УДК 004.94

ИНФОРМАЦИОННАЯ СИСТЕМА АНАЛИЗА И ЛИНГВИСТИЧЕСКОЙ ОБРАБОТКИ ЛИТЕРАТУРНОГО ТЕКСТА

ГАМИДУЛЛАЕВА ЛЕЙЛА АЙВАРОВНА
ПКИТ (филиал) ФГБОУ ВО МГУТУ им. К.Г. Разумовского (ПКУ), к.э.н., доцент кафедры прикладной и бизнес информатики, Россия, г. Пенза, 440000, ул. Володарского, 6; gamidullaeva@gmail.com

ТАКТАШКИН ДЕНИС ВИТАЛЬЕВИЧ
ПКИТ (филиал) ФГБОУ ВО МГУТУ им. К.Г. Разумовского (ПКУ), к.т.н., доцент кафедры прикладной и бизнес информатики, Россия, г. Пенза, 440000, ул. Володарского, 6; e-mail: dead-corpse@mail.ru.

Аннотация.

Современные писатели вынуждены использовать специализированное программное обеспечение по работе с текстом, которое отличается от типовых текстовых процессоров. В статье представлен анализ существующих программ для писателей. Выявлены ключевые недостатки использования данных программ. Показано, что имеется потребность в разработке информационной системы анализа и лингвистической обработки литературного текста, оптимизирующей работу автора над литературным текстом. Сделан вывод о том, что для разработки такого программного средства требуется наличие формализованной модели и соответствующих алгоритмов фонетического анализа.

Ключевые слова: программное обеспечение; лингвистическая обработка; анализ текста; программы ассистент литератора; инструментарий писателя; фонетические алгоритмы.

Введение.

В современном мире без работы с электронными средствами обработки информации не обходится ни одна компания, отрасль или физическое лицо. С каждым годом инженеры выпускают все более новые и улучшенные компьютерные устройства. Несмотря на достаточно малый период развития, компьютеры прочно укрепились в жизни общества – они стали лучшими помощниками практически во всех важных вопросах. Количество операций, выполняемых техникой, продолжает расти с каждым днем. Компьютерные технологии практически заменили различные печатные издания, такие как справочники, учебники, словари и тому подобное. Ведь гораздо проще включить компьютер и найти интересующую для себя информацию, чем ходить по магазинам в поисках нужной книги. В сравнении с процессом написания литературного произведения вручную, компьютерные программные средства помогают работать с произведением в электронном виде, исправлять различные стилистические, а также грамматические ошибки, вставлять в текст изображения и т.д. Персональные компьютеры помогают авторам создавать произведения и отправлять их на принтер. Для пользователей, работающих с литературным текстом, этот фактор играет ключевую роль. Программы для писателей – это десктопные приложения, с помощью которых можно писать, редактировать и проводить анализ литературных произведений разных жанров. Примерами десктопных приложений являются Microsoft Word, Блокнот и т.д. Для их работы требуются ресурсы персонального компьютера, а также наличие самого приложения и набора необходимых библиотек, содержащих требуемые функции для работы с программным средством.

На данный момент существует ограниченное количество подобных программ для писателей, но большинство из них распространяются на коммерческой основе, являются разработками иностранных компаний и не ориентированы на специфическую работу с русским литературным текстом.

Одним из главных недостатков является отсутствие поддержки русского языка. Если в одном из таких приложений запустить, к примеру, фонетический анализ текста, то в большинстве случаев результат обработки русскоязычного текста окажется неудовлетворительным, так как алгоритмы анализа текстов на разных языках значительно отличаются. Также к недостаткам иностранного программного обеспечения можно отнести неполную русификацию интерфейса и отсутствие справочных материалов на русском языке.

На сегодняшний момент программы для писателей становятся основным средством для создания качественных и оригинальных литературных произведений.

Данные программные средства можно смело считать новой ступенью развития современных писательских инструментов. Однако перед тем, как появились электронные средства обработки и выдачи данных, люди были вынуждены обходиться ручными инструментами для переноса слов на долговременные носители информации.

Первыми инструментами для изображения символов были палка и камень. Палкой и камнем можно было создавать изображения на земле или песке. Около 4000 лет до н.э. люди начали использовать для письма смоченные водой глиняные дощечки, на которых они писали деревянной или бронзовой палочкой. Спустя тысячу лет до н.э. в Египте изобрели новую форму письма в виде иероглифов. Для рисования на папирусе люди применяли тонкие тростниковые кисти.

В период с 600-х до 1800-х годов н.э. распространение пергамента привело разработке общедоступных пишущих инструментов. Примером такого инструмента стало заточенное гусиного перо, с помощью которого можно стало изменить стиль письма. Гусиные перья (рисунок 1) просуществовали до конца восемнадцатого века. На данный момент это рекордный период для всех существующих пишущих инструментов.

В 1790 году австралийцы и французы независимо друг от друга предложили для письма новый инструмент – карандаш с грифелем. Это и послужило началом возникновения индустрии канцелярских товаров.

Спустя 80 лет страховой агент Левис Эдсон Ватерман изобретает металлическую ручку заправляемую чернилами, с помощью пипетки. Несмотря на то, что изобретение чернильной ручки заметно продвинула прогресс развития пишущих инструментов, большая часть писателей склонялась к использованию в своей работе печатного варианта письма.

Первая печатная машина была изготовлена в 1808 году итальянцем Пеллегрино Тури для графини Каролины Фантони да Фивизоно. Каролина была слепа, а при помощи такого аппарата она могла переписываться со своими родственниками.

Следующая попытка создать пригодное для комфортной работы с печатным текстом устройство была предпринята в России, когда Михаил Иванович Алисов разработал наборно-пишущую машину. Михаил Иванович хотел упростить и облегчить процедуру переписывания рукописей и оригиналов. Реализация этой идеи завершилась успехом, однако высокая стоимость изобретения остановила дальнейшее развитие этого изделия.

Стоит отметить, что Л.Н. Толстой был большим поклонником печатных машинок. Свои романы он писал только на машинке от «Ремингтон», в то время как русский поэт В.В. Маяковский предпочитал печатать на «Андервуде» [1].

В настоящее время писатели работают с различными текстовыми редакторами. Начиная от стандартного «Блокнота» и заканчивая таким мощным программным обеспечением для офисных работ как «MS Word». В данной работе продукт от Microsoft целенаправленно называется инструментом для офисных работ, так как в этой программе отсутствует специализированный, писательский функционал. Примерами таких специализированных инструментов могут являться: база данных персонажей произведения; система таймлайна; функция удаления лишних символов в тексте, которая позволяет

сокращать количество знаков в произведении; модуль по работе со словарями, позволяющий выяснить значение слова или подобрать к нему нужные синонимы; инструмент формирования статистики использованных слов в тексте, помогающий автору избавится от «слов-паразитов» в произведении; модуль фонетического анализа текста, обеспечивающий анализ литературного произведения на фонетические тавтологии [2, 3].

На данный момент все нужные писателям инструменты разбросаны по разному программному обеспечению, такому как: yWriter, Liquid Story Binder, Fresh Eye, Wordstat, Reword и т.п. Все это вынуждает писателя устанавливать по отдельности огромное количество программных средств. Такой подход несомненно может отрицательно сказаться на качестве работы писателя, так как приходится отвлекаться на управление несколькими запущенными приложениями. Кроме этого большинству современных программ для писателей присуще такие недостатки как: отсутствие поддержки работы с русскими текстами, частичная русификация интерфейса, отсутствие справочных материалов на русском языке. Такие проблемы связанны главным образом с тем, что на рынке программного обеспечения отсутствуют российские аналоги программ, относящихся к категории – ассистент-литератора.

Основная часть.

Рассмотрим примеры проведения фонетического анализа фрагментов литературных произведений, с помощью различных алгоритмов реализованных в программном средстве для писателей «Сюжет» [4]. В качестве первого фрагмента предлагается оценить один абзац из книги Серовой М.С. «Десять карат несчастий». Как видно из результата проверки (рис. 1), на 135 слов данного абзаца автор использовал 6 раз слово «что» и 1 раз созвучное ему слово «чтобы». В отличие от Серовой М.С., братья Стругацкие в своем произведении «Пикник на обочине» позволяли себе использовать слово «что» не более чем 9 раз на более чем 1300 слов художественного текста (рис. 2).

– Мадемуазель, это вам, – произнес он три коротких слова, приятно согревших мне душу. Кто бы мог подумать, что Владик может быть таким очаровательным? Потом он передал мне конверт, объяснив, что это аванс. Я тщательно пересчитала деньги и спрятала их в надежном месте. – По-моему, вчера я слишком разоткровенничался, – сказал Владик нерешительно, когда я вернулась с кухни без драгоценной ноши в руках. Розу я поместила в высокий хрустальный стакан, наполнив его водой. День начинался потрясающе. Я была полна решимости сделать все, что в моих силах, чтобы Владик сумел справиться с неприятностями с наименьшими потерями. У него ведь на меня вся надежда, и я вся холодела при одной мысли, что с ним может что-то случиться. Он начинал мне нравиться, и я уже не раскаивалась, что взялась за это дело. Больше он не повторял своих нахальных попыток.

Рис. 1 Пример фонетического анализа

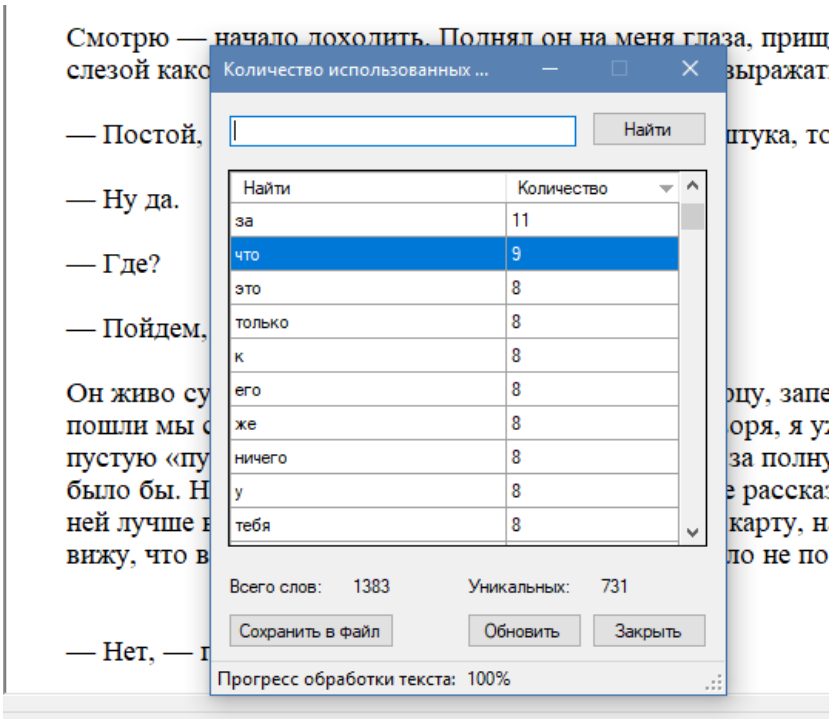

Рис. 2 Анализ фрагмента «Пикник на обочине»

Приведем пример анализа одного и того же фрагмента текста из произведения «Десять карат несчастий» с помощью разных фонетических алгоритмов обработки. Проанализируем результат работы адаптированного к русскому языку алгоритма Double metaphone (рис. 3). Как видно из приведенного анализа были обнаружены всего 3 фонетически схожих комбинации слов:

– случайный, случайность, случайностей;
– вероятность, вероятным;
– намеренный, намерения.

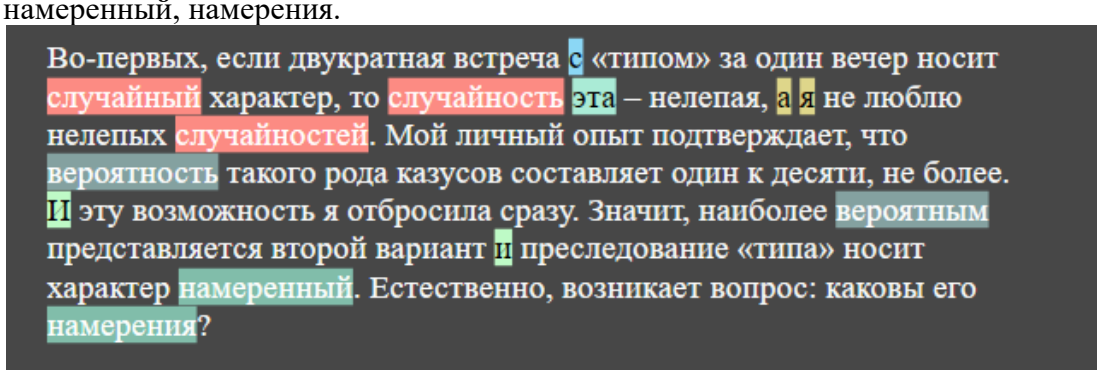

Рис.3 Анализ фрагмента алгоритмом Double metaphone

При анализе данного фрагмента текста комплексным фонетическим алгоритмом можно получить более точную фонетическую картину текста (рис. 4). В дополнение к предыдущим аллитерациям были также найдены:

– нелепая, нелепых;
– более, наиболее;
– составляет, представляется.

Во-первых, если двукратная встреча с «типом» за один вечер носит случайный характер, то случайность эта – нелепая, а я не люблю нелепых случайностей. Мой личный опыт подтверждает, что вероятность такого рода казусов составляет один к десяти, не более. И эту возможность я отбросила сразу. Значит, наиболее вероятным представляется второй вариант и преследование «типа» носит характер намеренный. Естественно, возникает вопрос: каковы его намерения?

Рис. 4 Результат комплексного анализа

Многие авторы часто используют в своих произведениях такой известный литературный прием, как аллитерация. Он заключается в том, что автор намеренно использует созвучные слова с одинаковым количеством согласных звуков, тем самым создавая нужный ему звуковой эффект. К примеру, передавая рев или клокотание бурлящей воды. К такой технике очень часто прибегал А.С. Пушкин (рис. 5).

Нева вздувалась и ревела, Котлом клокоча и клубясь…

Рис. 5 Пример целенаправленной аллитерации

В противовес данному техническому приему случайная аллитерация раздражает и отвлекает от сути текста. Использование фонетических алгоритмов обработки литературного текста позволяют избежать подобных случаев и исключить случайную аллитерацию из произведения.

В заключение проведем сравнительный анализ на фонетическую оригинальность двух произведений «Сашка» Михаила Юрьевича Лермонтова и «Волшебный свет» Татьяны Устиновой. Результаты проведенной проверки можно увидеть на рисунке 6.

Рис. 6 Анализ фонетической оригинальности

Для тестирования были выбраны одинакового объема фрагменты текста, включающие 1257 слов. Учитывая то, что «Сашка» представляет собой нравственную поэму, а «Волшебный свет» детективную прозу можно было бы ожидать от произведения, написанного в стихах довольно низкий процент фонетической оригинальности. В какой-то мере это обуславливается особенностями жанра, требующими от автора поэтичности и

певучести изложения, а также особенностями формирования рифм. Но, несмотря на это фонетическая оригинальность поэмы составила 89 % против 81 % современной детективной прозы.

Заключение.

К сожалению, в настоящее время публикуется очень немного работ посвященных разноаспектному анализу литературного текста. Большинство публикаций являются достаточно узконаправленными в плане их практического применения. Для того, чтобы анализ текста художественных произведений получил свое распространение, необходимо построить его формальную модель, разработать методы и алгоритмы обработки текста.

На данный момент можно заметить положительные тенденции в применении современными авторами новых информационных технологий при работе над своими литературными произведениями. При этом общепризнанных комплексных средств и методов анализа до сих пор не существует. Пока существуют только отдельные программные решения, позволяющие решать только отдельные узконаправленные вопросы, связанные с текстом произведения.

Ввиду выше обозначенных причин имеется потребность в разработке информационной системы анализа и лингвистической обработки литературного текста, оптимизирующей работу автора над литературным текстом. Однако для разработки такого программного средства требуется наличие формализованной модели и соответствующих алгоритмов фонетического анализа.

Большинство фонетических алгоритмов, в том числе и рассмотренные выше, в первую очередь ориентированы на использование в виде соответствующих плагинов для различных СУБД или для поисковых движков, таких как Apache Lucene. Изначальная область их применения очень специфична и ограничивается на данный момент только рамками быстрого поиска фамилий в базах данных.

Модификация фонетических алгоритмов позволяет решить и еще одну очень важную проблему, с которой сталкиваются авторы при редактировании художественных произведений. К примеру, благодаря высокой стабильности, эффективности в работе и хорошей приспособленности к правилам русского языка данные алгоритмы могут быть успешно применены для поиска фонетически схожих сегментов литературного текста. Это позволит авторам вычищать из художественных текстов такие стилистические погрешности, как паронимия или случайная тавтология.

Список литературы.

[1] Такташкин, Д.В. История развития рабочих инструментов писателя / Д.В. Такташкин, И.А. Масенко // Гуманитарные научные исследования. 2016. № 12 (64). С. 55-61.

[2] Такташкин Д.В., Масенко И.А. Проектирование и разработка структуры специализированного программного средства для писателей «Сюжет»/ Д.В. Такташкин, И.А. Масенко // Современные научные исследования и инновации. 2017. № 3 [Электронный ресурс]. URL: http://web.snauka.ru/issues/2017/03/77761 (дата обращения: 23.09.2018).

[3] Такташкин Д.В., Масенко И.А. Модель вариантов использования программы для писателей «Сюжет» / Д.В. Такташкин, И.А. Масенко // Современные научные исследования и инновации. 2016. № 3 [Электронный ресурс]. URL: http://web.snauka.ru/issues/2016/03/64882 (дата обращения: 23.09.2018).

[4] Такташкин, Д.В. Особенности реализации интерфейса программы для писателей «Сюжет» / Д.В. Такташкин, И.А. Масенко // Современная техника и технологии. 2016. № 5 [Электронный ресурс]. URL: http://technology.snauka.ru/2016/05/9904 (дата обращения: 24.09.2018).

INFORMATION SYSTEM ANALYSIS AND LINGUISTIC PROCESSING OF LITERATURE TEXT

GAMIDULLAEVA LEYLA AYVAROVNA
PhD, Associate Prof., K.G. Razumovsky Moscow State University of technologies and management (the First Cossack University), RF, Penza, Volodarskogo 6,

TAKTASHKIN DENIS VITALIEVICH
PhD, Associate Prof., K.G. Razumovsky Moscow State University of technologies and management (the First Cossack University), RF, Penza, Volodarskogo 6, e-mail: dead-corpse@mail.ru.

Abstract.
Modern writers are forced to use specialized software for working with text that differs from typical text processors. The article presents an analysis of existing programs for writers. The key disadvantages of using these programs are identified. It is shown that there is a need to develop an information system for the analysis and linguistic processing of literary text that optimizes the author's work on literary text. It is concluded that the development of such software requires the presence of a formalized model and corresponding phonetic analysis algorithms.

Keywords: software; linguistic processing; text analysis; assistant writer programs; toolkit writer; phonetic algorithms.

References.

[1] Taktashkin, D.V. The history of the development of working tools of the writer / D.V. Taktashkin, I.A. Masenko // Humanitarian scientific research. 2016. № 12 (64). Pp. 55-61.

[2] Taktashkin D.V., Masenko I.A. Design and development of the structure of specialized software for writers "Plot" / D.V. Taktashkin, I.A. Masenko // Modern scientific research and innovation. 2017. № 3 [Electronic resource]. URL: http://web.snauka.ru/issues/2017/03/77761 (appeal date: 09/23/2018).

[3] Taktashkin D.V., Masenko I.A. Model of the use of the program for writers "The plot" / DV Taktashkin, I.A. Masenko // Modern scientific research and innovation. 2016. № 3 [Electronic resource]. URL: http://web.snauka.ru/issues/2016/03/64882 (appeal date: 09/23/2018).

[4] Taktashkin, D.V. Features of the implementation of the program interface for writers "Theme" / D.V. Taktashkin, I.A. Masenko // Modern technology and technology. 2016. № 5 [Electronic resource]. URL: http://technology.snauka.ru/2016/05/9904 (appeal date: 09/24/2018).