученных текстов путём морфологического анализа слов, в него входящих, а также их взаимного расположения, выделяются словосочетания, которые могут быть многокомпонентными терминами. Далее пользователь может вручную отобрать термины, классифицировать их и сохранить в базе данных, состоящей из нескольких текстовых файлов, с которыми работает программа и которые можно анализировать с помощью функции поиска. Программная реализация системы извлечения многокомпонентных терминов из параллельных текстов представляет собой интерфейс для поиска терминов, соответствующих определённому набору характеристик.

На рис. 1 и 2 показан алгоритм работы системы извлечения многокомпонентных терминов из параллельных научно-технических текстов, представленных в виде IDEF0-диаграмм второго и третьего уровней декомпозиции разработанного алгоритма.

Интерфейс программы для извлечения многокомпонентных терминов из параллельных научно-технических текстов представлен на рис. 3. Работа в программе начинается с выбора способа ввода текстов: загрузка в виде файлов в форматах doc, docx, txt, pdf, или ручной ввод текстов на английском и русском языках (рис. 3а).

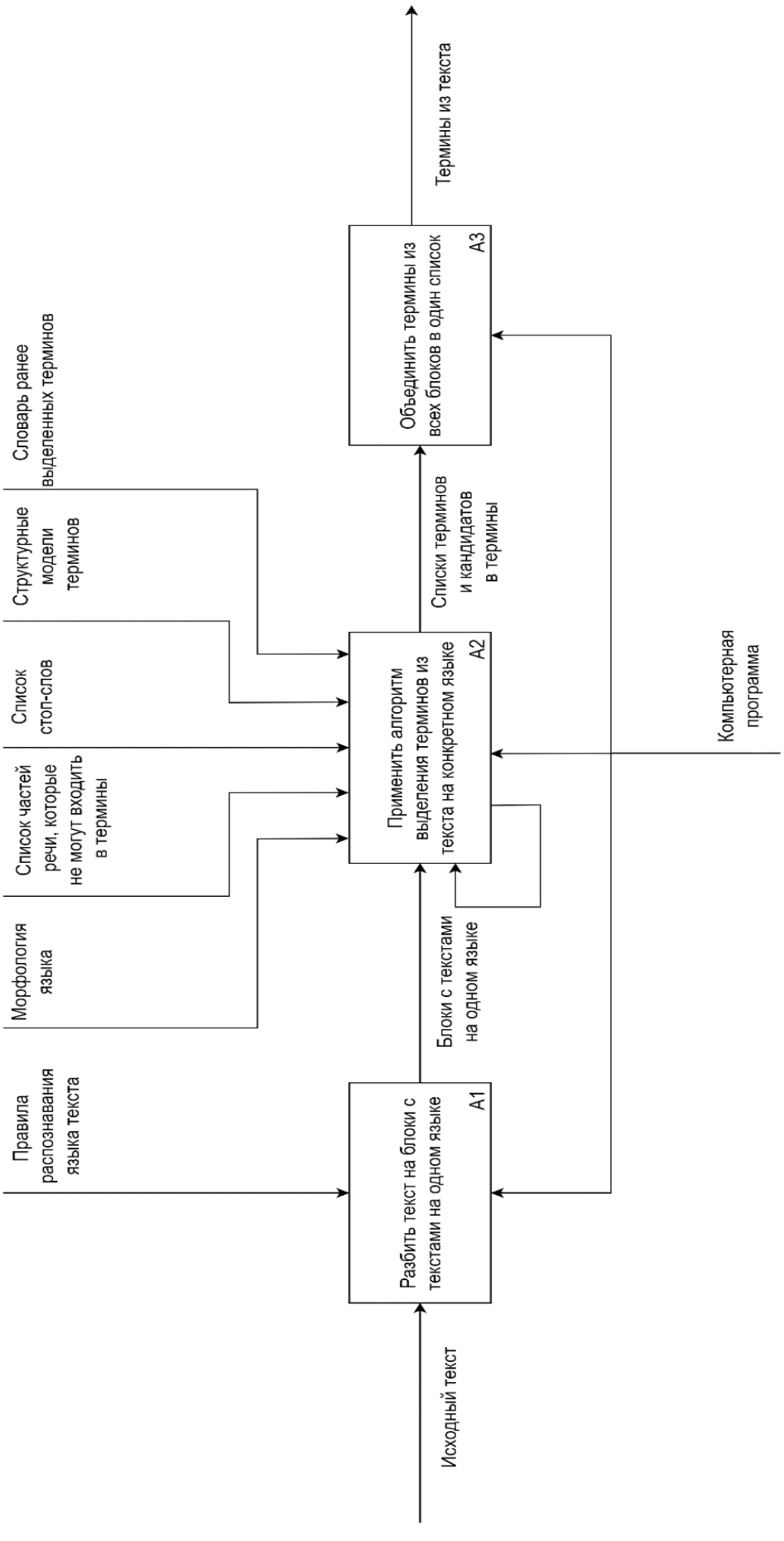


Рис. 1. Второй уровень декомпозиции алгоритма работы системы извлечения многокомпонентных терминов

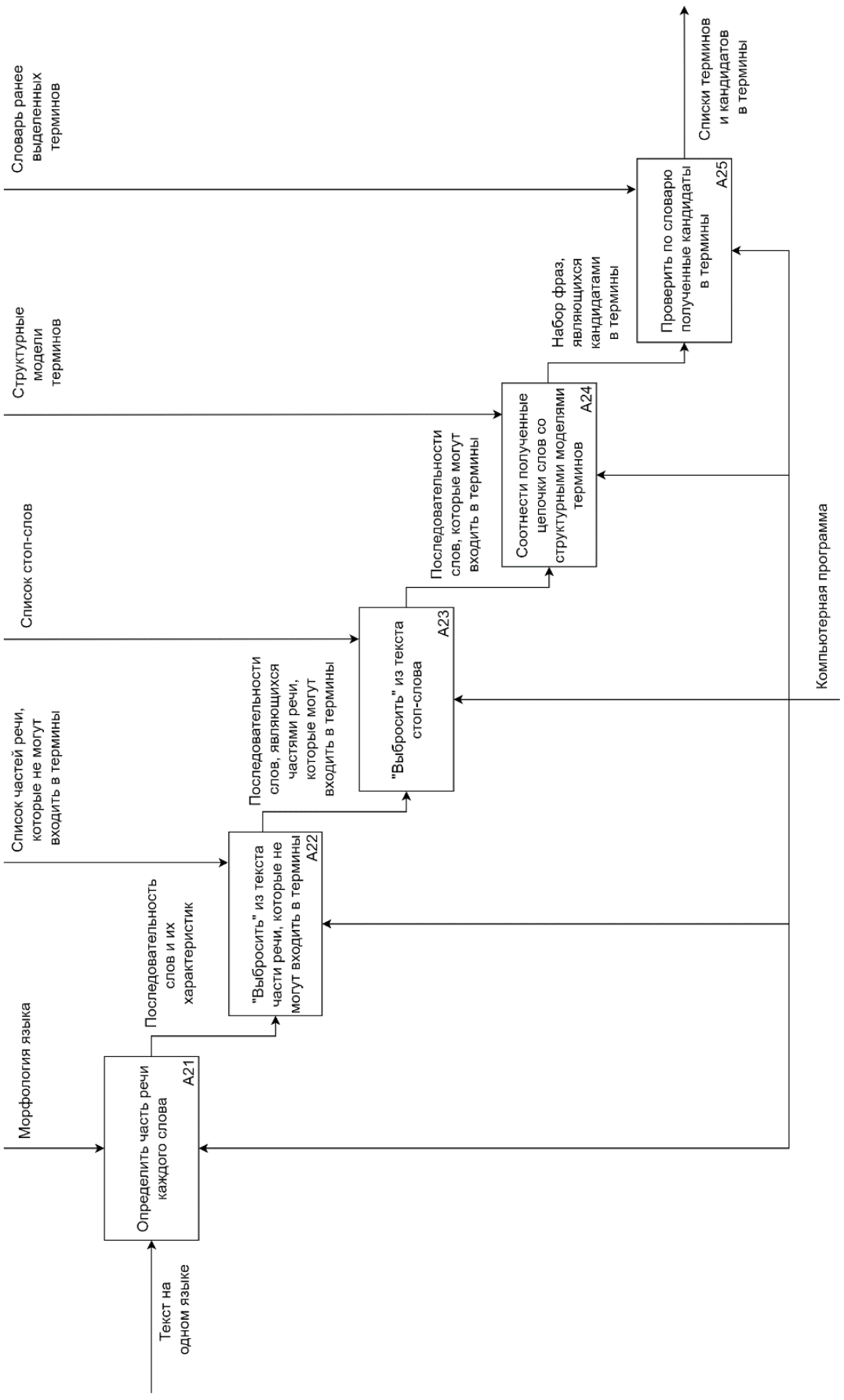
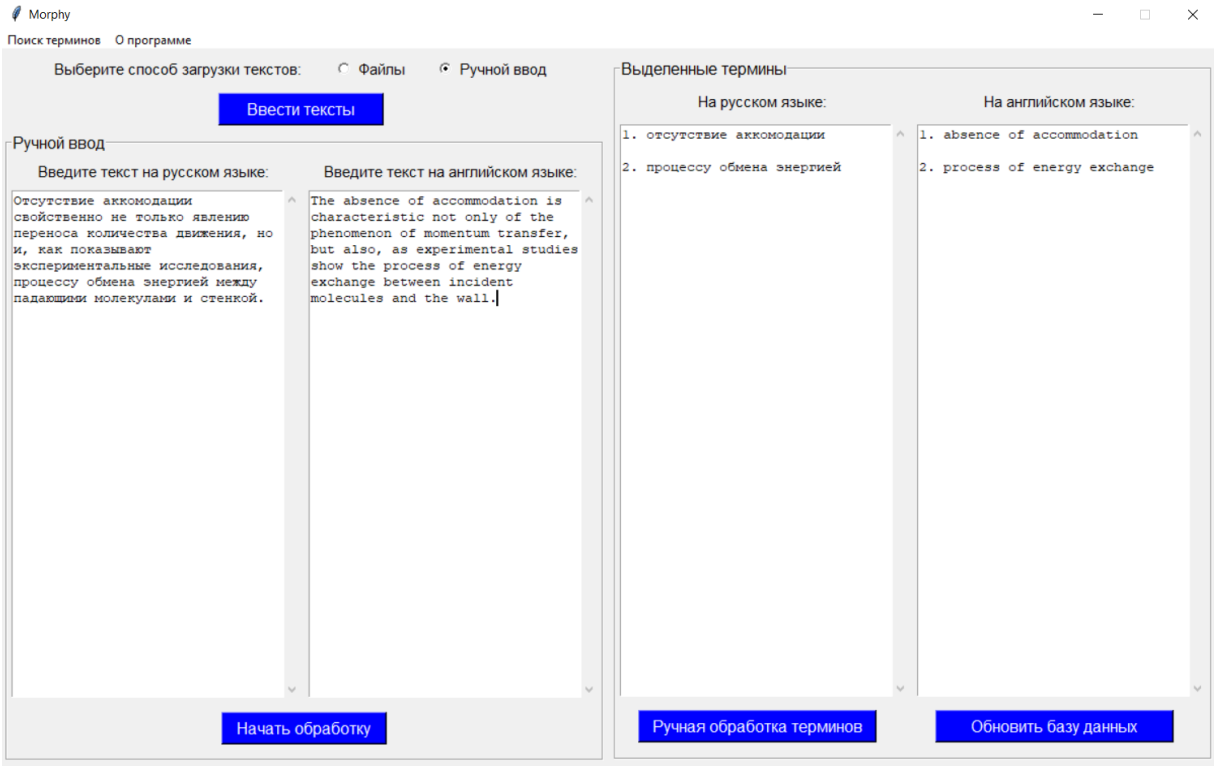


Рис. 2. Третий уровень декомпозиции алгоритма работы системы извлечения многокомпонентных терминов



а) *б)*

Рис. 3. Интерфейс системы извлечения многокомпонентных терминов из параллельных научно-технических текстов

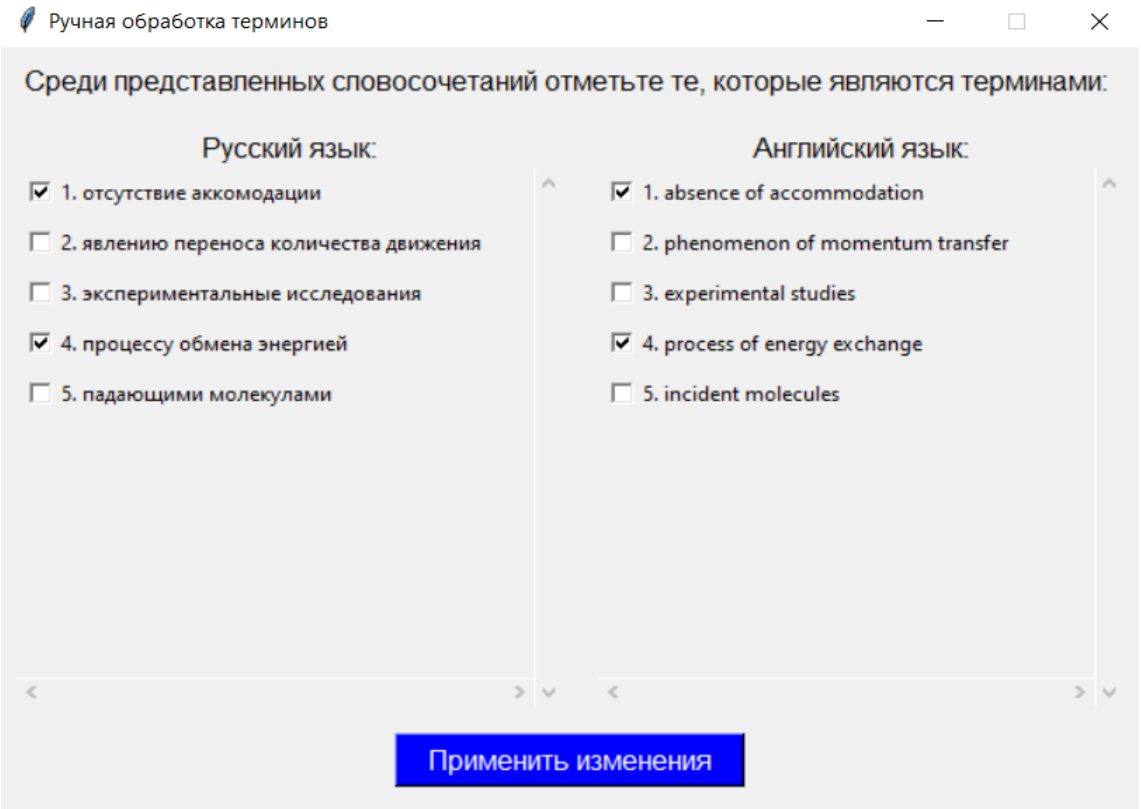


Рис. 4. Диалоговое окно выбора терминологических единиц

В программе реализована функция выявления количества повторений терминов в каждом тексте – повторяющийся по тексту термин будет выведен один раз, а количество его употреблений указывается в скобках рядом с термином. Термины могут извлекаться из текстов как на двух языках, так и на одном языке. После ввода данных следует нажать «Начать обработку».

По окончании обработки текстов откроется диалоговое окно, представленное на рис. 4, где пользователю будет дана возможность вручную выбрать требуемые терминологические единицы.

После выделения терминологических единиц пользователь может приписать им лексические и грамматические характеристики (рис. 5), и впоследствии сохранить в базу данных. Результаты отбора терминов появятся в окошке, представленном на рис. 3б. На основе грамматических и лексических характеристик можно проводить поиск терминов в базе данных по определенным параметрам.

Возможности практического использования предлагаемой нами системы извлечения многокомпонентных терминов рассмотрим на примере научно-технической статьи по космонавтике и ее переводной версии. В табл. 1 представлены результаты такого анализа.

Ручная проверка результатов работы рассматриваемой системы показала (табл. 2), что при обработке научно-технических текстов случаются и ошибки.

Анализируя результаты работы предлагаемой нами системы, можно выделить некоторые положительные и отрицательные моменты. Из текста на русском языке программа довольно хорошо извлекает термины, автоматически преобразовывая их ядерные элементы в именительный падеж единственного числа, однако не распознает аббревиатуры как в русском тексте, так и в английском, а также часто определяет только часть многокомпонентных терминов. Возможная причина этого в том, что базе нет определенной модели.

Дальнейшая перспектива развития системы излечения многокомпонентных терминов из параллельных научно-технических текстов представляется нам в исследовании структурных моделей литеральных терминов, например: *К метод, Е-слой, N тело, точка либрации L2, F-область,* и номенклатурных наименований, например, *Бион М-1, Луна-25, "Матрешка-Р", Космос-321* [15, 16].

Таблица 1

Результаты анализа формальной структуры терминов научно-технической статьи по космонавтике

Морозов В. М. Каленова В. И. Управление спутником при помощи магнитных моментов: управляемость и алгоритмы стабилизации // Космические исследования. – 2020. – № 3. – С. 199-207	Язык	
	русский	английский
Всего слов в тексте	1037	1404
Найдено терминов-кандидатов всего	285	302
Найдено терминов вручную	56	35
Из них **однокомпонентных**	11	6
RU: спутник, магнитометр, орбита, время, работоспособность, эффективность, матрица, нуль, столбцы, моделирование, коэффициент EN: satellite, magnetometers, system, controllability, time, dimensions		
Из них **двухкомпонентных**	32	13
RU: круговая орбита, линеаризованная система, уравнение движения, магнитная система, малый спутник, магнитная катушка, космический аппарат, магнитный момент, линеаризованная модель, управляющий момент, геомагнитное поле, нестационарная система, стационарная система, математическое моделирование, гравитационное поле, система координат, уравнение движения, орбитальная система, плоскость орбиты, направляющий косинус, единичный вектор, угол наклона, плоскость орбиты, плоскость экватора, угол поворота, линеаризованное уравнение, управляемое движение, система уравнений, определитель матрицы, независимые столбцы, матрица коэффициентов, квадратичный функционал EN: LQR method, magnetic systems, magnetic coils, control moment, floquet theory, stationary system, mathematical modeling, gravitational field, circular orbit, coordinate systems, OY-axis, orbital plane, independent columns		
Из них **трехкомпонентных**	10	10
RU: линейная нестационарная система, собственный магнитный момент, внешнее магнитное поле, линейная нестационарная система, приближенная стационарная система, замкнутая периодическая система, главная центральная ось, постоянное магнитное поле, линейная обратная связь, стационарная управляемая система EN: magnetic orientation system, intrinsic magnetic moment, external magnetic field, spacecraft attitude control, angle of inclination, equations of motion, determinant of matrix, time-varying system, moments of inertia, resulting time invariant		
Из них **четырехкомпонентных**	0	2
RU: - EN: stabilization of relative equilibrium, linearized system of equations		

Таблица 2

Анализ ошибок системы извлечения многокомпонентных терминов из параллельных научно-технических текстов

Морозов В. М. Каленова В. И. Управление спутником при помощи магнитных моментов: управляемость и алгоритмы стабилизации // Космические исследования. 2020 №3. С. 199-207		Язык	
		русский	английский
Всего действительных терминов		56	35
Всего терминов, не извлеченных системой		3	4
Результат программы	Термин	Описание проблемы и возможное решение	
1. Снабженная магнитная системой	Магнитная система ориентации	Программа, возможно, определила, что трехкомпонентный термин должен включать в себя именно прилагательное «снабженная», а не существительное «ориентации» по модели прил+прил+сущ, данную модель стоит поменять на прил+сущ+сущ.	
2. Function of geomagnetic field	Geomagnetic field	Были присоединены ненужные существительное «Function» и предлог «of», возможно ошибка в модели, стоит создать четкую модель двухкомпонентного термина прил+сущ	
3. Original nonstationary system	Nonstationary system	Ситуация, идентичная с пунктом 4, из модели прил+прил+сущ следует убрать первое прилагательное	
4. ЛКР	ЛКР	Программа не определяет аббревиатуру как термин, следует добавить соответствующую модель	
5. LQR	LQR	Ситуация, идентичная с пунктом 4	
6. Однородная система с постоянными коэффициентами	Линейная однородная система	Программа ошибочно определила термин, не определилось нужное прилагательное и добавились 3 ненужные части речи, в модели пятикомпонентного термина присутствует ошибка, оттуда нужно исключить предлог	
7. MATLAB	MATLAB	Программа не определяет аббревиатуру как термин, следует добавить соответствующую модель	

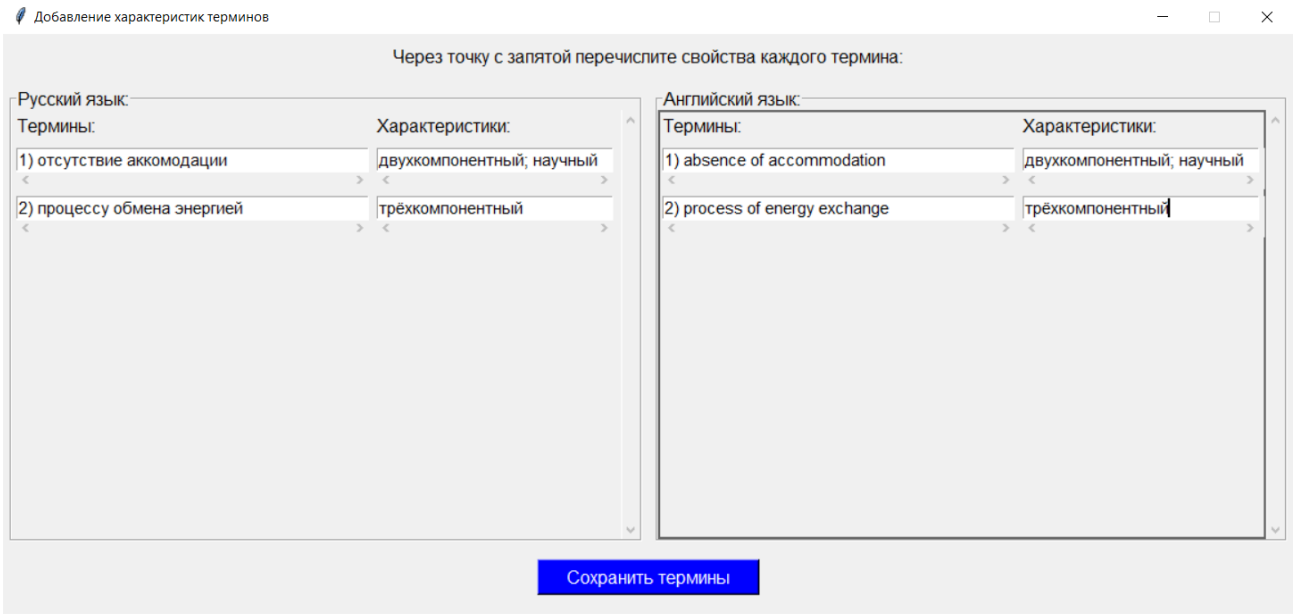


Рис. 5. Окно ввода лексической и грамматической информации к терминам

ВЫВОДЫ

В настоящей статье представлена система извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов на основе структурных моделей терминологических единиц с последующим применением ограничений на лексическую сочетаемость общеупотребительных слов и терминов. В языке терминологическая единица формально может быть представлена словом или словосочетанием, включающим в себя ядерный элемент и несколько левых и / или правых определений. Кроме того перечислены структурные модели и изложен метод извлечения многокомпонентных терминов из научно-технических текстов на английском и русском языках. Изложен метод извлечения многокомпонентных терминов из научно-технических текстов на английском языке на основе структурных моделей многокомпонентных терминов. Программно реализованная система извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов может работать как с одноязычными, так и с параллельными текстами. Алгоритм работы системы включает этапы ручной обработки научно-технических терминов. В качестве примера приведены результаты работы системы извлечения многокомпонентных терминов из параллельных текстов по космонавтике объемом 1307 слов в русском языке и 1404 – в английском. Обработка текстов позволила значительно снизить объем работ по поиску терминов в текстах – до 285 в русском и 304 английском языках. Результаты ручной проверки работы системы помогли выявить и ошибки в работе системы, которые чаще всего были вызваны сложностями при обработке аббревиатур, которые могут как представлять собой как термин-слово, так и входить в состав многокомпонентного термина.

СПИСОК ЛИТЕРАТУРЫ

1. Когаловский М.Р Энциклопедия технологий баз данных. – Москва: Финансы и статистика, 2002. – 800 с.
2. Алферова Т.К., Леонов А.В., Филиппов К.В. О состоянии автоматизированной базы данных терминов и определений в области ОП // Компетентность. – 2014.– № 7(118). – С. 10-14.
3. Горбач Т.А., Грибова В.В., Окунь Д.Б., Петряева М.В., Шалфеева Е.А., Шахгельдян К.И. База терминов нейрохирургии для интеллектуальной обработки биомедицинских данных // Сборник материалов XIII международной научной конференции: «Системный анализ в медицине». – Благовещенск, 2019. – С. 82-85.
4. Кузнецов И.О. Автоматическое извлечение двусловных терминов по тематике «Нанотехнологии в медицине» на основе корпусных данных // Научно-техническая информация. Сер. 2. – 2013. – № 5. – С. 25-33.
5. Becerro F. B. Phraseological variations in medical-pharmaceutical terminology and its applications for English and German into Spanish translations // SciMedicine Journal. – 2020. – № 2(1). – P.22-29. DOI: 10.28991/SciMedJ-2020-0201-4.
6. Simon N. I., Kešelj V. August. Automatic term extraction in technical domain using part-of-speech and common-word features // Proceedings of the ACM Symposium on Document Engineering. – 2018. – P. 1-4. DOI:10.1145/3209280.3229100.
7. Клышинский Э.С., Кочеткова Н.А., Карпик О.В. Метод выделения коллокаций с использованием степенного показателя в распределении Ципфа // Новые информационные технологии в автоматизированных системах. – 2018. – № 21. – С. 220-225.
8. Кочеткова Н.А. Метод извлечения технических терминов с использованием усовершенствованной меры странности // Научно-техническая информация. Сер. 2.– 2015. –№ 5. – С. 25-32; Kochetkova N.A. A Method for Extracting Technical Terms Using the Modified Weirdness Measure // Automatic Documentation and Mathematical Linguistics. – 2015 – Vol. 49, № 3. – P. 89-95.
9. Захаров В.П., Хохлова М.В. Автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода как инструмент создания тезаурусов // Структурная и прикладная лингвистика. – 2012. – № 9. – С. 222-233.
10. Terryn A., Hoste V., Lefever E. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora // Language Resources and Evaluation. – 2020. – Vol. 54, № 2. – P. 385-418. DOI:10.1007/s10579-019-09453-9.
11. Лейчик В.М. Оптимальная длинна и структура термина // Вопросы языкознания. – 1981. – № 2. – С. 63-73.
12. Лейчик В.М. Исходные понятия, основные положения, определения современного терминоведения и терминографии // Вестник Харьковского политехнического университета. – 1994. – № 1. – С. 147-180.
13. Гринев-Гриневич С.В., Сорокина Э.А. Опыт описания формальной структуры термина (на материале английской терминологии лексикологии // Вестник Московского государственного областного университета. Серия: Лингвистика. – 2020. – № 5. – С. 74-85. DOI: 10.18384/2310-712X-2020-5-74-85.
14. Бутенко Ю.И. Строганов Ю.В., Сапожков А.М. Метод извлечения русскоязычных многокомпонентных терминов в корпусе научно-технических текстов // Прикладная інформатика. – 2021. – № 6. – С. 21-27. DOI: 10.37791/2687-0649-2021-16-6-21-27.
15. Бутенко Ю.И., Лукьянова Г.О. Особенности разметки научно-технических текстов в аспекте создания специализированного корпуса // Филологические науки. Научные доклады высшей школы. – 2022. – № 1. – С. 14-20. DOI 10.20339/PhS.1-22.014.
16. Бизюкова Н.Ю., Тарасова О.А., Рудик А.В., Филимонов Д.А., Поройков В.В. Автоматическое распознавание названий химических соединений в текстах научных публикаций // Научно-техническая информация. Сер. 2.– 2020. –

№ 11. – С. 36-46. DOI: 10.36535/0548-0027-2020-11-5; Biziukova N.Y., Tarasova O.A., Rudik A.V., Filimonov D.A., Poroikov V.V. Automatic Recognition of Chemical Entity Mentions in Texts of Scientific Publications // Automatic Documentation and Mathematical Linguistics. – 2020. – Vol. 54, № 6. – P. 306–315. DOI: 10.3103/S0005105520060023.

Материал поступил в редакцию 16.07.22.

Сведения об авторах

БУТЕНКО Юлия Ивановна – кандидат технических наук, доцент кафедры теоретической информатики и компьютерных технологий Московского государственного университета им. Н.Э. Баумана.
e-mail: iubutenko@bmstu.ru

СТРОГАНОВ Юрий Владимирович – старший преподаватель кафедры программного обеспечения ЭВМ и информационных технологий Московского государственного университета им. Н.Э. Баумана.
e-mail: stroganovyv@bmstu.ru

САПОЖКОВ Андрей Максимович – студент факультета информатики и систем управления Московского государственного университета им. Н.Э. Баумана.
e-mail: andreysapozhkov535@gmail.com