DOI: https://doi.org/10.15688/jvolsu2.2017.2.4

UDC 81'33
LBC 81.1

Submitted: 10.03.2017
Accepted: 08.05.2017

AUTOMATION OF THE PROCESS FOR OBTAINING LINGUISTIC INFORMATION: STATE-OF-THE-ART CAPABILITIES

Andrey V. Svetlov
Volgograd State University, Volgograd, Russian Federation

Anatoly S. Komendantov
Volgograd State University, Volgograd, Russian Federation

Abstract. The paper is devoted to the process automation for solution of some problems in linguistic analysis. The review part of the article describes the variety of current linguistic software. We give its classification as follows: electronic dictionaries and thesauri, text conversion programs and text generators, programs for analysis and linguistic processing of documents, natural language processing systems. For each group we mention some examples of relevant applications or web services. In addition, we discuss current capabilities of the software, their scope of use and development prospects. In the main part of the work we overview the add-on we created for the MyStem stemming utility by Ilya Segalovich. The application adds to the features of the utility a user-friendly graphical interface that is easy to learn and intuitive to users who do not specialize in information technology. The algorithm implemented in the software is based on using the results of stemming process to solve some specific problems. It intercepts the output of the MyStem utility, then reformats it and run some specific analysis. The results of this analysis are the basis for main processes of the add-on. This way we can get the frequency analysis of the text, can extract any certain parts of speech, and select inciting words in the text. The examples in this part of paper show the results of all units of the software. In conclusion we made several remarks on the prospects for the development of our application.

Key words: automation, linguistic analysis, morphological analysis, automation of linguistic analysis, automation of morphological analysis, stemming, graphical interface, software shell.

Citation. Svetlov A.V., Komendantov A.S. Automation of the Process for Obtaining Linguistic Information: State-of-the-Art Capabilities. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2, Yazykoznanie [Science Journal of Volgograd State University. Linguistics]*, 2017, vol. 16, no. 2, pp. 39-46. (in Russian). DOI: https://doi.org/10.15688/jvolsu2.2017.2.4.

УДК 81'33
ББК 81.1

Дата поступления статьи: 10.03.2017
Дата принятия статьи: 08.05.2017

АВТОМАТИЗАЦИЯ ПРОЦЕССА ПОЛУЧЕНИЯ ЛИНГВИСТИЧЕСКОЙ ИНФОРМАЦИИ: СОВРЕМЕННЫЕ ВОЗМОЖНОСТИ

Андрей Владимирович Светлов
Волгоградский государственный университет, г. Волгоград, Российская Федерация

Анатолий Сергеевич Комендантов
Волгоградский государственный университет, г. Волгоград, Российская Федерация

Аннотация. Статья посвящена проблемам автоматизации решения некоторых задач лингвистического анализа. Описано многообразие существующего лингвистического программного обеспечения. Приведена

его классификация: электронные словари и тезаурусы; программы преобразования текстов и генераторы текстов; программы анализа и лингвистической обработки документов; системы обработки естественного языка. Для каждой группы даны примеры соответствующих приложений или веб-сервисов, обсуждаются современные возможности программ, сферы их использования и перспективы развития. Основная часть работы посвящена созданной авторами статьи надстройке над утилитой для стемминга MyStem И. Сегаловича. Приложение добавляет к возможностям утилиты удобный графический интерфейс, простой для освоения и интуитивно понятный пользователям, не специализирующимся в информационных технологиях. Функционирование приложения связано с использованием результатов стемминга для решения некоторых специфических задач. Оно перехватывает вывод утилиты MyStem, специальным образом переформатирует и анализирует его. В число задач, которые решаются на основании этой обработки, входит частотный анализ текста, выборка определенных частей речи, выборка побуждений. На примерах продемонстрированы результаты работы всех модулей программы. В заключении намечены некоторые перспективы развития созданного приложения.

Ключевые слова: автоматизация, лингвистический анализ, морфологический анализ, автоматизация лингвистического анализа, автоматизация морфологического анализа, стемминг, графический интерфейс, программная оболочка.

Цитирование. Светлов А. В., Комендантов А. С. Автоматизация процесса получения лингвистической информации: современные возможности // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2017. – Т. 16, № 2. – С. 39–46. – DOI: https://doi.org/10.15688/jvolsu2.2017.2.4.

1. Многообразие лингвистического программного обеспечения

Для решения лингвистических задач анализа текстов часто требуется выполнение большого объема однообразных операций, которые успешно поддаются автоматизации посредством специализированных компьютерных программ [Всеволодова, 2007; Гольдин, Крючкова, 2011; Щипицина, 2013]. Причем в настоящее время круг задач, решение которых можно передать ЭВМ, включает не только относительно простые (построение частотного анализа, синтаксического и морфологического разборов текста), но и более сложные (семантический анализ, автоматическое определение стиля текста или даже его возможного автора). Существует много лингвистических программных продуктов, и лишь часть из них представлена в уже значительно устаревшем каталоге [Логичев]. Условно эти продукты можно разделить на четыре группы.

1. Словари и тезаурусы. Всевозможные электронные формы их бумажных аналогов гораздо более удобны для работы в силу автоматизации процесса поиска. При этом остается актуальным создание снабженных специфическими поисковыми инструментами вариантов таких ресурсов [Андриянов, 2015].

2. Программы преобразования текстов и генераторы текстов. Простейшие приложения этого типа встроены в любой текстовый редактор – они позволяют выполнить автозамену частей документа, удовлетворяющих определенной маске. К этой же группе можно отнести программы для автоматического реферирования и аннотирования текстов – они тоже довольно хорошо знакомы любому квалифицированному пользователю офисных пакетов типа Microsoft Office. Однако специализированные приложения обладают более широкими возможностями. Пожалуй, наиболее масштабно в настоящее время они используются при создании веб-сайтов и их поисковой оптимизации. Они позволяют на основе заданного текста получить множество его псевдоуникальных дубликатов. Работать такие программы могут на основе либо заранее определенного шаблона, либо встроенных словарей синонимов. В качестве примера можно назвать сервис SeoGenerator.ru, приложения Generating the Web, SEO Anchor Generator. Впрочем, существуют и программы полностью самостоятельной генерации текста заданной тематики. Например, проект https://yandex.ru/referats позволяет создать действительно уникальный текст практически на любую тему. Смысловая наполненность его в настоящее время, конечно, находится на уровне шутки.

3. Программы анализа и лингвистической обработки текстов. Элементарные приложения этого типа также известны любому пользователю офисных пакетов. Они позволяют выполнять проверку орфографии, расста-

новку переносов, грамматическую и стилистическую проверку текста, простейший статистический анализ. Кроме того, востребованы программы для построения индексов, конкордансов, частотного анализа – достаточно простые с точки зрения программирования (их существует огромное количество, поэтому мы позволим себе не приводить примеры). При этом компьютеризированию поддаются и более сложные задачи: лемматизация (нормализация, приведение к исходной форме) слов, синтаксический и морфологический анализ. В основе таких приложений, как правило, лежат алгоритмы стемминга, о чем мы будем говорить ниже. Помимо этого, в данную группу лингвистических программ следует включить утилиты для автоматической идентификации и классификации текстов. Они позволяют решать задачи, связанные с выявлением плагиата (например, сервис antiplagiat.ru и его аналоги), определением возможного автора незнакомого текста (например, Лингвоанализатор – www.rusf.ru/books/analysis), установлением функционального стиля текста (например, Худломер – teneta.rinet. ru/2001/hudlomer).

4. Системы обработки естественного языка. Эта группа включает наиболее сложное лингвистическое программное обеспечение. Конечно, оно основано на решении описанных выше задач, но перед ним ставятся более серьезные цели: понимание смысла текста и генерирование грамотного осмысленного ответа. По сути, это направление развития систем искусственного интеллекта. Основное приложение здесь – построение естественноязыкового интерфейса для компьютеров. Примеры таких программ хорошо известны: Siri, Cortana, OKGoogle, как, впрочем, и неидеальное качество их работы. При этом данное направление сейчас активно развивается, подобно другим сферам применения технологии нейронных сетей.

2. Утилиты для стемминга и их приложения

При всем разнообразии имеющегося программного обеспечения часто возникает потребность разработки собственного, адаптированного для решения определенного набора задач. Именно описанию разработанного адаптированного программного обеспечения посвящена данная статья.

Перед нами была поставлена задача создания программы с тремя основными функциями: проведения частотного анализ текста, выборки определенных частей речи, выборки побуждений. Все эти функции для их успешной реализации требуют установления основы каждого слова в тексте (термин «основа слова» в данном случае понимается не традиционно, как это принято в лингвистике, а как компьютерный аналог корневой морфемы). Данная проблема хорошо известна специалистам в области компьютерных наук [Коваленко, 2002; Lovins, 1968; Porter, 1980; Segalovich, 2003] еще с 60-х гг. прошлого века. Разработано несколько удачных алгоритмов для реализации стемминга, то есть отсечения от слова префиксов, суффиксов и окончаний для получения его основы. Все эти алгоритмы, по сути, отличаются способом представления правил, по которым определяется, какие части слова не входят в его основу и могут быть отсечены. Помимо этих правил алгоритмы могут использовать таблицы поиска, словари или корпус языка ([Коваленко, 2002; Lovins, 1968; Porter, 1980; Segalovich, 2003], а также https://tech.yandex.ru/mystem/; http://snowballstem.org/; https://nlpub.ru/Stemka).

Для русского языка наиболее эффективными являются стеммеры Snowball (стеммер Портера), Stemka и MyStem. Все они распространяются бесплатно и могут быть использованы для морфологического анализа текста и для создания собственного программного продукта на их основе. Однако они не имеют графического интерфейса и потому не очень удобны для применения конечным пользователем, не имеющим определенной квалификации в области информационных технологий. Учитывая результаты анализа функциональных возможностей и производительности различных алгоритмов [Segalovich, 2003], для решения поставленной перед нами лингвистической задачи мы выбрали стеммер MyStem, разработанный И. Сегаловичем (сооснователем Яндекса), и построили для него визуальный интерфейс.

Приложение написано на языке C# для операционной системы MSWindows А.С. Комендантовым, В.А. Ксензом, А.Г. Матвее-

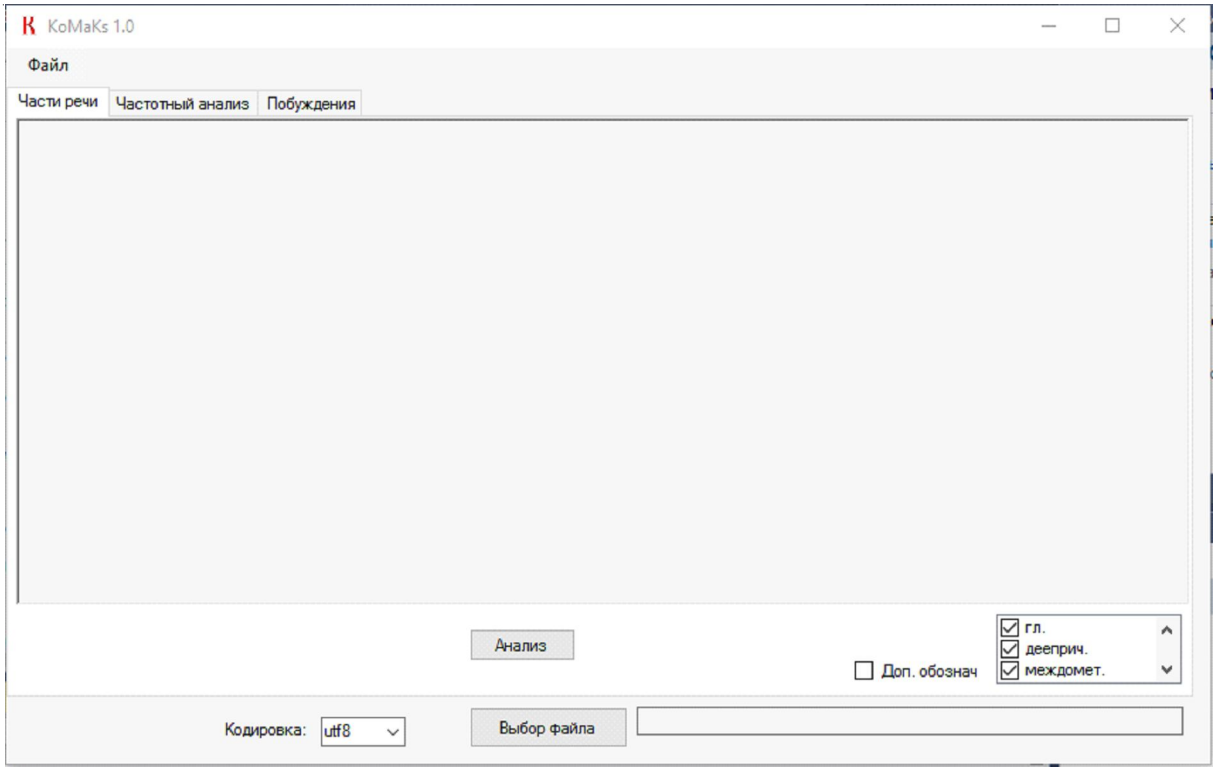

Рис. 1. Общий вид интерфейса приложения

вым под руководством канд. физ.-мат. наук, доц. А.В. Светлова и д-ра филол. наук, проф. С.В. Ионовой.

Интерфейс представляет собой окно, имеющее три вкладки: части речи, частотный анализ, побуждения (см. рис. 1). В каждой из вкладок содержатся функции, требуемые для данного вида работы.

Чтобы начать работу с программой, требуется определить путь к файлу. Для этого нужно нажать на кнопку «Выбор файла». На данный момент программа поддерживает импорт только из txt файлов. Далее нужно выбрать из выпадающего списка кодировку файла. После этого в зависимости от открытой вкладки определяются дополнительные параметры. Например, во вкладке «Части речи» нужно выбрать требуемые для выборки части речи, а во вкладке «Частотный анализ» можно определить, нужно ли подсчитывать служебные части речи.

На следующем шаге после выбора параметров при нажатии кнопки «Анализ» происходит вызов исполняемого файла MyStem с определенными опциями, которые зависят от установленных параметров и вида анализа. Стеммер анализирует текст, затем возвращает xml-код, содержащий вероятные начальные формы слов и граммемы, грамматические признаки слов.

Например, xml выглядит таким образом:

– для существительного:

<w>ночей<analex=“ночь” gr=“S,жен,неод=род,мн” /></w>

– для глагола:

<w>спал<analex=“спать” gr=“V,несов,нп=прош,ед, изъяв,муж” /><analex=“спадать” gr=“V,нп=прош,ед, изъяв,муж,сов” /></w>

В тегах ana перечисляются все вероятные разборы. В параметре gr указываются граммемы. Тегами <se></se> обозначаются предложения.

Программа читает xml-вывод стеммера, используя набор расширений LINQ. Выбирается разбор с наибольшим весом, то есть наиболее вероятная начальная форма и набор граммем. Далее в зависимости от выбранной задачи программа производит требуемый анализ.

Например, при установлении частей речи из всех граммем, определенных стеммером, выбираются параметры с грамматическими метками. Их обозначения переводятся на русский язык, приводятся к привычному виду. Если была поставлена галочка «Доп. обозначения»,

к некоторым словам добавляются имеющиеся уточнения, определяемые стеммером. У имен собственных устанавливается их тип: фамилия, имя, отчество, географическое название; помечается обсценная лексика, разговорные формы, искаженные формы, сокращения и др.

Полученные данные структурируются, объединяются в строку и выводятся на экранную форму, в графический интерфейс программы. Пользователь может получить информацию прямо из интерфейса программы в удобном и наглядном виде, а также может воспользоваться функцией сохранения в файл (доступны форматы txt и rtf). На рисунке 2 на примере текста поэмы А.Т. Твардовского «Василий Теркин» показана работа программы по автоматическому определению частей речи.

При частотном анализе действует следующий алгоритм: для каждой начальной формы слова проверяется, встречалась ли она ранее. Если нет, начальная форма этого слова добавляется в список и число ее появлений в тексте становится равно единице. Если начальная форма уже встречалась, то число ее появлений увеличивается на единицу. Затем из полученного списка формируется таблица и выводится на экранную форму (см. рис. 3, в качестве текста для анализа, как и выше, используется поэма А.Т. Твардовского «Василий Теркин»). В этой таблице доступны изменения порядка сортировки – по алфавиту или по количеству появлений слова в тексте.

При извлечении из текста побуждений программа ищет в источнике глаголы. Далее проверяется наклонение глагола. Формы повелительного наклонения отмечаются цветом. Затем происходит вывод исходного текста на экранную форму с выделенными глаголами повелительного наклонения. На рисунке 4 показан пример автоматического поиска побуждений в текстах песен Б. Окуджавы «Веселый барабанщик» и «Последний троллейбус». Заметим, что форматирование исходного текста программа не сохраняет.

3. Перспективы работы

Разработанная программа требует усовершенствования. Она допускает ошибки при определении части речи в случаях омонимии форм, для разграничения которых необходим анализ контекста употребления слова (например, *пищи* в зависимости от контекста должно быть определено как форма повелительного наклонения

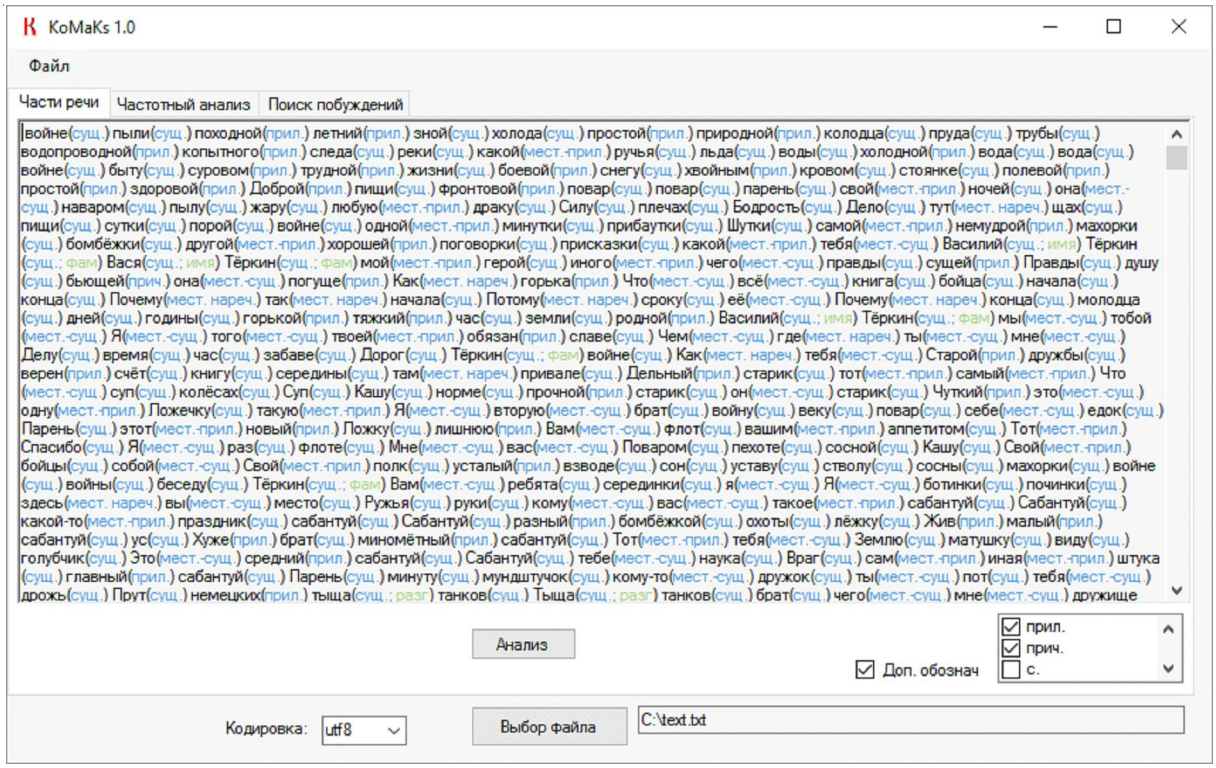

Рис. 2. Пример автоматического определения частей речи

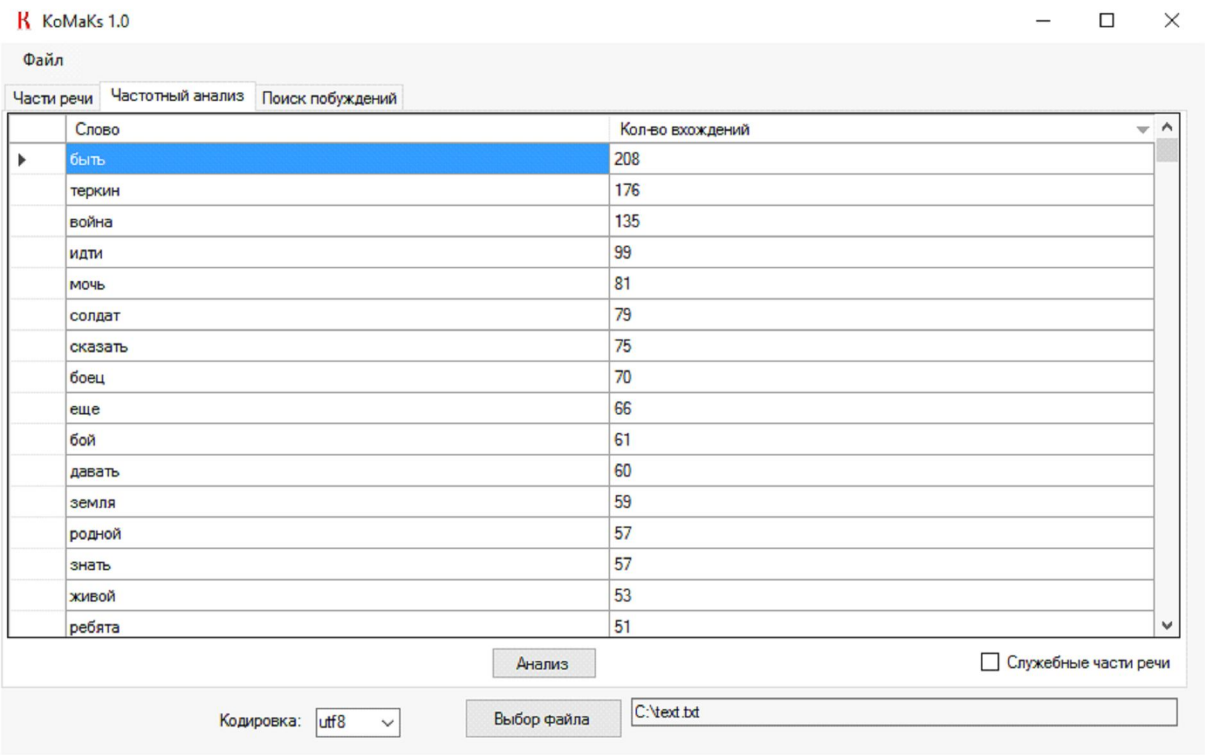

Рис. 3. Пример построения частотного анализа

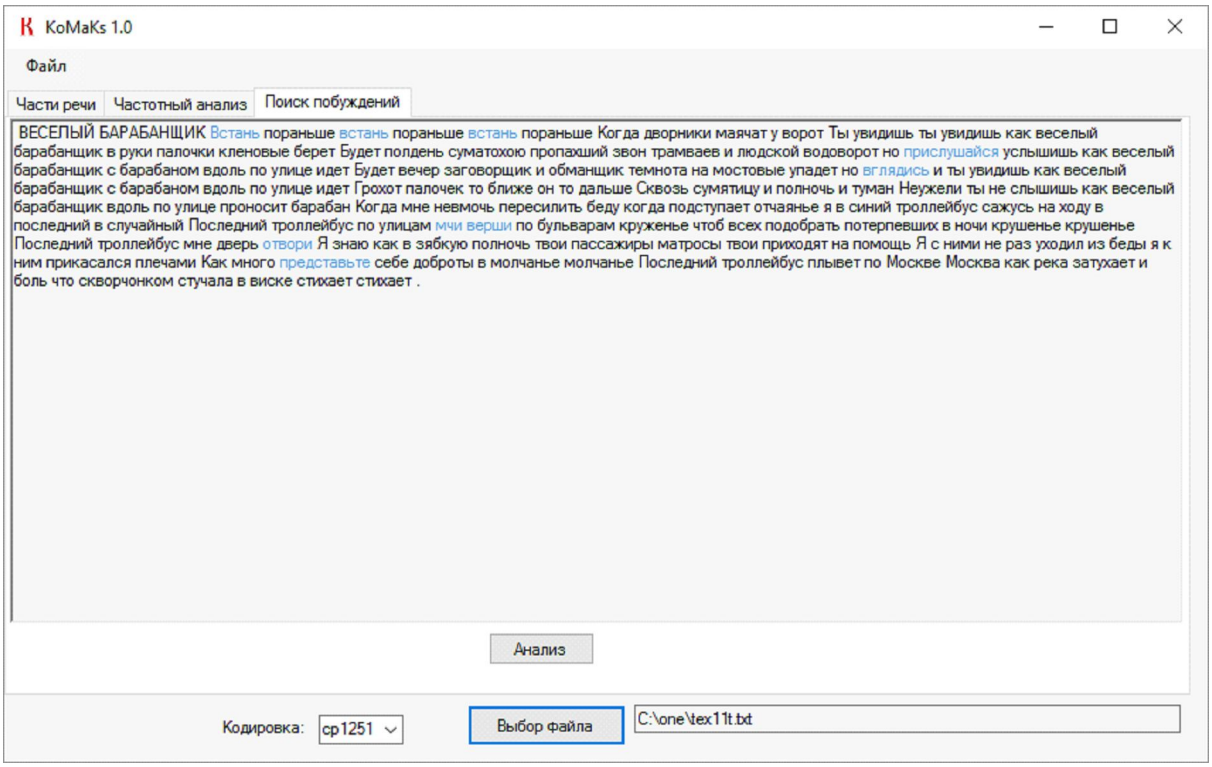

Рис. 4. Пример поиска побуждений в тексте

глагола или как форма родительного падежа существительного, однако программа в таких случаях делает практически случайный выбор варианта). Поэтому требуется вмешательство конечного пользователя и коррекция результата. Следовательно, необходимо минимизировать подобные ошибки, совершенствуя алгоритмы стемминга, позволяющие решать более сложные задачи, в частности задачу семантического анализа. Кроме того, планируется развитие программы, предполагающее добавление новых функций, прежде всего выбора части текста для проведения анализа, отделения дескриптивных слов, которые не связаны с содержанием текста (например, указаний на говорящего в диалогах, ремарок).

СПИСОК ЛИТЕРАТУРЫ

Андриянов, Д. В. Проектирование информационной системы для выборки словарных статей по стилистическим пометам / Д. В. Андриянов // Актуальные направления научных исследований XXI века: теория и практика. – 2015. – Т. 3, № 7, ч. 3. – С. 304–307.

Всеволодова, А. В. Компьютерная обработка лингвистических данных / А. В. Всеволодова. – М. : Флинта : Наука, 2007. – 96 с.

Гольдин, В. Е. Введение в электронные лингвистические ресурсы / В. Е. Гольдин, О. Ю. Крючкова. – Саратов : Изд-во СГУ, 2011. – 63 с.

Коваленко, А. Вероятностный морфологический анализатор русского и украинского языков / А. Коваленко // Системный администратор. – 2002. – № 1. – С. 66–75.

Логичев, С. В. Каталог лингвистических программ и ресурсов в Сети / С. В. Логичев // Русская виртуальная библиотека. – Электрон. текстовые дан. – Режим доступа: http://rvb.ru/soft/catalogue/catalogue.html. – Загл. с экрана.

Щипицина, Л. Ю. Информационные технологии в лингвистике / Л. Ю. Щипицина. – М. : Флинта : Наука, 2013. – 126 с.

Lovins, J. B. Development of a stemming algorithm / J. B. Lovins // Mechanical Translation and Computational Linguistics. – 1968. – Vol. 11, № 1–2. – P. 22–31.

Porter, M. F. An algorithm for suffix stripping / M. F. Porter // Program. – 1980. – Vol. 14, iss. 3. – P. 130–137.

Segalovich, I. A fast morphological algorithm with unknown word guessing inducedby a dictionary for a web search engine / I. Segalovich // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. – Las Vegas : CSREA Press, 2003. – P. 273–280.

REFERENCES

Andriyanov D.V. Proektirovanie informatsionnoy sistemy dlya vyborki slovarnykh statey po stilisticheskim pometam [Designing an Information System for the Selection of Entrieson Stylistic Litters]. *Aktualnye napravleniya nauchnykh issledovaniy XXI veka: teoriya i praktika* [Up-to-Date Areas of 21st Century Research: Theory and Practice], 2015, vol. 3, iss. 7, part 3, pp. 304-307.

Vsevolodova A.V. *Kompyuternaya obrabotka lingvisticheskikh dannykh* [Computer Processing of Linguistic Data]. Moscow, Flinta Publ.; Nauka Publ., 2007. 96 p.

Goldin V.E., Kryuchkova O.Yu. *Vvedenie v elektronnye lingvisticheskie resursy* [Introduction to Electronic Linguistic Resources]. Saratov, Izd-vo SGU, 2011. 63 p.

Kovalenko A. Veroyatnostnyy morfologicheskiy analizator russkogo i ukrainskogo yazykov [Probabilistic Morphological Analyzer for the Russian and Ukrainian Languages]. *Sistemnyy administrator* [System Administrator], 2002, no. 1, pp. 66-75.

Logichev S.V. Katalog lingvisticheskikh programm i resursov v Seti [Catalog of Linguistic Programs and Resources on the Internet]. *Russkaya virtualnaya biblioteka* [Russian Virtual Library]. URL: http://rvb.ru/soft/catalogue/catalogue.html.

Shchipitsina L.Yu. *Informatsionnye tekhnologii v lingvistike* [Information Technology in Linguistics]. Moscow, Flinta Publ.; Nauka Publ., 2013. 126 p.

Lovins J.B. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 1968, vol. 11, no. 1-2, pp. 22-31.

Porter M.F. An Algorithm for Suffix Stripping. *Program*, 1980, vol. 14, iss. 3, pp. 130-137.

Segalovich I. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications*. Las Vegas, CSREA Press, 2003, pp. 273-280.

Information About the Authors

Andrey V. Svetlov, Candidate of Sciences (Physics and Mathematics), Associate Professor, Department of Mathematical Analysis and Function Theory, Volgograd State University, Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation, andrew.svetlov@volsu.ru, matf@volsu.ru, http://orcid.org/0000-0002-8764-6132.

Anatoly S. Komendantov, Student, Institute of Mathematics and IT, Volgograd State University, Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation, matf@volsu.ru, http://orcid.org/0000-0001-5009-498X.

Информация об авторах

Андрей Владимирович Светлов, кандидат физико-математических наук, доцент кафедры математического анализа и теории функций, Волгоградский государственный университет, просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация, andrew.svetlov@volsu.ru, matf@volsu.ru, http://orcid.org/0000-0002-8764-6132.

Анатолий Сергеевич Комендантов, студент института математики и информационных технологий, Волгоградский государственный университет, просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация, matf@volsu.ru, http://orcid.org/0000-0001-5009-498X.