УДК 004.89, 004.912

А.М. Наместников, И.В. Арзамасцева

ОНТОЛОГИЧЕСКИЙ ПОДХОД К СЕНТИМЕНТ-АНАЛИЗУ ПРОГРАММНЫХ СИСТЕМ[1]

Наместников Алексей Михайлович, *доктор технических наук, доцент, окончил радиотехнический факультет Ульяновского государственного технического университета. Профессор кафедры «Информационные системы» УлГТУ. Имеет около 100 работ в области автоматизированного проектирования и интеллектуальных систем. [e-mail: nam@ulstu.ru].*

Арзамасцева Иветта Вячеславовна, *кандидат технических наук, окончила филологический факультет Саратовского государственного университета. Доцент кафедры «Прикладная лингвистика» УлГТУ. Имеет статьи, изобретения в области статистических исследований и математического моделирования в лингвистике. [e-mail: lingua@ulstu.ru].*

Аннотация

В работе предложен оригинальный подход к решению задачи определения тональности отзывов на программные системы (сентимент-анализа), которые могут быть представлены на нескольких языках (русский, английский и немецкий). Особенностью подхода является реализация процедуры выделения шаблонов, которые представляют собой устойчивые сочетания одного, двух или более слов, связанных по смыслу и грамматически. В статье приводятся примеры русскоязычных шаблонов для предметной области разработки компьютерных игр. Определяемые лексико-семантические шаблоны являются составной частью онтологии сентимент-анализа программных систем, формальное описание которой приведено в данной работе.

Ключевые слова: сентимент-анализ, программная система, лексико-семантический шаблон, онтология, байесовский классификатор.

doi: 10.35752/1991-2927-2021-2-64-34-39

AN ONTOLOGY APPROACH TO THE SENTIMENT ANALYSIS OF SOFTWARE SYSTEMS

Aleksei Mikhailovich Namestnikov, *Doctor of Sciences in Engineering, Associate Professor; graduated from the Radioengineering Faculty of Ulyanovsk State Technical University; Professor of the Department of Information Systems of UlSTU; an author of more than 100 articles in the field of computer-aided design and intelligent systems. e-mail: nam@ulstu.ru.*

Ivetta Viacheslavovna Arzamastseva, *Candidate of Sciences in Engineering; graduated from the Faculty of Language and Literature of Saratov State University; Associate Professor at the Department of Applied Linguistics of UlSTU; an author of articles and inventions in the field of statistical investigations and mathematical modeling in linguistics. e-mail: lingua@ulstu.ru.*

1 Работа выполнена при финансовой поддержке РФФИ, грант 19-47-730003.

Abstract
The article proposes an unorthodox approach to opinion mining for software systems (sentiment analysis) that can be in Russian, English or German languages. The main feature of this approach is an extraction of patterns represented by collocations consisting of one, two or more words and connected by common meaning or grammar. The article gives some examples of patterns in Russian in the area of computer games development. The determining lexical and semantic patterns are included in the ontology of the software system sentiment analysis described formally in the article.

Keywords: sentiment analysis, software system, lexical and semantic pattern, ontology, Bayes classificator.

Введение

Интернет-пространство является на сегодняшний день одним из основных коммуникативных пространств, поэтому отзывы о различных интернет-продуктах становятся все более важными, т. к. они предоставляют пользователям возможность узнать о достоинствах и недостатках определенного приложения задолго до его скачивания.

При анализе мнений в отзывах для поиска субъективности, содержащейся в выражениях, все чаще используется сентимент-анализ, поскольку он позволяет узнать, что думают пользователи о тех или иных приложениях, как описывают их преимущества и недоработки.

По мнению Сарбасовой А.Н., сентимент-анализ (англ. sentiment analysis) представляет собой «выявление тональности комментария при помощи методов NLP (обработка естественного языка), статистики, машинного обучения. Иногда упоминается как opinion mining, хотя в этом случае делается акцент на извлечение нужного отрывка текста» [1].

Термин Sentiment означает базовую или, зачастую, скрытую положительную или отрицательную эмоцию, которая подразумевает под собой мнение [2].

Немецкий ученый Мелани Зигель определяет тональность текста как «набор методов для определения эмоциональной окраски лексики текстов, эмоций автора по отношению к объекту и других свойств. Решение этой задачи компьютерной лингвистики позволит понимать текстовую информацию и упростит дальнейшее использование данных, полученных в результате ее систематизации и обработки».

Основной задачей анализа тональности является поиск мнений (лексических тональностей в тексте), извлечение шаблонов (лексических сентиментов, слов-сентиментов) и определение их свойств с целью дальнейшей классификации документов этого корпуса при помощи найденных слов-сентиментов. Лексема же понимается как экземпляр последовательности символов в определенном документе, объединенных в семантическую единицу для обработки. Данная задача также называется задачей классификации полярности документов, которая определяет, является ли имеющееся в документе мнение позитивным или негативным (в самом простом случае).

По мнению Бинг Лиу, существует три уровня, на которых возможно проведение сентимент-анализа [2]:

1) на уровне документа: на этом уровне весь документ целиком классифицируется как положительный, негативный или нейтральный (или в соответствии с выбранной шкалой полярностей);

2) на уровне предложения: каждое предложение текста классифицируется как положительное, негативное или нейтральное (или в соответствии с выбранной шкалой полярностей). Решение задачи на данном уровне характерно для сравнительных предложений;

3) на уровне характеристик: нахождение всех мнений, высказанных об объекте или его характеристиках; определение тональностей мнений. По сути, задача, тождественная задаче извлечения мнений.

На основе исследований различных иностранных ученых-лингвистов, М.В. Клековкина и Е.В. Котельников выделили следующие подходы для автоматизированного анализа тональности:

1) на основе правил с использованием шаблонов (rule-based with patterns), где подход заключается в генерации правил, на основе которых будет определяться тональность текста. Для этого текст разбивается на слова или последовательности слов, а затем полученные данные используются для выделения часто использующихся шаблонов, которым присваивается позитивная или отрицательная оценка;

2) машинное обучение без учителя (unsupervised learning). Данный подход основан на идее, что наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции. Выделив эти термины и определив их тональность, можно сделать вывод о тональности всего текста целиком;

3) машинное обучение с учителем (supervised learning). В этом подходе требуется наличие обучающей коллекции размеченных в рамках эмотивного пространства текстов, на базе которой строится статистический или вероятностный классификатор;

4) гибридный метод (hybrid method). Данный метод сочетает все или несколько подходов, рассмотренных выше, и заключается в применении классификаторов на их основе в определенной последовательности.

Смешение мнений и настроений в одном фрагменте текста представляет собой сложную исследовательскую задачу для систем текстового и веб-анализа, потому что в зависимости от цели соответствующие части должны быть целенаправленно идентифицированы и извле-

чены: приложение для извлечения информации ищет фактическую информацию в отличие от приложения, которое проводит анализ настроений, направленный на обнаружение эмоций и мнений. Косвенные и неоднозначные высказывания в мнениях не имеют явного и четкого утверждения сентимента, но только через интерпретацию и связь высказываний становится очевидным настроение мнения. Таким образом, автоматическое обнаружение косвенных высказываний в комментариях гораздо сложнее, чем определение прямых высказываний. Более того, распознавание субъективных сравнительных мнений в высказываниях также более трудоемко, нежели прямых простых мнений.

Что касается конкретно методов извлечения оценочных высказываний из мнений, то Лукашевич Н.В. подразделяет их на [3]:

1. ручной метод;
2. корпусный подход:
- на основе шаблонов и конструкций;
- на основе статистики взаимной встречаемости;
3. метод на основе словарей;
4. комбинация приведенных выше подходов.

1 Языковые шаблоны

Лексический шаблон – декларативная структура, структурный образец языковой конструкции, который отображает её лексические и поверхностные синтаксические свойства [4]. Иными словами, это описание некоторого смыслового явления в виде чёткого клише. Такие шаблоны называют также лексико-синтаксическими или лексико-семантическими.

Лексико-семантический шаблон представляет собой структурный образец целевой языковой конструкции с указанным составом и лексико-семантическими свойствами. В случае успешного сопоставления шаблона с фрагментом текста формируется лексический объект, которому приписываются формальные (позиционные) и семантические (класс и свойства) характеристики. Лексико-синтаксические шаблоны позволяют создавать новые языковые конструкции. Шаблон состоит из логической структуры и семантического описания [5].

Рабчевский Е.А. считает, что «лексико-синтаксические шаблоны представляют собой характерные выражения (словосочетания), конструкции из определенных элементов языка. Такие шаблоны позволяют построить семантическую модель, соответствующую тексту, к которому они применяются» [6].

Под лексико-семантическим шаблоном, по мнению Тимофеева П.С. и Сидоровой Е.А., понимается «структурный образец целевой языковой конструкции с указанным составом и лексико-семантическими свойствами. В случае успешного сопоставления шаблона с фрагментом текста формируется лексический объект, которому приписываются формальные (позиционные) и семантические (класс и свойства) характеристики» [7].

2 Определение лексико-семантических шаблонов

В данной работе представлены результаты отбора шаблонов из отзывов по компьютерным играм. Для извлечения лингвистических оценочных конструкций из текстов отзывов на первом этапе применялся ручной метод (на основе сплошной выборки). Затем, после выявления некоторых первичных шаблонов, применялся корпусный подход извлечения выражений из мнений на основе классификации Большаковой Е.И. и Лукашевич Н.В. Далее все шаблоны были распределены на несколько категорий и исследовалась частотность их употребления и анализ особенностей. В результате была собрана статистика частотности использования данных шаблонов.

На первом этапе объём материала исследования составил 300 отзывов по компьютерным играм, где 150 являются отрицательными, а 150 – положительными. Отзывы отбирались методом сплошной выборки на основе оценок пользователей: положительные были взяты из отзывов, имеющих оценку 5 звезд, а отрицательные – соответственно 1 звезду. Временные рамки отбора отзывов: февраль-март 2020 года.

Выделенные шаблоны представляют собой устойчивые сочетания одного, двух или более слов, связанных по смыслу и грамматически.

В процессе исследования все отзывы были разделены на 4 категории:

- Преимущества,
- Сравнения,
- Критика,
- Мнения.

На рисунке 1 представлена схема шаблонов категории «Критика».

Были выявлены следующие примеры самых частотных шаблонов с положительной оценкой (рис. 2).

Далее представлены самые частотные отрицательные шаблоны (рис. 3).

Остальные шаблоны встречались всего один раз и не планировались к включению в состав онтологии сентимент-анализа программных систем.

3 Формальная модель онтологии

Формальное представление онтологии сентимент-анализа программных систем запишем следующим образом:

$$MO = < C, T, R, F >,$$

где C – множество понятий онтологии; T – множество шаблонов, извлеченных из отзывов, состоящих из трех подмножеств:

$$T = T^{ru} \cup T^{en} \cup T^{de},$$

где каждое подмножество определяет лингвистические шаблоны, относящиеся к различным языкам (ru – русский, en – английский и de – немецкий языки);

R – множество отношений, включающее в себя следующие подмножества:

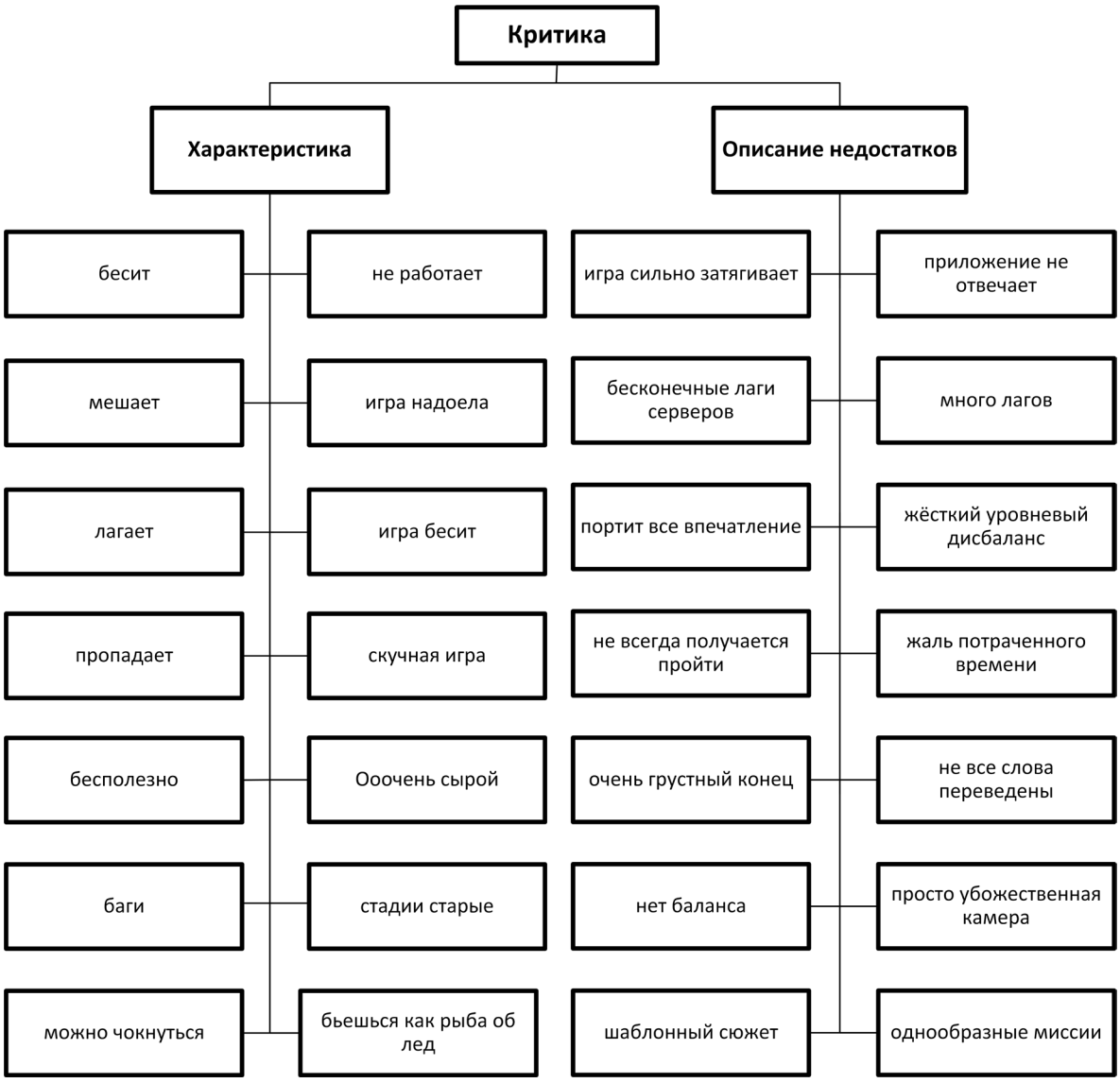

Рис. 1. Основные шаблоны категории «Критика»

$$R = R^{isA} \cup R^{isLang} \cup R^{hasALabel} \cup$$
$$\cup R^{hasAPosTemp} \cup R^{hasANegTemp},$$

где R^{isA} – отношения обобщения («is_A»);

R^{inLang} – отношения, связывающие понятия предметной области и термин, заданный на определенном языке;

$R^{hasALabel}$ – отношения, определяющие язык представления концепта;

$R^{hasAPosTemp}$, $R^{hasANegTemp}$ – подмножества отношений, определяющие положительные и отрицательные шаблоны для понятий предметной области, соответственно.

F – множество функций интерпретаций, имеющее следующий вид:

$$F = F^{Lang} \cup F^{Temp} \cup F^{Parent},$$

где

$$F^{Lang}: LC \to LabLang,\ LC \subset C,\ LabLang \subset C,$$

где LC – термин определенного языка, представляющий концепт онтологии;

$LabLang$ – множество языковых меток;

$F^{Temp}: T \to C$ – множество функций, определяющих по лингвистическому шаблону концепт предметной области;

$F^{Parent}: C \to C$ – множество функций, позволяющих осуществлять переход от дочерних концептов к родительским концептам.

Алгоритм онтологической предобработки текстов отзывов представлен следующим образом:

Шаг 1. Выполнение стемминга и приведение к формату словаря Python (ключ – лексема, значение – частота встречаемости);

Шаг 2. Приведение терминов текста отзыва к независимому от языка представлению (только для тех терминов, которые присутствуют в онтологии);

Шаг 3. Если есть термины в тексте, у которых одинаковый родительский концепт в онтологии, то

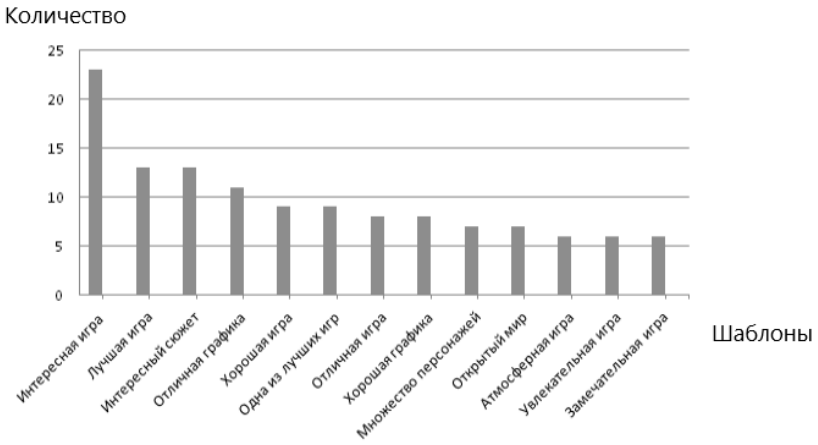

Рис. 2. Частотность употребления шаблонов с положительной оценкой

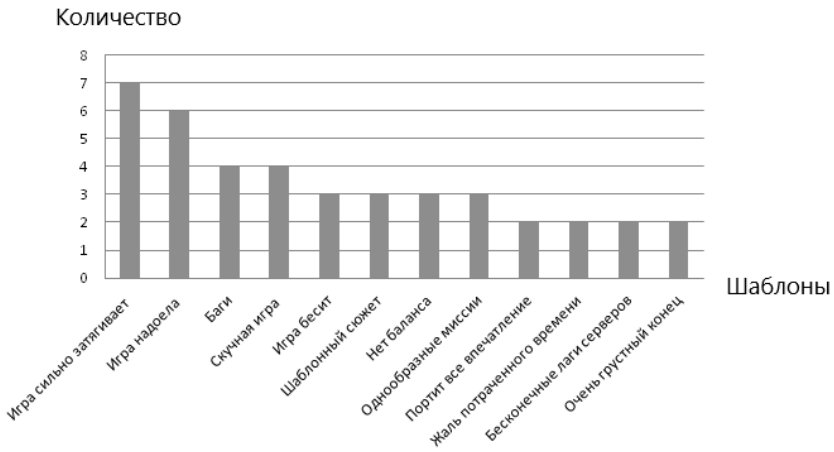

Рис. 3. Частотность употребления шаблонов с отрицательной оценкой

выполняется замена данных терминов на родительский;

Шаг 4. Если в тексте встречается шаблон, соответствующий категории, то отзыву присваивается соответствующая метка («Преимущества», «Сравнение», «Критика», «Мнения»). Категория заранее известна при выполнении классификации.

4 Реализация программной системы и проведение вычислительных экспериментов

Реализация программной системы выполнена на платформе Django 2.0 (фреймворк для построения Web-приложений на языке Python 3.7) с использованием библиотеки NLTK (библиотека для обработки естественного языка).

Реализованы классификаторы отзывов на программные продукты следующих типов:

1) Наивный байесовский классификатор (NB);

2) Деревья принятия решений (DT);

3) Логистическая регрессия (LR).

Вычислительные эксперименты выполнялись на 300 отзывах (150, относящихся у категории «Критика», и 150, к указанной категории не относящихся). Каждый тип классификаторов исследовался совместно с онтологическим модулем (+O) и без него. Результаты вычислительных экспериментов приведены в таблице.

Наиболее значимый результат отмечается для классификатора LR+O (логистическая регрессия с онтологическим модулем).

Заключение

Рассмотренные в данной статье методология, комбинация этапов инженерного подхода к извлечению шаблонов и полученные результаты подтверждают то, что сентимент-анализ является одним из наиболее важных инструментов при извлечении информации из текстов, т. к. на основе выявления оценочных мнений легко сделать вывод о том, какую коннотацию приобретает целый отзыв.

Дальнейшие перспективы данного исследования предполагают углубленное изучение лексико-семантических шаблонов для решения задач анализа тональности текста, а также других методов сентимент-анализа для более эффективного извлечения оценочных мнений из текстов естественного языка.

Таблица

№ п/п	Тип классификатора	Точность классификации	Вклад онтологии
1	NB	0,83	
2	NB+O	0,82	-
3	DT	0,91	
4	DT+O	0,93	+-
5	LR	0,9	
6	**LR+O**	**0,94**	+

СПИСОК ЛИТЕРАТУРЫ

1. Сарбасова А.Н. Исследование методов сентимент-анализа русскоязычных текстов // Молодой ученый. 2015. № 8 (88). С. 143–146. URL: https://moluch.ru/archive/88/17413 (дата обращения: 11.10.2020).

2. Liu. B. Sentiment Analysis and Subjectivity // Handbook of Natural Language Processing, N. Indurkhya & F.J. Damerau. (Eds.), 2010. С. 1–38. URL: https://www.researchgate.net/publication/228667268_Sentiment_analysis_and_subjectivity (дата обращения: 29.09.2020).

3. Лукашевич Н.В., Четверкин И.И. Построение модели для извлечения оценочной лексики в различных

предметных областях // Моделирование и анализ информационных систем. 2013. Т. 20. №2. С. 70–79. URL: http://foresight.ifmo.ru/ict/shared/files/201309/1_15.pdf (дата обращения: 18.09.2020).

4. Большакова Е.И. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии : тр. Междунар. конф. «Диалог», 2007. URL: http://www.dialog-21.ru/digests/dialog2007/materials/html/11.htm (дата обращения: 30.09.2020).

5. Большакова Е.И. Язык лексико-синтаксических шаблонов LSPL: опыт использования и пути развития // Программные системы и инструменты : тематич. сб., 2014. № 15. URL: http://www.lspl.ru/articles/Paper_19_LSPL.pdf (дата обращения: 19.08.2020).

6. Рабчевский Е.А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска // Тр. 11-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2009. Петрозаводск, 2009, С. 69–77. URL: http://rcdl.ru/doc/2009/069_077_DIIS-seminar-1-2009-1.pdf (дата обращения: 11.09.2020).

7. Тимофеев П.С., Сидорова Е.А. Лексико-семантические шаблоны как инструмент декларативного описания языковых конструкций и лингвистического анализа текста // System Informatics (Системная информатика). 2018. No. 13. С. 35–47. URL: https://system-informatics.ru/files/article/diglex-timofeevsidorova_v2.pdf (дата обращения: 07.10.2020).

REFERENCES

1. Sarbasovs A.N. Issledovanie metodov sentiment-analiza russkoiazychnykh tekstov [The Study of Sentiment Analysis Methods for Russian Language Text]. *Molodoi uchenyi* [Young Scientist], 2015, no. 8 (88), pp. 143–146. Available at: https://moluch.ru/archive/88/17413 (accessed 11.10.2020).

2. Liu. B. Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, N. Indurkhya & F.J. Damerau (Eds.), 2010, pp. 1–38. Available at: https://www.researchgate.net/publication/228667268_Sentiment_analysis_and_subjectivity (accessed 29.09.2020).

3. Lukashevich N.V., Chetviorkin I.I. Postroenie modeli dlia izvlecheniia otsenochnoi leksiki v razlichnykh predmetnykh oblastiakh [Construction of a Model for the Cross-Domain Opinion Word Extraction]. *Modelirovanie i analiz informatsionnykh sistem* [Modeling and Analysis of Information Systems], 2013, vol. 20, no. 2, pp. 70–79. Available at: http://foresight.ifmo.ru/ict/shared/files/201309/1_15.pdf (accessed 18.09.2020).

4. Bolshakova E.I. Leksiko-sintaksicheskie shablony v zadachakh avtomaticheskoi obrabotki tekstov [Lexico-Syntactic Templates in Automated Language Processing Tasks]. *Kompiuternaia lingvistika i intellektualnye tekhnologii. Tr. Mezdunar. konf. "Dialog"* [Proc. of Int. Sci. Conf. "Dialog". Computer Linguistics and Intelligent Technologies]. 2007. Available at: http://www.dialog-21.ru/digests/dialog2007/materials/html/11.htm (accessed 30.09.2020).

5. Bolshakova E.I. Iazyk leksiko-sintaksicheskikh shablonov LSPL: opyt ispolzovaniia i puti razvitiia [Language of Lexico-Syntactic LSPL Templates. Use Experience and Development Trends]. *Programnye sistemy i instrumenty. Tematich. sb.* [Proc. on Software Systems and Tools]. 2014, no. 15. Available at: http://www.lspl.ru/articles/Paper_19_LSPL.pdf (accessed 19.08.2020);

6. Rabchevskii E.A. Avtomaticheskoe postroenie ontologii na osnove leksiko-sintaksicheskikh shablonov dlia informatsionnogo poiska [Computer-Aided Ontology Building Based on Lexico-Syntactic Templates for Information Search]. *Tr. 11-i Vseros. nauch. konf. "Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii" – RCDL'2009* [Proc. of the 11th Russian Sci. Conf. on Electronic Libraries: Prospective Methods and Technologies, Electronic Collections. RCDL'2009]. Petrozavodsk, 2009, pp. 69–77. Available at: http://rcdl.ru/doc/2009/069_077_DIIS-seminar-1-2009-1.pdf (accessed 11.09.2020);

7. Timofeev P.S., Sidorova E.A. Leksiko-semanticheskie shablony kak instrument deklarativnogo opisaniia iazykovykh konstruktsii i lingvisticheskogo analiza teksta [A Lexico-Semantic Templates as a Tool for Declarative Description Language Constructs Linguistic Text Analysis]. *Sistemnaia informatika* [System Informatics], 2018, no. 13, pp. 35–47. Available at: https://system-informatics.ru/files/article/diglex-timofeevsidorova_v2.pdf (accessed 07.10.2020).