УДК 004.855.2/004.5 DOI: 10.17587/it.24.719-724

А. Б. Сорокин, канд. техн. наук, доц., e-mail: ab_sorokin@mail.ru,
А. П. Кушнарев, магистрант, e-mail: brainzeater@gmail.com,
Московский технологический университет (МИРЭА)

Морфологический анализатор текста для выявления полноты информации

Рассматриваются вопросы применения технологии автоматической частеречной разметки русскоязычных текстов, представленных в цифровом виде, в целях определения переизбытка или недостатка информации в тексте, выявления и построения концепта. Основное внимание уделяется этапу морфологического анализа как одному из наиболее сложных этапов в анализе текста в силу специфических особенностей морфологии русского языка, связанных с неоднозначностью соответствия слов той или иной части речи.

Ключевые слова: *концептуальная структура, компьютерная лингвистика, частеречная разметка, автоматическая обработка документов, обработка текстов на естественном языке, алгоритм Портера*

Введение

Развитие искусственного интеллекта и кибернетических идей привело к появлению множества научных областей, которые являются порождением часто не связанных между собой наук. Одной из таких научных областей является компьютерная лингвистика, которая позволила найти решение множества задач, с которыми прежде мог справиться только человек. Анализ текста естественного языка — одна из таких задач.

В современном мире человек вынужден существовать в бесконечном информационном потоке. В таких условиях возникает проблема зашумленности информации, определения разницы между значимой и несущественной информацией. При этом значительная часть информации представлена в виде цифровых текстовых документов. Вполне вероятно, что большинство из них обладают лишней, избыточной информацией, которая может затруднять чтение и усваивание информации из текста. Вместе с тем, текст может обладать недостаточностью для понимания информации.

Для структурирования и формализации информации необходимо выделить из нее концепт как некую логическую структуру. При этом концептуальная структура должна позволять проводить оценку полноты и адекватности информации, формировать базу знаний на программном уровне. Описание такой структуры на основе ситуационного анализа представлено в предыдущих статьях [1, 2]. Очевидно, что для построения такой структуры должен быть разработан морфологический анализатор текста, который выявляет различные части речи и соотносит их с различными элементами структуры.

1. Постановка задачи для разработки морфологического анализатора

Исходя из концептуальной структуры ситуационного анализа [3] необходимо из текста выделить следующие элементы (рис. 1): субъект действия *Xas*; действие *Xa*; объект действия *Xao*; компоненты действия $\{Xac_1, Xac_2, ..., Xac_N\}$; взаимодействия $\{Rsc_1, Rsc_2, ..., Rsc_N\}$; отношения $\{Rco_1, Rco_2, ..., Rco_N\}$; наборы свойств: $\{Pas_1...Pas_N\}$, $\{Pao_1...Pao_N\}$, $\{Pac_{11}...Pac_{1N}\}$, $\{Pac_{21}...Pas_{2N}\}$ и $\{Pac_{N1}...Pac_{NN}\}$; соотношения $Rp\{Pac_{11}...Pac_{1N}\}$ и $\{Pao_1...Pao_N\}$, $Rp\{Pac_{21}...Pac_{2N}\}$ и $\{Pac_{N1}...Pac_{NN}\}$ [1].

Выдвигается гипотеза, что каждый элемент концептуальной структуры может быть сопоставлен с определенной частью речи. При этом частью речи называется класс слов, объединенных общим грамматическим значением,

набором морфологических признаков и синтаксической ролью в предложении.

Однако сложность анализа любого элемента русского языка объясняется характерными особенностями структуры языка: неоднозначностью формализации языковых правил, законов склонения слов, произвольным порядком слов в предложении и т. д. Данные особенности не позволяют применять упрощенные алгоритмы обработки и анализа русскоязычных текстов. Однако они не являются последними в списке препятствий на пути к достижению корректного морфологического анализа [4]. Например, существует несколько точек зрения по поводу выделения причастий и деепричастий в отдельные части речи. Данная неоднозначность вызвана расхождением взглядов ученых-лингвистов относительно суффиксов, образующих причастия и деепричастия: приверженцы первой точки зрения утверждают, что суффиксы причастий и деепричастий относятся к формообразующим, в то время как их противники придерживаются второй точки зрения, которая гласит, что эти суффиксы являются словообразующими [5].

Существуют и другие аспекты, касающиеся морфологии русского языка, по которым ученые не могут прийти к единому мнению. В данной статье рассматривается традиционная и наиболее распространенная классификация частей речи русского языка.

В русском языке принято различать междометные и немеждометные классы слов. Соответственно, междометия не используются в концептуальной конструкции в связи с их эмоциональным содержанием. Поэтому рассмотрим немеждометные классы слов, которые делятся на служебные и самостоятельные. К служебным частям речи относятся предлоги, частицы и союзы. Самостоятельные части речи можно разделить на две большие группы: именные и глагольные. К именным относятся следующие части речи: имя существительное, имя прилагательное, имя числительное и местоимение. Глагольными частями речи называют глагол, причастие, деепричастие и наречие [6]. Таким образом, будут рассмотрены 11 частей речи, на которые будут отображены элементы концептуальной структуры (рис. 2).

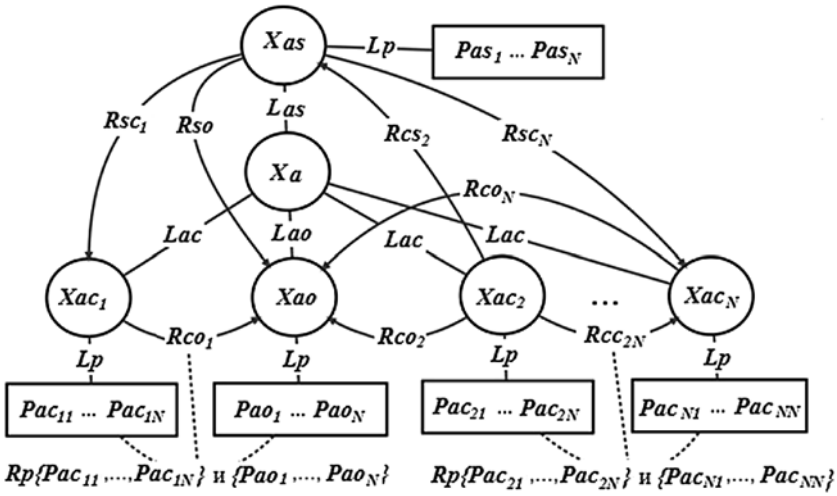

Рис. 1. Концептуальная структура на основе ситуационного анализа

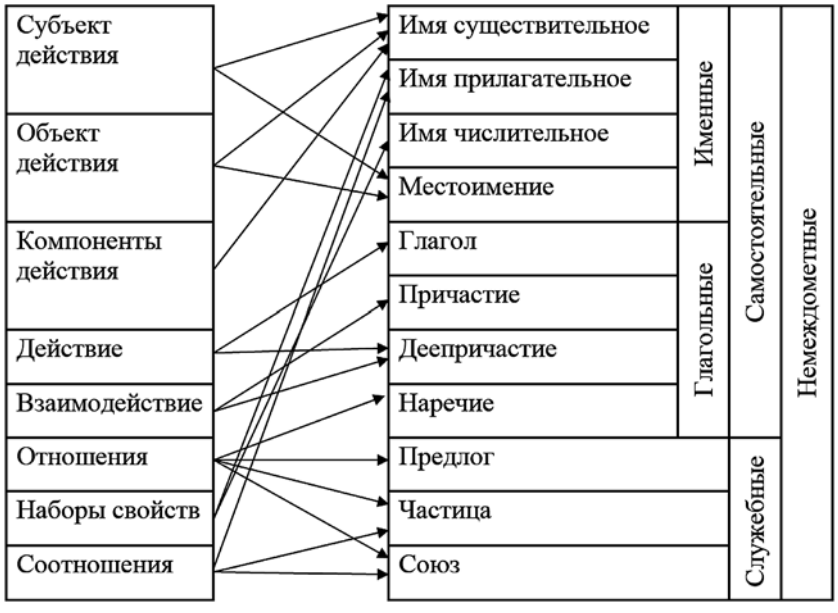

Рис. 2. Отображение элементов концептуальной структуры на части речи

Отображение элементов концептуальной структуры на части речи не является биективным, т. е. взаимно однозначным. Поэтому для снятия противоречий "многие к одному" и "один ко многим" предлагается использовать в разрабатываемом программном обеспечении сочетание словарного и бессловарного морфологических анализаторов.

Задачей морфологического анализа является определение по словоформе нормальной формы слова (леммы), от которой была образована данная словоформа, и набора параметров, приписанных к данной словоформе [7].

Морфологический параметр — это пара <имя параметра, значение

параметра>. Именем параметра может служить род, число, время, склонение, краткость формы прилагательного и другие признаки слов, принятые в данном языке. Значение параметра — это конкретное значение, которое может принимать данный признак. Так, например, род может быть мужским, женским, средним. При этом возможна ситуация, при которой одной словоформе может быть сопоставлено несколько параметров [8].

Важным аспектом при разработке анализатора является такое понятие, как омонимия. Омонимами называют разные по значению, но одинаковые по написанию единицы языка [5]. В терминах компьютерной лингвистики омонимию можно определить как ситуацию, при которой одной словоформе может быть приписано несколько значений для одной грамматической характеристики.

2. Выбор методов морфологического анализа

Морфологический анализ реализуется за счет двух методов: словарной и бессловарной морфологий.

Словарная морфология подразумевает наличие словаря, который содержит набор морфологических параметров для каждой словоформы, а также ее нормальную форму [4]. Соответственно, для реализации словарной морфологии требуется выделить область памяти для хранения словаря. При этом слова, отсутствующие в словаре, проанализировать не представляется возможным. Вдобавок, время поиска будет пропорционально объему базы данных и средней длине слова.

Однако при всех недостатках словарного метода снятие омонимии может быть достигнуто только путем составления дополнительного словаря омонимов, в котором будут содержаться наиболее часто употребляемые словоформы, обладающие свойством омонимии [9]. При этом уменьшение неоднозначности произойдет при построении концептуальной структуры и генерации прямого логического вывода в программном комплексе (ПК) "Оформитель + Решатель" [1]. Напомним, что в данном комплексе в случае обнаружения каких-либо синтаксических и семантических ошибок в концептуальных структурах пользователю выдаются соответствующие сообщения. Таким образом, ПК "Оформитель + Решатель" используется как семантико-синтаксический анализатор, что упрощает устранение конфликтов. Происходит анализ не только неоднозначного слова, но и окружающего контекста. Соответственно, подобным образом решается задача неоднозначного отображения элементов концептуальной структуры на части речи.

Необходимо заметить, что слова могут группироваться по парадигмам. Парадигма — это правила, согласно которым можно получить все формы слов в лексеме (набор словоформ одного слова) для данного стема (неизменяемая часть слова). Бессловарные морфологии хранят только парадигмы слов. При этом в парадигме в качестве постфикса (единица слова, располагающаяся в слове после корня) может храниться только окончание. Также в бессловарной морфологии может храниться набор приставок и суффиксов с привязанной к ним морфологической информацией [4].

Существенным плюсом бессловарных морфологий является то, что они могут предсказать морфологические характеристики практически любого слова, если его парадигма изменения попадает под одну из хранимых. Бессловарные морфологические анализаторы позволяют сэкономить расходы оперативной памяти. Скорость анализа выше, чем у словарных морфологий, а объем хранимых баз значительно сокращается.

К сожалению, такой подход не лишен недостатков, главным из которых является высокий процент ошибок. Так, например, слово "кровать" за счет своего постфикса может быть идентифицировано как глагол в начальной форме. Поэтому были разработаны специальные алгоритмы выявления постфиксов в словах в определенной последовательности, что позволило снизить процент ошибок.

Среди бессловарных методов морфологии выделяют системы, основанные на стемминге (процесс приведения словоформы к неизменяемой форме). В случае стемминга нередко отбрасывается вся морфологическая информация, а в качестве нормальной формы берется неизменяемая псевдооснова, называемая стем. Так, для слова "стена" стемом будет являться строка "стен". Именно эта основа и используется в дальнейшем для идентификации слова во всех его формах.

Стеммер Портера является одним из современных вариантов реализации бессловарной морфологии в чистом виде. В данной реализации применяется алгоритм, предложенный Мартином Портером [10].

3. Программная реализация частеречной разметки текста на основе алгоритма Портера

Под частеречной разметкой текстов (POS-tagging — Part-Of-Speech Tagging) понимается этап автоматического определения частей речи слов в тексте, представленном в цифровом виде. В процессе анализа и обработки текстов присвоение каждому слову соответствующей части речи является наиболее существенной частью морфологического анализа.

Алгоритм Портера не использует при работе базу основ слов, но функционирует за счет последовательного отсечения окончаний и суффиксов согласно определенным алгоритмом правилам [11].

Для полного понимания работы алгоритма введем некоторые определения:

1. Гласные буквы — а, е, и, о, у, ы, э, ю, я. Буква "е" считается равнозначной букве "е".

2. RV — область слова после первой гласной буквы. Она может быть пустой, если гласные в слове отсутствуют.

3. R1 — область слова после первого сочетания "гласная—согласная".

4. R2 — область R1 после первого сочетания "гласная—согласная"

Например, в слове "противоестественном": RV = "тивоестественном"; R1 = "ивоестественном"; R2 = "оестественном".

Среди существующих суффиксов и окончаний выделяются множества, разбитые по группам: PERFECTIVE GERUND (деепричастие), ADJECTIVE (имя прилагательное), PARTICIPLE (причастие), REFLEXIVE (возвратность), VERB (глагол), NOUN (имя существительное), SUPERLATIVE (превосходная степень), DERIVATIONAL (словообразовательные), ADJECTIVAL (PARTICIPLE + ADJECTIVE) (свойства причастия и имени прилагательного).

При поиске окончания из всех возможных выбирается наиболее длинное. Все проверки проводятся над областью RV. Буквы перед RV не участвуют в проверках вообще.

Шаг 1. Найти окончание PERFECTIVE GERUND. Если оно существует — удалить его и завершить этот шаг. Иначе удалить окончание REFLEXIVE (если оно существует). Далее удалить, если существуют, в следующем порядке окончания: ADJECTIVAL, VERB, NOUN. Как только одно из них найдено — шаг завершается.

Шаг 2. Если слово оканчивается на "-и", то удалить "-и"

Шаг 3. Если в R2 находится окончание DERIVATIONAL, то удалить его.

Шаг 4. Возможен один из трех вариантов:

1. Если слово оканчивается на "-нн", то удалить последнюю букву.

2. Если слово оканчивается на SUPERLATIVE, то удалить его и снова удалить последнюю букву, если слово заканчивается на "-нн".

3. Если слово оканчивается на "-ь", то удалить его.

Данный алгоритм используется для нахождения неизменяемой части слова — стема. Однако на его основе с определенной долей погрешности можно выделить пять частей речи: Имя существительное, Имя прилагательное, Глагол, Причастие и Деепричастие, причем для некоторых слов можно получить дополнительные морфологические признаки, такие как "превосходная степень" или "возвратность".

Разработка программного обеспечения (ПО) "Анализатор" для частеречной разметки текстов проводили в программной среде IntelliJ IDEA. Данный выбор обусловлен тем, что IntelliJ IDEA — интегрированная среда разработки программного обеспечения на многих языках программирования, в частности Java, и является одним из наиболее мощных редакторов исходного кода за счет множества встроенных функций, таких как: умное автодополнение, инструменты для анализа качества кода, удобная навигация, расширенные возможности рефакторинга и форматирования [12].

На основе работы ПО "Анализатор" выдается отчет, где текст разбит на слова и для каждого слова определяется не только часть речи, но и элемент концептуальной структуры. Например, слово: Мастер <*Имя существительное*> — [*Субъект Объект Компонент*].

Таким образом, алгоритм Портера успешно применен для решения задачи частеречной разметки русского языка, однако его реализация для русского языка применима только для пяти частей речи.

Для проведения оценки точности работы ПО "Анализатор" и для составления словарей небольшого объема, на которых частично будет основываться анализ, необходимо выбрать словарь, содержащий информацию о словоформах русского языка. Подобные словари называются корпусами и используются при работе словарных морфологических анализаторов.

Корпус — подобранная и обработанная по определенным правилам совокупность текстов, используемых в качестве базы для исследования языка. Они используются для стати-

стического анализа и проверки статистических гипотез, подтверждения лингвистических правил в данном языке [5].

К наиболее популярным корпусам русского языка относятся "Открытый корпус", на основе которого был проведен анализ частоты употреблений словоформ каждой части речи [13] (см. таблицу).

Частота употреблений словоформ каждой части речи в словаре

Часть речи	Число употреблений (словоформ)
Имя существительное	1 006 853
Причастие	899 250
Имя прилагательное	680 557
Глагол	412 235
Деепричастие	68 237
Наречие	4057
Междометие	305
Имя числительное	236
Союз	193
Местоимение	180
Предлог	139
Частица	133

Исходя из результатов проведенного анализа можно заключить, что наибольшим числом словоформ обладают пять частей речи: Имя существительное, Причастие, Имя прилагательное, Глагол, Деепричастие. Именно эти части речи и позволяют классифицировать алгоритм Мартина Портера.

Остальные семь частей речи — Наречие, Междометие, Имя числительное, Союз, Местоимение, Предлог и Частица — целесообразно хранить в форме словаря, так как данные части речи не обладают достаточным числом характерных окончаний, которые позволили бы классифицировать их с достаточной точностью. Следует отметить, что суммарно данные семь частей речи (5243) составляют лишь 0,17 % от общего числа словоформ в словаре (3 072 375). Хранение словаря, содержащего упомянутые семь частей речи, требует всего лишь 578 Кбайт. Данный факт значительно уменьшает объем, занимаемый программным обеспечением. При этом напомним, что междометия в концептуальных структурах не используются.

Заключение

Для определения точности распознавания частей речи проведен анализ сравнения результатов работы разработанного программного обеспечения с результатами работы словарного морфологического анализатора. Каждое совпадение хотя бы с одной из интерпретаций принадлежности слова к той или иной части речи с множеством интерпретаций из словаря помечалось как успешное определение части речи. Если же выявленная в результате анализа часть речи не совпадала ни с одной из тех, что приписывались слову словарем, то такое событие отмечалось как неудачное определение части речи. На основании проведенного сравнения процент точности разработанного проекта, в сравнении со словарем, составил 85,5 %.

Также одним из возможных применений разработанного алгоритма может быть включение его в алгоритм семантико-синтаксического анализа SemSyn [14—16]. Входной язык этого алгоритма составляют вопросы многих видов, команды и утверждения (описания ситуаций) на русском языке. Входные предложения могут включать причастные обороты и придаточные определительные предложения. Алгоритм SemSyn является композицией алгоритмов BuildMatr и BuildSem. Первый алгоритм строит семантико-синтаксическое представление входного текста в виде некоторой строково-числовой матрицы, а второй алгоритм строит по матрице семантическое представление входного текста. Основные процедуры алгоритма BuildMatr запускаются в зависимости от части речи, к которой относится анализируемая лексическая единица. При этом главную роль играют процедуры, обрабатывающие глаголы, причастия, существительные и прилагательные. А это именно те части речи, где хорошо проявил себя представленный в данной статье алгоритм.

Список литературы

1. **Сорокин А. Б., Смольянинова В. А.** Концептуальное проектирование экспертных систем // Информационные технологии. 2017. Т. 23, № 9. С. 634—641.
2. **Сорокин А. Б., Лобанов Д. А.** Концептуальное проектирование интеллектуальных систем // Информационные технологии. 2018. Т. 24, № 1. С. 3—10.
3. **Болотова Л. С.** Системы искусственного интеллекта: модели и технологии, основанные на знаниях. М.: Финансы и статистика. 2012. 663 с.
4. **Боярский К. К.** Введение в компьютерную лингвистику: учеб. пособие. СПб.: НИУ ИТМО, 2013. 72 с.
5. **Большакова Е. И.** Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 272 с.
6. **Соснина Е. П.** Введение в прикладную лингвистику: учеб. пособие. Ульяновск: УлГТУ, 2012. 110 с.
7. **Кронгауз М. А.** Семантика: Учебник. М.: Издательский центр "Академия", 2005. 352 с.

8. **Коваль С. А.** Лингвистические проблемы компьютерной морфологии. СПб.: Изд-во С.-Петербург. ун-та, 2005. 151 с.

9. **Гребенева Ю. Н.** Словарь омонимов и омоформ русского языка. М.: Айрис-пресс, 2011. 351 с.

10. **Porter M. F.** An algorithm for suffix stripping // Program. 1980, vol. 14, no. 3 (14). P. 130—137.

11. **Willett P.** The Porter stemming algorithm: then and now // Program: Electronic Library and Information Systems. 2006, vol. 40, no. 3 (40). P. 219—223.

12. **Ефимов А.** IntelliJ IDEA. Профессиональное программирование на Java. СПб.: БХВ-Петербург. 2005. 800 с.

13. **Открытый** корпус. URL: http://opencorpora.org/, свободный.

14. **Фомичев В. А**. Новый метод преобразования естественно-языковых тестов в семантические представления // Информационные технологии. 2005. Т. 11, № 10. С. 25—35.

15. **Фомичев В. А.** Формализация проектирования лингвистических процессоров. М.: МАКС Пресс, 2005. 368 с.

16. **Фомичев В. А., Разоренов А. А.** Значение теории К-представлений для исследований по автоматическому выявлению семантических ролей // Информационные технологии. 2015. Т. 21. № 6. С. 403—411.

A. B. Sorokin, PhD in Technique, Associate Professor, e-mail: ab_sorokin@mail.ru,
A. P. Kushnarev, master student, e-mail: brainzeater@gmail.com,
Moscow Technological University (MIREA)

Morphological Text Analyzer for Revealing the Completeness of Information

This article discusses the use of technology of automatic part-of-speech tagging of Russian-language texts, presented in digital form, in order to determine the excess or lack of information in the text, the identification and construction of the concept. The main attention is paid to the stage of morphological analysis, as one of the most difficult stages in the analysis of the text, due to the specific features of the morphology of the Russian language, associated with the ambiguity of matching words of a particular part of the speech. Improving the accuracy of the analysis of Russian-language texts is achieved by identifying new patterns among the five parts of speech and by adding new inflectional endings to the existing ones in Porter algorithm. Removal of homonymy is achieved by creating an additional dictionary of homonyms, which will contain the most commonly used word forms that have the property of homonymy. Identification of excessive and insufficient information for understanding the text occurs in the process of constructing a conceptual structure and generating direct logical inference in the software package "Designer + Solver".

Keywords: *conceptual structure, computational linguistics, part-of-speech tagging, POS tagging, automatic processing of documents, processing of texts in natural language, Porter stemming algorithm*

DOI: 10.17587/it.24.719-724

References

1. **Sorokin A. B., Smol'janinova V. A.** Konceptual'noe proektirovanie jekspertnyh system (Conceptual design of expert systems), *Informacionnye Tehnologii*, 2017, vol.23, no. 9, pp. 634—641 (in Russian).

2. **Sorokin A. B., Lobanov D. A.** Konceptual'noe proektirovanie intellektual'nyh sistem (Conceptual design of intelligent systems), *Informacionnye Tehnologii*, 2018, vol. 24, no.1, pp. 3—10 (in Russian).

3. **Bolotova L. S.** *Sistemy iskusstvennogo intellekta: modeli i tehnologii, osnovannye na znanijah* (Systems of artificial intelligence: models and technologies based on knowledge), Moscow, Finansy i Statistika, 2012, 663 p. (in Russian).

4. **Bojarskij K. K.** *Vvedenie v komp'juternuju lingvistiku* (Introduction to computer linguistics), SPb, Publishing house of NIU ITMO, 2013, 72 p. (in Russian).

5. **Bol'shakova E. I.** *Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i komp'juternaja lingvistika* (Automatic processing of texts in natural language and computer linguistics), Moscow, MIJeM, 2011, 272 p. (in Russian).

6. **Sosnina E. P.** *Vvedenie v prikladnuju lingvistiku* (Introduction to Applied Linguistics), Ul'janovsk, UlGTU, 2012, 110 p. (in Russian).

7. **Krongauz M. A.** Semantika (Semantics),Moscow, Izdatel'skij centr "Akademija", 2005, 352 p. (in Russian).

8. **Koval' S. A.** Lingvisticheskie problemy komp'juternoj morfologii (Linguistic Problems of Computer Morphology), SPb., Publishing house of S.-Peterburg. un-ta, 2005, 151 p. (in Russian).

9. **Grebeneva Ju. N.** Slovar' omonimov i omoform russkogo jazyka (Dictionary of homonyms and omoform of the Russian language), Moscow, Ajris-press, 2011, 351 p. (in Russian).

10. **Porter M. F.** An algorithm for suffix stripping, *Program*, 1980, vol. 14, no. 3, pp. 130—137.

11. **Willett P.** The Porter stemming algorithm: then and now, Program, Electronic Library and Information Systems, 2006, vol. 40, no. 3, pp. 219—223.

12. **Efimov A.** IntelliJ IDEA. Professional'noe programmirovanie na Java (Professional programming in Java), SPb., BHV-Peterburg, 2005, 800 p. (in Russian).

13. **Otkrytyj** korpus (Open case) (materialy sajta), available at: http://opencorpora.org/, svobodnyj.

14. **Fomichev V. A.** Novyj metod preobrazovaniya estestvennoyazykovyh tekstov v semanticheskie predstavleniya (A new method for converting natural language tests into semantic representations), *Informacionnye Tehnologii*, 2005, vol. 11, no. 10, pp. 25—35 (in Russian).

15. **Fomichev V. A.** Formalizaciya proektirovaniya lingvisticheskih processov (Formalization of the design of linguistic processors), Moscow, MAKS Press, 2005, 368 p. (in Russian).

16. **Fomichev V. A., Razorenov A. A.** Znachenie teorii K-predstavlenij dlya issledovanij po avtomaticheskomu vyyavleniyu semanticheskih rolej (The importance of the theory of K-representations for research on the automatic detection of semantic roles), *Informacionnye Tehnologii*, 2015, vol. 21, no. 6, pp. 403—411 (in Russian).