УДК 004.912

ПРОГРАММНЫЕ СРЕДСТВА ДЛЯ РЕДАКТОРСКОЙ ДЕЯТЕЛЬНОСТИ. ПОИСК И АНАЛИЗ ФРАГМЕНТОВ ТЕКСТА

© Л. Г. Ильина, аспирантка, Л. Н. Марголин, доцент, МГУП, Москва, Российская Федерация

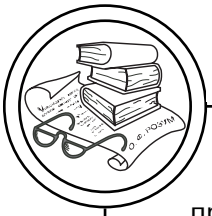
Розглядаються функціональні можливості нового пакету програм (мова VBA) для середовища текстового процесора MSO Word. Пакет призначено для оптимізації деяких етапів аналізу тексту, який проводиться при редагуванні.

The article deals with the functional possibilities of new software package (VBA language) for the medium of the text processor MSO Word. This package is aimed at improvement of certain stages of the text analysis performed during its editing.

В то время как информационные технологии полностью охватили полиграфические производственные процессы, в работе редактора над рукописью они используются значительно реже. И хотя уже давно ведётся разработка систем искусственного интеллекта, в подготовку произведения к изданию они не внедрены. Это объясняется тем, что редакторский анализ текста носит творческий характер и поэтому трудно поддаётся формализации. Ещё в 70-х годах прошлого века разработчики первых программ обработки текста пришли к выводу, что наиболее эффективны не чисто машинные способы решения задач, а методы, обеспечивающие тесное взаимодействие человека и вычислительной машины. При редакторском анализе существует ряд операций, которые могут быть значительно более успешно, а главное, — неизмеримо более быстро и точно — осуществлены машиной, а не человеком, при этом за ним остаётся контроль за результатами и выбор оптимальных решений из числа предлагаемых программой. Вместе с тем, для выполнения некоторых несложных для человека операций, например, выделения корня слова, в программу требуется загрузить огромную базу знаний, соответствующую обычной грамотности человека — носителя языка.

По такому принципу совместной работы построен программный пакет «Поиск и анализ фрагментов текста». Он предназначен для работы с документами в среде MSO Word 2003 (и более поздних версий) и состоит из пяти автономных программ, имеющих различное функциональное назначение: «Заимствования», «Повторы», «Цитаты», «Образы», «Словари» (рис. 1).

— Программа «Заимствования» сравнивает два документа между собой и отображает похожие фрагменты текста, имеющиеся в обоих документах.



— Программа «Повторы» просматривает документ на предмет повторения в нём одного и того же либо сходного фрагмента текста.

— Программа «Цитаты» позволяет найти в документе-первоисточнике и проверить имеющуюся в редактируемом документе цитату по некоторому набору слов, из которых она состоит. При осуществлении поиска подлежащая проверке цитата может быть приведена неточно, неполно, иногда лишь передавая смысл фразы первоисточника несколькими ключевыми словами.

— Программа «Образы» ищет в тексте документа фрагменты, где встречается набор слов, заданных пользователем, причём слова могут указываться без окончаний и в произвольной последовательности.

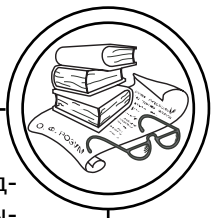
— Программа «Словари» создаёт список 1) отдельных слов или 2) двух рядом стоящих слов, из которых составлен текст исследуемого документа, с подсчетом количества случаев их употребления суммарно во всех словоформах.

Программы объединены в один пакет, поскольку все они направлены на поддержку редакторского анализа текста произведения, но имеют дело с разными по величине фрагментами текста — от абзацев и предложений до отдельных словосочетаний и слов. Во всех программах предусмотрено два варианта отображения результатов их работы: на мониторе и в протоколе. При выборе варианта «на монитор» пользователь может следить за работой программы, поскольку искомые фрагменты текста показываются в предназначенном для этого окне. При варианте «в протокол» создаётся отдельный файл с протоколом, который содержит все результаты и используется для последующего редактирования текста.

В любой из первых четырех программ в качестве критерия поиска задается минимальный процент совпадения, найденные фрагменты в тексте выделяются с указанием их адреса (страница, строка). Поскольку адреса относятся к конкретному размещению текста на страницах документа, после проведения поисковых операций не следует вносить в текст какие-либо изменения, в том числе форматирование, иначе зафиксированные адреса станут неверными и потребуется повторение операции.

В основу алгоритма поиска в первых трех программах положена идея обнаружения в тексте не отдельных слов, а сочетаний двух рядом стоящих слов из проверяемого фрагмента. Эти сочетания являются гораздо более характерным признаком

Рис. 1. Титульная панель программного пакета «Поиск и анализ фрагментов текста»



фрагмента, чем отдельные слова, и их перебор осуществляется гораздо быстрее. В то же время еще более характерные трехсловные сочетания оказываются менее эффективными при поиске, поскольку возможные трансформации проверяемого фрагмента по сравнению с оригиналом могут привести к его «неузнаваемости». Иными словами, трехсловные сочетания менее устойчивы к отличиям в сравниваемых фрагментах, чем двухсловные. Программа «Образы» работает несколько медленнее, чем предыдущие, поскольку осуществляет последовательный просмотр всего текста на предмет обнаружения элементов образа в заданном порядке и сочетании. В отличие от трех первых программ, она ориентирована не на текстуальное сходство, а на поиск фрагментов текста, содержащих максимальное количество элементов (слов) из задаваемого поискового образа, что обеспечивает смысловой поиск. Далее рассматриваются области применения и функциональные возможности каждой из подпрограмм.

Программа «Заимствования» может быть использована как для проверки легальных случаев заимствования текста со ссылками на источники и употреблением кавычек для цитат, так и для обнаружения прямого плагиата и других злоупотреблений, имеющих место в информационном поле. При наличии доступа к ранее опубликованным документам в электронной форме можно проверить редактируемое произведение сходной тематики, созданное позднее, на заимствования без ссылок.

В современном обществе самый быстрый и легкий доступ к информации предоставляют Интернет-технологии. При многих положительных возможностях они порождают, с одной стороны, избыточное количество информации вообще, а с другой — трудность выделения искомой информации. В результате поиска материалов по заданной теме в Интернете часто оказывается, что большинство из найденных страниц просто-напросто дублируют друг друга, создавая «белый шум». Скачивание реферата из Интернета и предоставление его в качестве собственной работы для студентов является вполне обыденным делом, благо в просторах Всемирной паутины имеется достаточное количество сайтов с соответствующими базами данных. Ситуация, когда ничего оригинального создавать не нужно, поскольку всё необходимое можно найти в Интернете, обесценивает контроль процесса обучения и не способствует выработке самостоятельности мышления будущего специалиста.

Естественно, всё это не могло долго оставаться без внимания со стороны педагогической общественности, на помощь которой пришли компьютерные технологии. Программы, разработанные для борьбы с плагиатом, имеются как в открытом, так и платном доступе. К ним относятся «Антиплагиат» [1], «Детектор Плагиата» [2], «Коллекция рефератов для поиска пла-

гиата» [3], «Плагиат-Информ» [4], «Copyscape» [5] и другие.

Представляемая программа «Заимствования» также нацелена на поиск похожих фрагментов в результате сравнения двух документов: первичного и производного. Особенностью предлагаемого в настоящей работе инструмента является то, что редактор или преподаватель может работать в привычной среде текстового процессора Word. Кроме того, оригинальный алгоритм поиска позволяет провести анализ текста весьма тщательно.

Таким образом, имея в своем распоряжении подобные программы и банк первоисточников в электронной форме, преподаватель сможет без особых усилий распознать степень самостоятельности той или иной работы. И если проверки на плагиат станут обязательным этапом при оценке отчетных и квалификационных работ, а при обнаружении прямого заимствования будут приниматься соответствующие меры, то у студентов появится веская мотивация выполнять всё собственными силами и при цитировании давать ссылки на источники.

Впрочем, продвинутые студенты нашли способы обмана программ, пытающихся их разоблачить. Среди этих способов замена букв русского алфавита одинаковыми по написанию латинскими, а также вставка бесцветных символов вместо пробелов. При этом взаимодействие преподавателя и студента скорее начинает напоминать антагонистическую компьютерную игру, чем процесс обучения предмету. Следует заметить, что обратная замена бесцветных символов на пробелы, обнаружение и исправление латиницы в русских словах — несложные задачи для компьютерной обработки и соответствующий блок «профилактически» включен в программу «Заимствования». Поскольку программа направлена в большей степени на подготовку произведения к изданию, чем на проверку студенческих работ, в ней предусмотрен режим «Редактирование» для удаления заимствований.

Помимо заимствований из чужих произведений без кавычек и ссылок авторы нередко допускают появление повторов в своём собственном тексте. Это происходит, когда при редактировании какой-то фрагмент переносится в другое место и по забывчивости не удаляется из исходного либо когда автор вновь обращается к подобной тематике, воспроизводя сходные фразы. Такие повторы не всегда могут быть замечены редактором, зато легко обнаруживаются программными средствами. Разработанную для этой цели программу «Повторы» рекомендуется использовать как автору, так и редактору. В программе предусмотрено создание протокола анализа текста. При обсуждении недочётов произведения с автором редактор имеет возможность аргументировать свои замечания с помощью этого протокола. Следует также учитывать, что в учебно-методических пособиях повторы могут помещаться сознательно, если это

дидактически оправдано. Решение об удалении или сохранении повтора в таком случае остаётся за автором.

В работе с программами «Заимствования» и «Повторы» имеется ряд сходных возможностей. Во-первых, перед поиском задается минимальный процент совпадения фрагментов, с которого будут фиксироваться результаты поиска, а также минимальная и максимальная длина фрагментов, которые должны быть найдены (рис. 2).

Во-вторых, созданы удобные условия для редактирования документов. В программе «Заимствования» оба документа открываются рядом и совпавшие или похожие фрагменты выделяются цветом. Можно просматривать эти фрагменты друг за другом и вносить правку в редактируемый документ, причем во избежание случайной ошибки документ с первоисточником заблокирован от изменений. В программе «Повторы» экран разделён по горизонтали на две части, в которых отображаются страницы исследуемого документа, содержащие повторы. Совпавшие фрагменты, выделенные цветом, отображаются одновременно друг над другом в обоих окнах (рис. 3). Это позволяет редактировать документ, изменяя один или оба фрагмента.

Программа «Цитаты» предназначена для поиска в авторских текстах фраз по нескольким словам, воспроизводящим смысл высказывания, но не гарантирующим его точности. Необходимость такой программы вызвана тем, что многие авторы цитируют тексты по памяти и смыслу сказанного. В задачи редактора входит проверка всего фактического материала, приводимого автором, а значит, и поиск по первоисточникам. В редакциях, как правило, имеется широкий набор справочной литературы: словарей, энциклопедий и т.д. — с которой приходится работать редактору. Однако на поиски необходимого материала ему приходится тратить много времени, большая часть которого уходит бесполезно. Естественно, что для использования данной про-

Рис. 2. Вкладка «Параметры» программы «Заимствования»

граммы редактору должны быть предоставлены необходимые источники в электронном виде, поэтому требуется составление полнотекстовых баз источников по тематике редакционного отдела.

В программе «Цитаты» также предусмотрена возможность определения общего числа и объёма цитат в данном документе, если цитаты приведены в кавычках. Для этого нужно указать примененный в документе тип кавычек, ограничивающих цитаты, а также тип кавычек внутри цитат. Чтобы программа не включала в число цитат слова в кавычках, к ним не относящиеся, следует оценить и условно установить минимальную длину учитываемой цитаты. Результаты анализа текста на процент наличия в нём цитат могут быть занесены в протокол, где приводится в табличной форме список найденных цитат в порядке возрастания длины.

Дополнительные по сравнению со стандартными, «интеллектуальные» возможности компьютерных технологий поиска по тексту оказываются намного эффективней. Поэтому в ситуации, когда нужно получить информацию о том, имеется ли в незнакомом тексте описание какого-либо конкретного вопроса, следует применять программу «Образы». Целесообразно искать фрагмент текста по набору слов, характерных для искомого фрагмента, другими словами, по некоему «поисковому образу». Под терми-

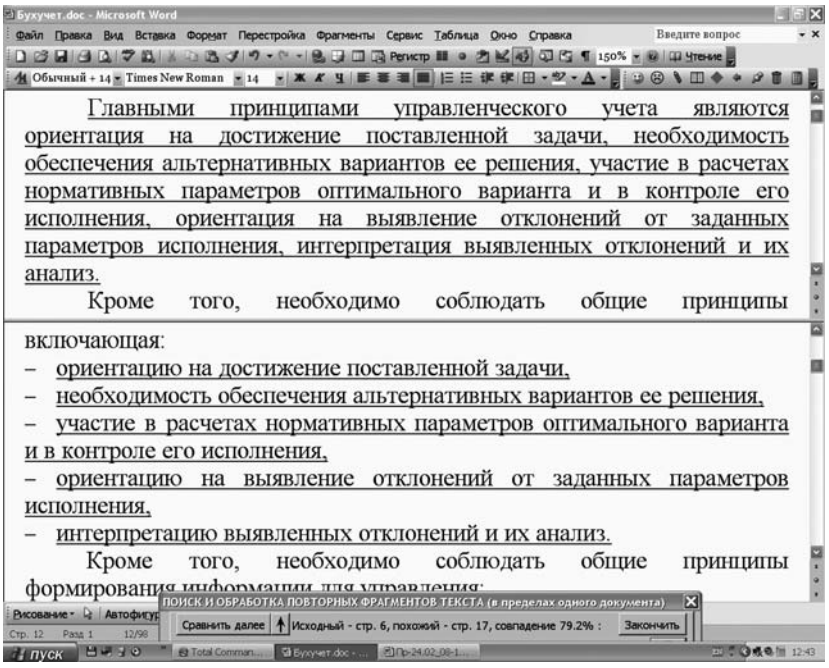


Рис. 3. Режим «Редактирование» в программе «Повторы»: выделение цветом здесь заменено подчеркиванием

ном «поисковый образ» понимается такой набор слов, который достаточно полно ограничивает смысловое наполнение понятия. Слово будет найдено не само по себе, а в нужном контексте.

Для успешного поиска в программах «Цитаты» и «Образы» пользователю предоставляется возможность задать ряд параметров (рис. 4).

1. Определяется минимальный процент совпадения поискового образа со словами в анализируемых фрагментах текста.

2. Предоставляется выбор режима поиска: с остановками, без остановок, автоматически. При режимах с остановками/без остановок промежуточные результаты выводятся/не выводятся в окне программы. При автоматическом режиме критерий совпадения игнорируется, программа отбирает фрагменты с наивысшим процентом совпадения и их отображает на мониторе. При завышенном критерии совпадения поискового образа со словами в анализируемых фрагментах текста результат поиска может быть отрицательным.

3. Указываются типы фрагментов, на которые должен быть разбит текст для анализа. Это могут быть абзац, предложение или фрагмент с заданным числом символов.

4. Задаётся минимальная длина слова, которое будет учитываться программой при поиске. Если в качестве образа взята целая фраза, то из нее при поиске автоматически будут удалены знаки препинания и слова с длиной менее заданной минимальной длины.

5. В программе «Образы» имеется опция, предусматривающая, чтобы были найдены

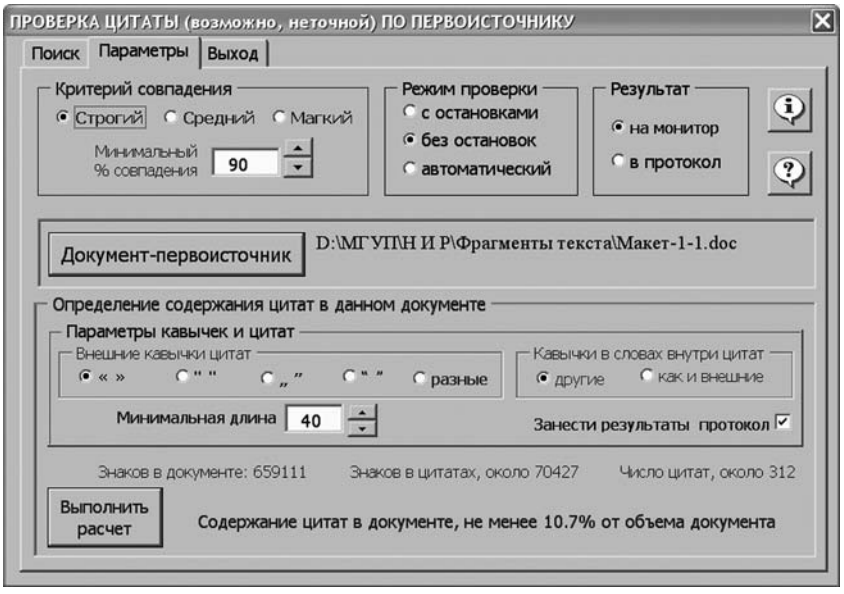
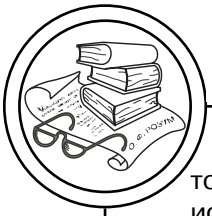


Рис. 4. Вкладка «Параметры» программы «Цитаты»



только те фрагменты текста, где искомые слова располагаются в той же последовательности, что и в поисковом образе.

При вводе слов, составляющих поисковый образ искомого фрагмента, их следует приводить без окончаний, чтобы предусмотреть все словоформы; ввод может осуществляться в произвольном порядке. Поскольку поиск в программе «Образы» должен быть значительно более гибким, чем в программе «Цитаты», там при составлении поискового образа можно вместо конкретного слова вводить его основу без приставок и флексий. Эта основа будет входить как часть в слова искомых фрагментов. Можно также использовать уточняющие шаблонные символы, которые ниже указаны в кавычках, тогда как в программе их следует вводить без них (рис. 5).

1. Символ «<», поставленный в начале слова из поискового образа, означает, что начало соответствующего слова в найденных фрагментах должно совпадать с началом слова из образа. Так, если задать поисковый образ «<описание», то фрагменты, где употреблены такие слова, как, например, «книгоописание», найдены уже не будут.

2. Если после основы слова из поискового образа поставить знак «[» (левая квадратная скобка) и за ней ввести в подбор буквы окончания слова, то поиск слова будет успешным только для тех случаев, когда в тексте это слово оканчивается на любую букву или группу из букв, стоящих после скобки. К примеру, в тексте нужно найти слово «книга» только в именительном падеже. Тогда следует составить образ: «книг[а». Чтобы включить в поиск слово с нулевым окончанием, в группу букв после символа «[» следует добавить символ «*» («звездочка»).

3. Знак «_» (подчерк), поставленный между двумя и более словами образа, приводит к тому, что в найденных фрагментах эти слова будут располагаться в указанном порядке. Для первого из этих слов также возможно одновременное употребление знаков «<» и «[», последующие слова их содержать не должны.

4. Знак «/» (слэш) между двумя и более словами делает их поиск альтернативным, то есть он будет считаться успешным при наличии хотя бы одного слова из группы. В поисковом образе допускается одна такая группа альтернативных слов. Эта опция удобна в тех случаях, когда понятие имеет несколько вариативных наименований. Примеры таких образов:

учебн/обучающ издани;

электронн/цифров/мультимедийн/виртуальн книг.

5. Если два слова и более заключены в круглые скобки «(слово1 слово2 …)», то эта группа рассматривается программой как одно слово; для первого слова в группе применим знак «<», для последнего — «[». Остальные слова внутри группы будут найдены точно в соответствии с образом. Эта опция позволяет включать в поисковый образ устойчивые выражения.

Кроме самих найденных фрагментов текста пользователю на мониторе или в протоколе (в соответствии с выбранным режимом отображения) предоставляется следующая информация: адрес фрагмента в документе (страница и строка/сноска), процент слов в найденном фрагменте от числа слов в поисковом образе, параметры поиска, элемент образа, по которому произошло первичное обнаружение фрагмента. Следовательно, программа «Образы» является мощным универсальным поисковым инструментом, значительно расширяющим стандартные возможности текстового процессора Word, даже с учётом его дополнительных функций поиска с применением подстановочных знаков [6].

В теории и практике редакторской деятельности иногда возникает потребность оценить состав слов, употребляемых в произведении, вычленить устойчивые словосочетания для их критического рассмотрения или для подготовки рубрик предметного указателя. Этой цели служит входящая в пакет программа «Словари», которая предназначена для анализа словарного состава редактируемого документа и просмотра контекста отдельных слов с возможной правкой. Программа формирует из текста документа список (словарь) слов или двухсловных сочетаний с указанием частоты их использования. Задача учета множества словоформ, характерных для русского языка, является весьма трудной для такой относительно небольшой программы, как «Словари». Однако удалось подобрать алгоритм и реализовать механизм морфологического анализа слов на предмет вычле-

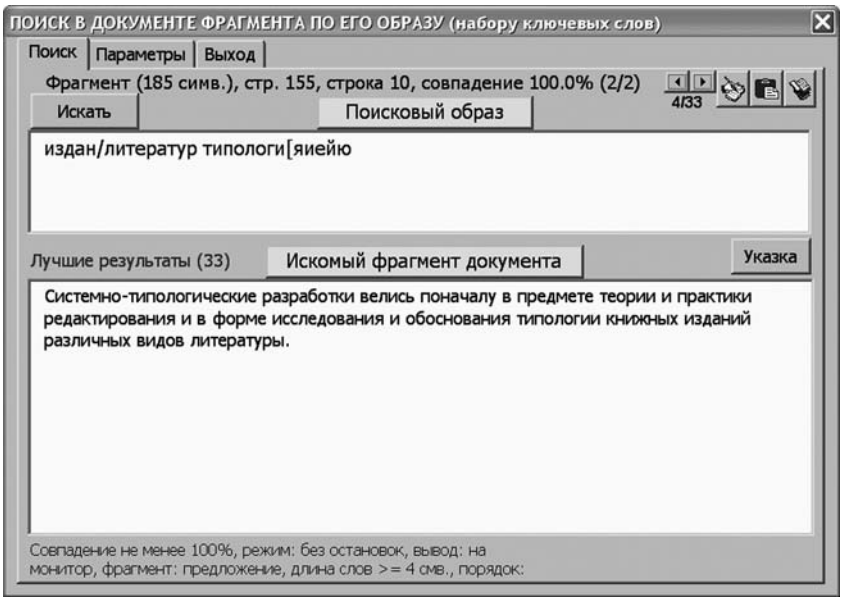


Рис. 5. Вкладка «Поиск» программы «Образы»

нения окончаний, который с эффективностью в 80 % позволяет проводить группировку слов по общей основе. Следует признать, что в данной, первой, версии программы этот анализ выполняется еще не всегда корректно, что особенно относится к окончаниям глаголов, причастий и деепричастий.

Записи в генерируемом программой словаре включают основу с указанием в прямых скобках встречающихся окончаний, в том числе и нулевых (знак «*»). Заключает запись суммарное число слов с вычлененной основой, например:

аспект|*, а, ам, ами, ах, е, ов, ы| 56

компонент|*, а, ам, ами, ах, ов, ом, ы| 43

Двухсловные сочетания образуются из слов, между которыми стоит пробел. Соседние слова, разделённые знаками препинания, кавычками и т.д., в список не заносятся. Слова, входящие в словарь двухсловных сочетаний, приводятся в два уровня, например:

издательск|ая, ий, их, ого, ое, ой, ому, ую| 51

— дел|а/о/у| 15

— практик|е/и/у| 8

— репертуар|*/а| 5

— задач|ей/и| 4

— книготоргов|ая/ого| 3

— оригинал|*/а| 3

— работ|а/е| 2

— способ|*/ов| 2

Пользователю предоставлена возможность задавать дополнительные условия формирования словарей (рис. 6).

1. Может быть произведена дискриминация слов по их минимальной длине. Например, в качестве минимальной можно указать длину в 4 символа. Тогда в созданном словаре не окажется слов, типа «и», «да», «но», «как», «что» и тому подобных.

2. В список могут быть не включены цифры, а также слова, набранные латиницей либо кириллицей.

3. Дискриминация также может производиться и по частям речи. В последнем случае программа учитывает файлы, содержащие отдельные списки с наиболее употребляемыми гла-

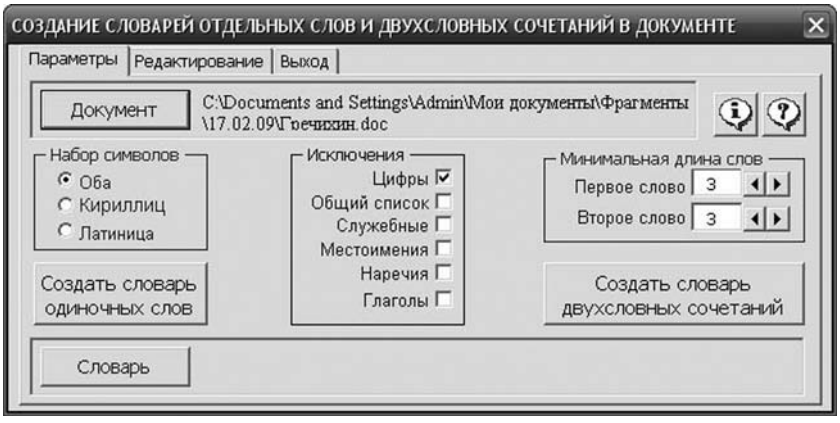
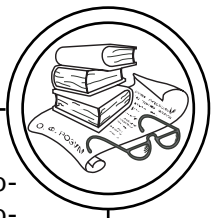


Рис. 6. Вкладка «Параметры» программы «Словари»



голами, местоимениями, наречиями, служебными словами в различных словоформах. Кроме того, имеется файл с общим списком исключаемых слов. Указанные файлы (в формате Word) могут редактироваться пользователем.

Возможность формирования словаря по частям речи позволяет анализировать текст на предмет слишком частого употребления тех или иных слов.

Словари одиночных слов и двухсловных сочетаний могут быть упорядочены как в алфавитном порядке, так и по частоте употребления слов в тексте. Последний вариант демонстрирует предметную область анализируемого текста. Это быстрый и удобный способ составления списка ключевых слов, который необходим для поиска текста по базам данных. Наиболее часто встречающиеся слова можно вносить в указатели и глоссарии, так что программа может быть полезна при составлении списка терминов для предметных указателей.

Учёт этой области применения словарей привёл к необходимости разработки средств их редактирования. Так, программа может удалить из списка слова, встретившиеся в тексте менее заданного пользователем количества раз. Причём из двухсловных словарей эта дискриминация осуществляется отдельно для первых и вторых слов. Словарь двухсловных сочетаний может быть преобразован в один уровень, что позволяет, после упорядочения по частоте, автоматически выделить наиболее устойчивые двухсловные выражения, например:

книготорговых|ого, ый, ым, ых| ассортимент|*/а/ом/ов| 029

книготорговых|ое, ом, ым, ых| дел|*/а/е/о/ом| 019

книготорговых|ое, ые| знан|ия/ие| 016

книготорговых|ая, ой| деятельност|и/ь| 009

книготорговых|ые, ых| процесс|ы/ах/ов| 007

книготорговых|ая, ые, ых| библиограф|ии/ия/ий| 005

Созданию словарей, пригодных для формирования в дальнейшем рубрик предметных указателей, служит также опция исключения частей речи, поскольку глаголы, местоимения, служебные слова в состав указателей, как правило, не входят.

Представленный пакет программ «Поиск и анализ фрагментов текста» призван устранить значительную часть рутинной деятельности редактора и тем самым освободить его время для творческой работы с текстом произведения. Целью дальнейших разработок в этой области должно явиться распространение компьютерных технологий на все формализуемые операции, необходимые при подготовке авторского произведения к изданию.

1. Антиплагиат [Электронный ресурс]. — Электрон. дан. и прогр. — М. : ЗАО «Анти-Плагиат», 2005. — Режим доступа : http://www.antiplagiat.ru. — Загл. с домашней страницы Интернета. 2. Детектор плагиата [Электронный ресурс]. — Электрон. дан. и прогр. — М., 2007. — Режим доступа : http://www.detector-plagiata.ru. — Загл. с

домашней страницы Интернета. 3. Коллекция рефератов для поиска плагиата [Электронный ресурс]. — Электрон. дан. и прогр. — М. : Клио Софт. — Режим доступа : http://www.2balla.ru. — Загл. с домашней страницы Интернета. 4. Плагиат-Информ [Электронный ресурс]. — Электрон. дан. — М. : СофтИнформ, 2007. — Режим доступа : http://www.plagiatinform.ru. — Загл. с домашней страницы Интернета. 5. Copyscape [Электронный ресурс]. — Электрон. дан. и прогр. — М. : Indigo Stream Technologies Ltd. — Режим доступа : http://www.copyscape.com. — Загл. с домашней страницы Интернета. — Текст на экране англ. 6. Клименко Б. Н. Microsoft Word : Комфортная работа с помощью макросов. Самоучитель / Клименко Б. Н., Розенберг М. М. — СПб. : БХВ-Петербург, 2006. — 496 с.

Надійшла до редакції 29.06.10