

DOI 10.37386/2305-4077-2024-2-158-168

Е. Г. Кузина¹*Национальный исследовательский университет
«Высшая школа экономики» (Санкт-Петербург)***Т. А. Литвинова²***Национальный исследовательский университет
«Высшая школа экономики» (Санкт-Петербург)*

NLP – ЭТО НЕ НЛО: СОСТАВЛЕНИЕ СЛОВАРЯ-ГЛОССАРИЯ ТЕРМИНОВ В СФЕРЕ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Целью настоящей статьи является описание теоретических и практических аспектов проекта по созданию англо-русского словаря-гlossария терминов, относящихся к сфере обработки естественного языка (Natural Language Processing (NLP)). Теоретические аспекты включают краткую характеристику NLP как современной области компьютерной лингвистики, а также описание существующих способов перевода терминологической лексики, связанной с данной сферой. Практическая часть исследования включает выбор источника для составления словаря, обозначение потенциальных проблем в ходе данного процесса и предложение возможных способов их решения.

Ключевые слова: прикладная лингвистика, обработка естественного языка, машинный перевод, терминологический словарь, перевод терминов, словарь-тезаурус

E. G. Kuzina*HSE University (Saint-Petersburg)***T. A. Litvinova***HSE University (Saint-Petersburg)*

NLP IS NOT AN UFO: COMPILING A GLOSSARY OF TERMS IN THE NLP SPHERE

The aim of the current article is presenting theoretical and practical aspects of the project on compiling an English-Russian glossary on NLP terms. The theoretical aspects are the following: a brief characteristic of NLP as a modern sphere of computer linguistics and a description of actual ways of translating the terminological vocabulary related to this sphere. The practical research aspect includes choosing a source for the dictionary compiling, stating potential problems during this process, and proposing possible solutions for them.

Key words: applied linguistics, natural language processing, machine translation, terminological dictionary, terms translation, thesaurus dictionary

¹ Елена Геннадьевна Кузина – магистрант программы «Языковые технологии в бизнесе и образовании» направления «Фундаментальная и прикладная лингвистика» Национального исследовательского университета «Высшая школа экономики» (Санкт-Петербург), egkuzina@edu.hse.ru.

² Татьяна Александровна Литвинова – доктор филологических наук, доцент, профессор Национального исследовательского университета «Высшая школа экономики» (Санкт-Петербург), tlitvinjva@hse.ru.

Введение

Данная статья посвящена проблеме унификации существующих в отечественной прикладной лингвистике переводов англоязычных терминов, относящихся к сфере обработки естественного языка (Natural Language Processing (NLP)). Основное внимание уделено практической стороне поставленного вопроса, созданию электронного двуязычного словаря – глоссария.

Актуальность этого проекта обусловлена следующими взаимосвязанными факторами. В силу того, что обработка естественного языка как направление компьютерной лингвистики появилось относительно недавно, значительная часть английской терминологии данной научной области не имеет общепризнанных эквивалентов на русском языке. В свою очередь, подобная неупорядоченность системы узкоспециализированной лексики приводит к трудностям понимания при работе с текстами, относящимися к рассматриваемой научной области. Особенно отчётливо данный недостаток отечественной терминологической системы NLP проявляется при применении систем автоматического перевода текстов.

Проект создания глоссария терминов NLP имеет важное практическое применение. Его потенциальной целевой аудиторией являются начинающие специалисты, работающие в области обработки естественного языка, а также практикующие разработчики систем машинного перевода. Начинающим специалистам в сфере NLP созданный электронный словарь будет полезен для ознакомления с базовой терминологией изучаемой области компьютерной лингвистики, а специалисты-практики смогут использовать этот инструмент для создания и обучения модели для автоматизированного перевода текстов указанной тематики.

В тексте статьи рассмотрены следующие аспекты обозначенной проблемы:

1. общая характеристика NLP как раздела прикладной лингвистики;
2. краткая характеристика машинного перевода как современного направления NLP;
3. анализ существующих способов перевода англоязычной терминологии NLP;
4. выбор оптимальных источников для составления глоссария;
5. описание потенциальных проблем при составлении словаря и способов их решения;
6. выбор оптимального способа визуализации продукта-результата проекта.

NLP как раздел прикладной лингвистики

Активное развитие обработки текстов на естественном языке как направления прикладной лингвистики связано с появлением компьютеров в 1940-х гг., но в течение продолжительного времени оно не воспринималось как самостоятельная дисциплина. Обособление данной отрасли во многом стало возможным благодаря контактам между сообществом компьютерных лингвистов и сообществом специалистов в сфере машинного обучения с начала 2000-х гг. Кроме этого, на становление NLP как самостоятельной области прикладной лингвистики в указанный период повлияли рост объёма доступных для работы язы-

ковых данных и тенденция к использованию компьютерных программ с целью ускорения процесса обработки подобного материала. В свою очередь, возможность применения компьютеров для работы с текстами на естественном языке была обусловлена увеличением вычислительных мощностей устройств. Наконец, необходимость формирования NLP как отдельного направления связано с трансформацией мобильных технологий, появлением т. н. речевых интерфейсов (устройств, управление которыми выполняется с помощью устных команд) [Вагин, 2019, с. 216–217].

Спектр задач обработки естественного языка включает задачи компьютерного анализа языка и синтеза речи [Близнюк, 2017, с. 15]. Оба направления, особенно синтез речи, тесно связаны с проблемой связи между языком и человеческим сознанием, а также исследованием особенностей человеческого интеллекта. Кроме этого, как было отмечено выше, развитие NLP в значительной степени обусловлено интеграцией методов, применяемых в компьютерной лингвистике, с методами машинного обучения. Поэтому значительное число современных исследований и разработок в области NLP связано с искусственным интеллектом [Вагин, 2019, с. 217].

Машинный перевод как современное направление NLP

Исторически системы автоматического (машинного) перевода являются одним из первых применений методов обработки естественного языка [Большакова, 2017, с. 9; Ерышов, 2023, с. 65]. При этом наиболее распространены системы статистического перевода, принцип работы которых состоит в поиске наиболее подходящего эквивалента предложений на основе анализа совокупности текстов на 2 языках. Необходимо отметить, что системы подобного типа полностью способны к самостоятельному обучению, но в то же время малоэффективны при работе с профессиональными узкоспециализированными текстами. Данная проблема связана с тем, что в подобных текстах некоторые слова употребляются в уникальном значении, не свойственном им при использовании в нейтральном стиле [Кузьмин, 2021, с. 45].

Недавно появившийся метод нейросетевого перевода также имеет существенные достоинства. Обучение переводческих систем, созданных на основе искусственного интеллекта, происходит на материале конкретных предложений, в которых употребляется то или иное слово. Подобный способ обучения позволяет системе научиться распознавать варианты перевода слов, характерные для контекстов различной тематики. С этим тесно связана способность подобных моделей к самостоятельному улучшению качества своей работы [Там же, с. 45]. Однако даже при переводе с использованием систем нейросетевого перевода в отдельных случаях возможны смысловые неточности. В частности, они могут возникать при переводе текстов, лексика из которых редко встречалась в материале, используемом при обучении модели [Дюжева, 2021, с. 25].

Таким образом, каждый из описанных методов машинного перевода характеризуется определёнными преимуществами и недостатками. Поэтому в настоящее время чаще всего используются модели, сочетающие оба рассмотренных

метода [Кузьмин, 2021, с. 46]. Однако использование машинного перевода является неэффективным средством для решения проблемы перевода англоязычной терминологии сферы NLP на русский язык. В качестве подтверждения сказанного приведён пример перевода фрагмента специализированного текста указанной тематики с английского языка на русский, выполненный с помощью онлайн-системы перевода с Яндекс. Жирным шрифтом выделены англоязычные термины, перевод которых сопровождается ошибками или неточностями, а также соответствующие им русские слова и словосочетания.

Consequently, **IE** methods allow for mining of the actual information present within the text rather than the limited set of **tags** associated with the documents. The **IE** process makes the number of different relevant **entities** and **relationships** on which the text mining is performed unbounded – typically thousands or even millions, which would be far beyond the number of **tags** any **automated categorization system** could handle. Thus, preprocessing techniques involving **IE** tend to create more rich and flexible representation models for documents in text mining systems. **IE** can be seen as a limited form of “complete text comprehension.” No attempt is made to understand the document at hand fully. Instead, one defines a priori the types of semantic information to be extracted from the document. **IE** represents documents as sets of **entities** and **frames** that are another way of formally describing the **relationships** between the **entities**. The set of all possible **entities** and **frames** is usually open and very big compared with the set of **categorization** keywords. It cannot be created manually. Instead, the **features** are extracted directly from the text. The hierarchy relation between the **entities** and **frames** is usually a simple tree. The root has several children – the **entity** types (e.g., “Company,” “Person,” “Gene”, etc.) under which the actual **entities** are automatically added as they are being discovered. The **frames** constitute structured objects, and so they cannot be directly used as **features** for text mining. Instead, the **frame** attributes and its label are used for features. The **frame** itself, however, may bypass the regular text mining operations and may be fed directly to the querying and visualization components.

[Следовательно, методы **IE** позволяют извлекать фактическую информацию, присутствующую в тексте, а не ограниченный набор **тегов**, связанных с документами. Процесс **IE** делает количество различных релевантных **объектов** и **связей**, для которых выполняется интеллектуальный анализ текста, неограниченным – обычно это тысячи или даже миллионы **тегов**, что намного превышает количество **тегов**, с которыми может справиться любая **автоматизированная система категоризации**. Таким образом, методы предварительной обработки с использованием **IE**, как правило, позволяют создавать более богатые и гибкие модели представления документов в системах интеллектуального анализа текста. **IE** можно рассматривать как ограниченную форму “полного понимания текста”. Не предпринимается попыток полностью понять имеющийся документ. Вместо этого априори определяются типы семантической информации, которую необходимо извлечь из документа. **IE** представляет документы в виде наборов **сущностей** и **фреймов**, которые являются еще одним способом формального описания **отношений** между **сущностями**. Набор всех возможных **сущностей** и **фреймов** обычно является открытым и очень большим по сравнению с набором ключевых слов для

категоризации. Его нельзя создать вручную. Вместо этого **элементы** извлекаются непосредственно из текста. Иерархическая связь между **сущностями** и **фреймами** обычно представляет собой простое дерево. Корень имеет несколько дочерних элементов – типов **сущностей** (например, “Компания”, “Человек”, “Ген” и т.д.), в которые фактические **сущности** автоматически добавляются по мере их обнаружения. **Фреймы** представляют собой структурированные объекты, и поэтому их нельзя напрямую использовать в качестве **объектов** для интеллектуального анализа текста. Вместо этого для **объектов** используются атрибуты **фрейма** и его метка. Однако сам **фрейм** может обходиться без обычных операций интеллектуального анализа текста и может передаваться непосредственно в компоненты запроса и визуализации].

На примере этого фрагмента можно выделить следующие случаи несоответствий при переводе англоязычной терминологии в области NLP.

Система машинного перевода не распознаёт аббревиатуру *IE*, заменяющую собой терминологическое словосочетание *information extraction* (извлечение информации).

В рамках одного и того же фрагмента существительное *entities* одновременно переводится 2 словами, *объекты* и *сущности*, причём вариант перевода *сущности* используется значительно чаще. Аналогичная ситуация наблюдается при переводе слова *relationships*. В рассматриваемом фрагменте происходит смешение его переводов *связи* и *отношения* с преобладанием последнего варианта. В обоих случаях нарушено важное требование к единству перевода терминологии.

Термины *tags* и *frames* переведены на русский язык путём транскрибирования английских слов. Это не нарушает вышеупомянутого требования единства перевода терминов, однако подобные переводы могут привести к сложностям при понимании текста у читателей, не обладающих достаточными знаниями по теме NLP. С другой стороны, использование развёрнутых словосочетаний, разъясняющих значение терминов *tags* и *frames*, внутри текста на русском языке затрудняет его восприятие в ещё большей степени. Следовательно, целесообразно вынести развёрнутое пояснение подобных терминов в словарь-гlossарий.

Переводы термина *categorization* и терминологического словосочетания *automated categorization system* представляют собой пословный перевод соответствующих английских слов. Подобный перевод сохранён на протяжении всего фрагмента текста, но с целью сохранения естественного звучания русскоязычных предложений предпочтительнее использовать перевод с помощью описательного словосочетания. Слово *categorization* более уместно перевести как *выделение категорий*, а словосочетание *automated categorization system* – как *систему автоматического выделения категорий*.

Наконец, слово *feature* переведено 2 различными словами: *элементы* и *объекты*, что является нарушением однородности при переводе терминологической лексики. Кроме этого, в обоих случаях при переводе пропущен важный компонент значения этого термина – «объект с характерными опознавательными признаками».

Способы перевода англоязычной терминологии NLP

Как было отмечено выше, направление прикладной лингвистики «обработка естественного языка» появилось относительно недавно, и именно этот факт обуславливает несформированность единой системы русскоязычных эквивалентов терминов рассматриваемой предметной области. Кроме этого, значительную часть терминологии данной отрасли составляют т. н. термины-фраземы, или многокомпонентные описательные термины. Научная лексика этого типа представляет значительные трудности в переводе из-за того, что подобные сочетания слов обозначают целостное понятие и значение всего составного термина не совпадает со значениями его изолированных элементов [Рябцева, 2023, с. 218]. Поэтому при переводе составных терминов обработки естественного языка применяются следующие приёмы [Там же, с. 220–222]:

- Импликация (пропуск избыточных с точки зрения русского языка элементов исходного описательного термина)

computer-assisted translation – автоматизированный перевод (прилагательное *автоматизированный* уже содержит элемент значения «выполняемый с помощью компьютерных технологий», поэтому соответствующая часть английского словосочетания не переводится дословно);

voice user interface – голосовой интерфейс (слово *интерфейс* означает «устройство, обеспечивающее взаимодействие пользователя с приложением или компьютером», поэтому перевод слова *user* опускается).

- Экспликация (перевод с добавлением необходимых слов на русском языке для передачи элементов значения, передаваемых в английском языке с помощью одного слова)

image captioning – генерация описания изображения (значение действия, выраженное в английском языке отглагольным существительным, переводится на русский язык с помощью словосочетания «генерация описания»);

supervised learning – обучение модели с учителем (прилагательное *supervised* имеет значение «выполняемый под контролем кого-либо», и при переводе этого словосочетания добавлено эксплицитное обозначение того, с чьей помощью осуществляется обучение модели).

- Нормализация (преобразование синтаксической структуры описательного словосочетания в соответствии с нормами русского языка)

sentiment analysis – анализ тональности текста (вместо первичного значения слова *sentiment*, связанного с человеческими эмоциями, в переводе использован узкоспециализированный оттенок этого значения, который относится к эмоциональной окраске текста);

word segmentation – разделение текста на слова (при переводе добавлено отсутствующее в составе англоязычного словосочетания уточнение о том, что данный процесс применяется ко всему тексту. Дословный перевод данного термина привёл бы к его некорректному пониманию: словосочетание «сегментация слов» можно понять как «разделение слов на отдельные буквенные символы»).

Источники для построения словаря-гlossария, возможные проблемы при его составлении и способы их решения

Как было отмечено во введении, целью проекта является систематизация принятых в русской лингвистической традиции терминов из сферы NLP в виде двуязычного гlossария. В англоязычной научной литературе описана возможность применения онтологий для решения подобных задач, однако на данный момент указанный метод составления предметного словаря применяется только в научных областях, не связанных с компьютерной лингвистикой [Tudorache, 2020, с. 125–126]. Кроме того, в научных работах отмечено, что применение онтологий возможно только при работе со структурированными данными [Там же, с. 127]. Таким образом, применение существующих онтологий (*Schema.org*, *Wikidata.org* и др.) для создания терминологического словаря в сфере NLP нельзя считать целесообразным главным образом из-за того, что тексты на русском и английском языках, содержащие необходимые термины, не являются структурированными данными.

Что касается выбора источника текстов, содержащих терминологию NLP на 2 языках, наиболее целесообразным представляется использование базы данных переводческой памяти. Названная база данных содержит часто встречавшиеся переводы отдельных фрагментов текстов данной тематики с использованием терминологических слов и словосочетаний. Использование данного источника представляется эффективным в силу его структурированности. Переводческая память представляет собой упорядоченный в виде двуязычных соответствий набор контекстов, что позволяет существенно ускорить поиск слов-эквивалентов.

Необходимо отметить, что при составлении списка соответствий терминов возможно возникновение следующих проблем.

Один и тот же термин может иметь несколько соответствий, и в этом случае логично выбрать в качестве перевода тот вариант, который встречается с большей частотой (см. в качестве примера рассмотренный в начале статьи пример перевода слов *entities* и *relationships*). С другой стороны, не следует исключать полное отсутствие русскоязычных эквивалентов отдельных терминов. В этом случае необходимо самостоятельно сформулировать их перевод с использованием приёмов, описанных в предыдущей рубрике.

При переводе многокомпонентных терминов посредством экспликации или нормализации возможно образование словосочетаний с ещё более сложной синтаксической структурой, чем исходные. В этом случае возможна замена перевода кратким комментарием, раскрывающим суть переводимого понятия.

Некоторые термины могут быть переведены на русский путём транскрипции (передачи английского произношения слов с помощью русской графической системы). В подобных случаях также целесообразно использовать перевод в виде краткого толкования сути транскрибированного понятия. Например, рассмотренные в начале статьи слова *tags* и *frames* можно перевести с помощью ком-

ментариев «ключевые слова, позволяющие однозначно определить смысловую категорию, к которой относится та или иная информация в тексте» и «обобщённое описание объектов с помощью набора ключевых признаков» соответственно.

Если соответствия терминов представляют собой калькирование (пословный перевод) английских словосочетаний, требуется редактирование подобных переводов путём нормализации (см. в качестве примера перевод словосочетания *automated categorization system* как *системы автоматического выделения категорий*).

Представление результатов проекта и усовершенствование составленного словаря

Оптимальным средством для визуализации составленного двуязычного глоссария терминов является приложение *Glossword*, позволяющее не только компилировать словари или глоссарии, но и публиковать их в открытом доступе. В качестве возможного направления усовершенствования итогового продукта логично рассматривать упорядочивание основных понятий обработки естественного языка в формате словаря-тезауруса. Как отмечают современные исследователи, представление терминологической системы какой-либо научной области в виде словаря такого типа позволяет более детально изучить её как лингвистическую знаковую систему, а также понять механизмы формирования терминологии в изучаемой сфере. Кроме этого, тезаурусная модель терминов имеет важное дидактическое значение, потому что она отражает отношения между понятиями в конкретной области [Журавлёв, 2021, с. 13–14]. Наконец, Ю. Н. Караулов указывает на тесную связь тезаурусных терминологических словарей с одной из задач сферы NLP, рассматриваемой в настоящей статье, с задачей автоматического извлечения информации из текстов. Словарь-тезаурус содержит лексику, отражающую актуальное состояние той или иной научной области, что позволяет применять данный инструмент для автоматического извлечения информации из текстов соответствующей тематической направленности, а также для их индексирования и реферирования [Караулов, 1981, с. 222].

Ниже приведена схема подобного словаря-тезауруса в виде многоуровневого нумерованного списка, представляющая собой группировку терминов по направлениям NLP.

I. Natural Language Processing (NLP) Обработка естественного языка

1. Machine Learning Машинное обучение

1.1. Supervised Learning Обучение с учителем

1.2. Unsupervised Learning Обучение без учителя

II. Speech Recognition Распознавание речи

1. Direct Voice Input (DVI) Голосовой ввод

2. Voice User Interface (VUI) Голосовой интерфейс

3. Word Processor Текстовый процессор

II. Morphological Analysis Морфологический анализ текста

1. **Tokenization (Word Segmentation)** Разделение текста на слова

1.1. **Lemmatization** Приведение слов к начальной форме

1.2. **Part-of-speech Tagging** Разметка частей речи в тексте

III. **Lexical Semantics** Описание значения слов в контексте

1. **Information Extraction (IE)** Извлечение информации

1.1. **Named Entity Recognition (NER)** Распознавание именованных сущностей

1.1.1. **Categorization** Выделение категорий

1.1.2. **Frame** Обобщённое описание объектов с помощью набора ключевых признаков

1.1.3. **Tag** Ключевое слово, позволяющее однозначно определить категорию понятия

1.2. **Terminology Extraction** Извлечение терминов

IV. **Machine Translation (MT)** Машинный перевод

1. **Neural Machine Translation (NMT)** Нейронный машинный перевод

2. **Statistical Machine Translation (SMT)** Статистический машинный перевод

Заключение

В данной статье были рассмотрены основные теоретические и практические аспекты, связанные с проблемой составления единого словаря англоязычных терминов в сфере обработки естественного языка.

Теоретический обзор литературы позволяет сформировать общее представление об NLP как современном направлении компьютерной лингвистики. Упомянутое направление связано с оптимизацией процесса обработки большого объёма текстовых данных с помощью компьютерных технологий, а также с применением искусственного интеллекта. Исторически данное направление связано прежде всего с использованием систем машинного перевода. Однако путём анализа фрагмента автоматизированного перевода специализированного текста, посвящённого обработке естественного языка, было установлено, что с помощью данного инструмента невозможно в полной мере достичь поставленной во введении цели проекта.

Практические результаты настоящей статьи включают выбор накопителя переводов в качестве оптимального источника для построения двуязычного словаря-гlossария, а также выделение возможных проблем, связанных с поиском русскоязычных эквивалентов терминов, и описание возможных способов их решения. Необходимо отметить, что проект, которому посвящена данная статья, является перспективным с точки зрения разработки и дальнейшего совершенствования. Одним из направлений его развития является группировка терминов по направлениям NLP в формате словаря-тезауруса. Примерная схема этого словаря представлена в настоящей работе.

СПИСОК ЛИТЕРАТУРЫ

1. **Близнюк, Б. О.** Современные методы обработки естественного языка / Б. О. Близнюк, Л. В. Васильева, И. Д. Стрельников, Д. С. Ткачук // Вестник Харьковского национального университета имени В. Н. Каразина. Серия: Математическое моделирование. Информационные технологии. Автоматизированные системы управления. – 2017. – № 36. – С. 14–26.
2. **Большакова, Е. И.** Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие. / Е. И. Большакова, К. В. Воронцова, Н. Э. Ефремова, Э. С. Клышевский, Н. В. Лукашевич, А. С. Сапин. – Москва.: Изд-во НИУ ВШЭ, 2017. – 269 с.
3. **Вагин, Н. А.** Общее понятие об обработке естественного языка / Н. А. Вагин // Фундаментальные и прикладные научные исследования: актуальные вопросы, достижения и инновации: сборник статей XXV Международной научно-практической конференции: в 2 ч., Пенза, 30 июля 2019 года. Часть 1. – Пенза: Наука и Просвещение, 2019. – С. 215–218.
4. **Дюжева, А. Н.** Информационные технологии в лингвистике / А. Н. Дюжева // Языковые и культурные реалии современного мира: Сборник материалов VI Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых, Пенза, 12–22 апреля 2021 года. – Пенза: Пензенский государственный технологический университет, 2021. – С. 19–26.
5. **Ерышов, А. А.** Достижения в области обработки естественного языка / А. А. Ерышов // Наука и технологии сегодня: сборник статей Международной научно-практической конференции, Петрозаводск, 23 августа 2023 года. – Петрозаводск: Международный центр научного партнерства «Новая Наука», 2023. – С. 63–67.
6. **Журавлёв, А. П.** Актуальность применения тезаурусного подхода для моделирования лексико-семантической структуры терминополья / А. П. Журавлёв // Известия Самарского научного центра Российской академии наук. Социальные, гуманитарные, медико-биологические науки. – 2021. – Т. 23. – № 78. – С. 13–16.
7. **Караулов, Ю. Н.** Лингвистическое конструирование и тезаурус русского языка / Ю. Н. Караулов – Москва: Наука, 1981. – 367 с.
8. **Кузьмин, О. И.** Перспективы современных систем машинного перевода / О. И. Кузьмин // Вестник Нижегородского государственного лингвистического университета им. Н. А. Добролюбова. – 2021. – № 53. – С. 41–52.
9. **Рябцева, Н. К.** Межъязыковые соответствия и актуальные проблемы терминографии / Н. К. Рябцева // Научный диалог. – 2023. – Т. 12. – № 8. – С. 211–227.
10. **Tudorache, T.** Ontology engineering: Current state, challenges, and future directions / T. Tudorache // Semantic Web. – 2020. – P. 125–138.

REFERENCES

1. **Bliznjuk, B. O.** Sovremennye metody obrabotki estestvennogo jazyka / B. O. Bliznjuk, L. V. Vasil'eva, I. D. Strel'nikov, D. S. Tkachuk // Vestnik Kharkovskogo natsionalnogo universiteta imeni V. N. Karazina. Serija: Matematicheskoje modelirovanije. Informacionnyje tekhnologii. Avtomatizirovannyje sistemy upravlenija. – 2017. – № 36. – S. 14–26.
2. **Bolshakova, E. I.** Avtomaticheskaya obrabotka tekstov na estestvennom jazyke i analiz dannykh: ucheb. Posobie / E. I. Bol'shakova, K. V. Voroncova, N. E. Efremova, E. S. Klyshevskij, N. V. Lukashevich, A. S. Sapin. – Moskva: Izd-vo NIU VSHE, 2017. – 2017. – 269 s.
3. **Djuzheva, A. N.** Informacionnyje tekhnologii v lingvistike / A. N. Djuzheva // Jazykovyje i kulturnyje realii sovremennogo mira: Sbornik materialov VI Vserossijskoj nauchno-prakticheskoy konferentsii studentov, aspirantov i molodykh uchjonykh, Penza, 12 – 22 aprelja 2021 goda. – Penza: Penzskij gosudarstvennyj tekhnologicheskij universitet, 2021. – S. 19–26.
4. **Jeryshov, A. A.** Dostizhenija v oblasti obrabotki estestvennogo jazyka / A. A. Jeryshov // Nauka i tekhnologii segodnja: sbornik statej Mezhdunarodnoj nauchno-prakticheskoy konferentsii, Petrozavodsk, 23 avgusta 2023 goda. – Petrozavodsk: Mezhdunarodnyj tsentr nauchnogo partnjorstva «Novaja nauka», 2023. – S. 63–67.
5. **Karaulov, J. N.** Lingvisticheskoje konstruirovanije i tezaurus russkogo jazyka / J. N. Karaulov. – Moskva: Nauka, 1981. – 367 s.
6. **Kuzmin, O. I.** Perspektivy sovremennykh system machinnogo perevoda / O. I. Kuzmin // Vestnik Nizhegorodskogo gosudarstvennogo lingvisticheskogo universiteta im. N. A. Dobroljubova – 2021. – № 53. – S. 41–52.
7. **Rjabtseva, N. K.** Mezhjazykovye sootvetstvija i aktualnye problemy terminografii / N. K. Rjabtseva // Nauchnyj dialog. – 2023. – № 8. – S. 211–227.
8. **Tudorache, T.** Ontology engineering: Current state, challenges, and future directions / T. Tudorache // Semantic Web. – 2020. – P. 125–138.
9. **Vagin, N. A.** Obshee ponyatie ob obrabotke estestvennogo yazyka / N. A. Vagin // Fundamental'nye i prikladnye nauchnye issledovaniya: aktual'nye voprosy, dostizheniya i innovacii: sbornik statej XXV Mezhdunarodnoj nauchno-prakticheskoy konferencii: v 2 ch., Penza, 30 iyulya 2019 goda. Chast' 1. – Penza: Nauka i Prosveshchenie, 2019. – S. 215–218.
10. **Zhuravlev, A. P.** Aktualnost primeneniya tezaurusnogo podkhoda dlja modelirovanija leksiko-semanticheskoy struktury terminopolja / A. P. Zhuravlev // Izvestija Samarskogo nauchnogo tsentra Rossijskoj akademii nauk. Sotsialnye, gumanitarnye, medico-biologicheskie nauki – 2021. – T. 23 – № 78. – S. 13–16.