ГОЛОВКО НИКОЛАЙ ВЯЧЕСЛАВОВИЧ

ОЦЕНКА СЕМАНТИЧЕСКОГО ПОТЕНЦИАЛА ТЕКСТОВ В АНАЛИТИЧЕСКИХ СИСТЕМАХ

Монография

2-е издание, стереотипное

Ставрополь
2014

УДК 81’322.2
ББК 83я73+32.965
Г 61

Рецензент:
Гусаренко Сергей Викторович
доктор филологических наук, профессор, декан факультета филологии, журналистики и межкультурной коммуникации ФГАОУ ВПО «Северо-Кавказский федеральный университет»

Головко, Н.В.
Г 61
Оценка семантического потенциала текстов в аналитических системах: монография. – Изд. 2-е, стереотипное. – Ставрополь: ООО ИД «ТЭСЭРА», 2014. – 204 с.
ISBN

В исследовании рассматривается один из перспективных способов практического использования данных о семантическом потенциале текста на русском языке – специфическом интегральном показателе, отражающем меру его смысловой неопределенности с точки зрения автоматизированного анализатора. Оценивание данного показателя посредством формально-семантического анализа многозначной лексики применено в интересах разработки минимизированного алгоритма автоматической классификации русскоязычных текстов по признаку их функционально-стилевой отнесенности, что, в свою очередь, позволяет в перспективе обеспечить оптимизацию производительности и эффективности основных разновидностей автоматизированных систем обработки текстов - средств понимания естественного языка, поисковых роботов и каталогизаторов, а также систем машинного перевода.

УДК 81’322.2
ББК 83я73+32.965

I S B N

© Головко Н.В., 2014
© Издательский дом «ТЭСЭРА», 2014

ВВЕДЕНИЕ

В конце XX века многие исследователи проявляли существенный интерес к разработкам в области искусственного интеллекта вообще и к вопросам автоматизированной обработки естественных языков в частности. Этот интерес сохраняется и поныне, регулярно находя выражение в инновационных алгоритмах, программных продуктах и программно-аппаратных решениях, призванных обеспечить машинное «понимание» тех или иных текстов.

Интенсивное развитие информационных технологий позволяет накапливать значительные объемы разнообразных данных, что автоматически ставит вопросы о надлежащей обработке упомянутых сведений, оперативном поиске требуемой информации, а также предельном упрощении человеко-машинной коммуникации вкупе с автоматизацией действий, на выполнение которых затрачиваются существенные временные ресурсы. Все это формирует предпосылки для развития и совершенствования автоматизированных систем обработки текстов (АСОТ).

Хотя первичные исследования, послужившие базой для дальнейших разработок, датируются концом XIX – началом XX века (в связи с этим могут быть упомянуты, к примеру, концепции Дж. Милля (1914), представляющие собой образцы раннего приближения к теории фреймов), период активного изучения подобных вопросов приходится главным образом на 80-е годы минувшего века. В это время наблюдается рост интереса советских и зарубежных ученых к лингвистическим возможностям вычислительных машин. Теоретические основания данных исследований закладывались Р. Шенком (1977), М. Минским (1978), Ч. Филлмором (1985), Т. Виноградом (1985), В.А. Звегинцевым (1985) и другими известными специалистами. В СССР регулярно публиковались аналитические обзоры и сборники статей по компьютерной

лингвистике, подготовленные С.М. Шевенко (1985), А.Н. Барановым (1987), Б.Ю. Городецким (1989), что свидетельствовало об актуальности рассматриваемой проблемы с точки зрения советских ученых. При этом вопросы машинного «понимания» естественных языков неразрывно увязывались с разработками в области лексической семантики и семантических языков (Б.Н. Плотников (1981), А.Ф. Лосев (1982), Д.Н. Шмелев (1973), Ю.Д. Апресян (1974), Н.Д. Арутюнова (1988), Е.С. Кубрякова (1992)). Велись исследования и в сфере формализации языковых образцов (А.М. Шахмайкин (1986)).

После того, как Советский Союз прекратил свое существование, объективные обстоятельства, связанные с политической обстановкой в государстве, оказали негативное воздействие на развитие отечественной компьютерной лингвистики, в результате чего активные разработки в области автоматизированных систем обработки текстов были практически полностью прекращены. Те из них, которые нашли применение в коммерческом секторе, как и более актуальные разработки зарубежных ученых, легли в основу разнообразных продуктов и решений, обеспечивающих их производителям конкурентное преимущество перед прочими игроками рынка; таким образом, упомянутые научные достижения обрели статус информации, охраняемой коммерческой тайной, что автоматически обусловило низкую степень их доступности для рядового исследователя.

Тем не менее, российские и зарубежные специалисты активно рассматривают вопросы формально-семантического анализа текстов (Леонтьева Н.Н. (2002); P. Cimiano, S. Staab, J. Tane (2003); D. Ferrucci (2004); И.В. Азарова, Е.А. Овчинникова (2005) и др.), автоматической классификации данных для нужд информационно-поисковых систем (Панков И.П., Захаров В.П (1996); П.И. Браславский (1999)), формализации языковых сообщений (В.А. Фомичев (2002); A. Mehler, U. Waltinger, A. Wegner (2007)). Кроме того, разработки зарубежных ученых манифестируются в создаваемых ими аналитических системах (например, о текущем состоянии ведомых американскими лингвистами и инженерами работ в данной области можно судить по особенностям вычислительного кластера IBM Watson (2010)). Из наиболее содержательно близких к на-

стоящему исследованию работ надлежит упомянуть диссертацию О.Г. Шевелева «Разработка и исследование алгоритмов сравнения стилей текстовых произведений» (2006).

Необходимо отметить, что в настоящее время существует потребность в оптимизации систем автоматизированного анализа текстов на естественных языках, которая связана с ростом объемов информации и необходимостью ее надлежащей классификации. Несмотря на активное исследование связанных вопросов и существенный объем накопленных за все время их изучения данных, современные АСОТ по-прежнему нуждаются не только в совершенствовании, но и в дальнейшем повышении скорости и производительности их работы. Кроме того, ряд исследователей отмечает, что собственно лингвистическое обеспечение упомянутых систем нередко характеризуется недостаточной степенью унифицированности, системности и разработанности, что приводит к не вполне верному пониманию сущности языковых явлений и, как следствие, к выбору неоптимальных аналитических критериев. Это приводит нас к заключению, что на данный момент востребованными являются исследования, в которых рассматривались бы потенциальные пути улучшения базовых показателей типичных АСОТ, основанные на достаточно надежном теоретико-лингвистическом фундаменте и обеспечивающие более корректное отражение специфики языковой системы.

Привлекая формально-семантический анализ многозначной лексики для исследования текстовых образцов, мы рассчитываем выйти тем самым на формирование у машины определенного эмулированного представления о семантике и о контексте. Наш анализ показывает, что разработчики современных автоматизированных классификаторов и обработчиков информации полагаются главным образом на те формальные признаки исследуемых текстов, которые не проникают глубже плана выражения и не имеют какой-либо опосредованной или непосредственной связи с планом содержания, хотя актуальность и необходимость обращения к значению и смыслу языковых единиц не раз обосновывалась в упомянутых нами выше исследованиях. В тех же случаях, когда учеными предпринимаются попытки выхода на план содер-

жания посредством, к примеру, изучения генерализованного грамматического значения, для подобных исследований привлекаются громоздкие методы и приемы, для поддержания работоспособности которых требуются сложные алгоритмы и объемные базы данных, что, в свою очередь, негативно сказывается на производительности АСОТ. Сложившееся положение требует разработки принципиально новых аналитических алгоритмов, способных претендовать на адекватное отражение семантических и контекстуальных особенностей предъявляемых им текстовых образцов и в то же время отличающихся высокой скоростью функционирования и экономичностью в отношении системных ресурсов.

Настоящее исследование фокусируется на лексико-семантических аспектах значения текста на естественном языке, в частности – семантическом потенциале текста как интегральном показателе многозначности входящих в его состав лексических единиц, определяемом средствами формально-семантического анализа. Подробному рассмотрению подвергаются системные внутриязыковые взаимоотношения, существующие, с одной стороны, между лексической многозначностью и полисемантичностью текста, и, с другой стороны, между неоднозначностью текста и его функционально-стилевой отнесенностью. Данные взаимоотношения выражаются в специфических характеристиках естественноязыковых текстов, которые, в свою очередь, манифестируются в машиночитаемых формальных маркерах и могут быть подвергнуты автоматизированному анализу.

В рамках работы будет рассмотрено одно из потенциальных направлений практического использования данных о семантическом потенциале естественноязыковых текстов и методов его оценивания, а именно – оптимизация автоматизированных систем обработки текстов посредством функционально-стилистической классификации исследуемых образцов. Соответственно, в ходе исследования мы намерены выявить, описать и систематизировать те средства формально-семантического анализа многозначной лексики, использование которых позволяет усовершенствовать лингвистическое обеспечение автоматизированных систем обработки текстов и оптимизировать их. Для этого надлежит определить харак-

теристики естественноязыковых текстов, обработка которых составляет наиболее эффективные потенциальные пути оптимизации АСОТ, после чего на примере русского языка сформировать и описать минимальный набор машиночитаемых формальных маркеров, позволяющих реализовать выбранные пути оптимизации АСОТ посредством классификации естественноязыковых текстов на основании упомянутых характеристик; изучить классификационные возможности выявленных маркеров, в том числе показателей потенциальной полисемантичности, находимых в результате формально-семантического анализа многозначной лексики, и разработать алгоритм анализа и классификации текстов по выявленным формальным маркерам, построить демонстрационный образец программного обеспечения, реализующего разработанный алгоритм, и провести испытания, позволяющие судить о степени эффективности построенного программного обеспечения, а также рассмотреть перспективы повышения эффективности построенного программного обеспечения по результатам проведенных испытаний.

Нами выдвинута следующая гипотеза: основанием для достоверной и эффективной функционально-стилистической классификации текстов, которая выступает в качестве источника оптимизации существующих и перспективных АСОТ, может служить упрощенный аналитический алгоритм, принцип работы которого базируется на определении и сопоставлении двух минимально необходимых диагностических показателей: средней длины слова и потенциальной полисемантичности текста, выявляемой посредством обработки его лексического массива. Потенциальная полисемантичность текста при этом рассматривается как мера его неопределенности (т.е. энтропия), которая образуется в результате суммирования неопределенностей лексических единиц, входящих в состав исследуемого текста.

В настоящем исследовании систематизированы, обработаны и сведены в единую дефиницию разрозненные определения формально-семантического анализа, представлен способ оценивания потенциальной полисемантичности текстов на основе расчета и сопоставления массовых долей специфических словарных групп (данное терминологическое выра-

жение применяется в рамках исследования для обозначения всякого конститутивного фрагмента словника лексикографического издания с алфавитным принципом сортировки, объединяющего заголовочные слова по признаку инициальной графемы), сформирован и описан минимальный набор машиночитаемых формальных маркеров функционально-стилевой принадлежности русскоязычных текстов, а также предложено обоснование для упрощенного исследовательского алгоритма, возможности которого обладают значительным потенциалом с точки зрения оптимизации АСОТ. Полученные результаты, разработанные алгоритмы и построенное на их основании программное обеспечение после необходимой адаптации могут быть применены для улучшения производительности и качества функционирования основных типов АСОТ – поисковых роботов и каталогизаторов, комплексов машинного перевода, а также средств человеко-машинного взаимодействия.

По нашему мнению, формально-семантический анализ многозначной лексики, выступающий эффективным средством автоматического классифицирования текстов на естественных языках по признаку их функционально-стилевой принадлежности, представляет собой наиболее эффективное направление оптимизации автоматизированных систем обработки естественноязыковых текстов в рамках комплексного подхода к формированию их лингвистического обеспечения.

Исследования по заявленной теме выполнялись в рамках реализации Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009-2013 годы (НК-523П) и были обусловлены Государственным контрактом № П122 от 13 апреля 2010 г. на выполнение поисковых научно-исследовательских работ для государственных нужд.

1. ФОРМАЛЬНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ МНОГОЗНАЧНОЙ ЛЕКСИКИ И АКТУАЛЬНЫЕ ПРОБЛЕМЫ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ОБРАБОТКИ ТЕКСТОВ (АСОТ)

1.1. СТРУКТУРНО-СЕМАНТИЧЕСКИЕ АСПЕКТЫ ЛЕКСИЧЕСКОГО ЗНАЧЕНИЯ СЛОВА

Вопросы, связанные с лексическим значением слова в целом и явлением лексической многозначности в частности, находились в сфере внимания многих известных отечественных и зарубежных лингвистов. Общетеоретические положения подобного рода разрабатывались А.Ф. Лосевым (1982), Ю.Д. Апресяном (1995), Е.С. Кубряковой (1992), Д.Н. Шмелевым (1973), Н.Д. Арутюновой (1988), А.Е. Кибриком (1992), Е.В. Падучевой (1985), Р. Якобсоном (1983), Дж. Лайонзом (1978), Е. Куриловичем (1962), А. Вежбицкой (1967) и др. Некоторые из вышеупомянутых положений, детальный обзор которых мы полагаем уместным представить в данной главе, способны выступить в качестве надежной основы для дальнейших исследований по лингвистическому обеспечению АСОТ и формально-семантическому анализу многозначной лексики.

В соответствии с концепциями А.Ф. Лосева [58], язык неизбежно предполагает смысловую подвижность и непрерывность. Если язык рассматривается, начиная с минимального фонетического значения и заканчивая полноценными структурами данных значений, то исследователю необходимо исходить из принципа вариативности; кроме того, поскольку речь идет о смысловом знаке, указанный принцип может привести к неожиданным семантическим комплексам, которые смогут быть соотнесены ни с изначальными звуками, ни с их разнообразными сочетаниями. В процессе изучения языкового контекста следует отметить разнообразие значений каждого отдельного слова; это значение может принимать различные формы, фактически – от нуля до бесконечности. В отдельных

случаях слово или даже отдельная морфема может иметь десятки возможных значений. В иных случаях представляется возможным говорить о бесконечном разнообразии допустимых значений, так что возникают затруднения при перечислении всех возможных семантических оттенков данного слова. В третьем случае вероятна полная потеря значения и приравнивание его к нулю.

Так, по мнению А.Ф. Лосева, объяснение специфических сем отдельных слов (напр., розоватый, мрачноватый и т.п.) возможно исключительно через привлечение некоторых иных элементов, выходящих за их пределы. Данное объяснение может быть осуществлено с привлечением словесного контекста рассматриваемого слова в определенной группе аналогичных слов, в определенной разновидности языка, где это слово употребляется в той или иной обстановке и, вероятно, в значительном историко-культурном контексте. Таким образом, основополагающим базовым принципом, определяющим употребление и значение того или иного знака, является контекст, а для языка — как языковой, так и внеязыковой контекст в наиболее широком смысле этого слова. Соответственно вышесказанному представляется возможным считать т.н. аксиомой знака тезис о его контекстной значимости: любой знак получает свою полноценную значимость только в контексте других знаков.

В соответствии с позицией А.Ф. Лосева, вместе с тем данная аксиома контекста требует рассмотрения вопросов, касающихся определения того, что такое знак. Представляется очевидным, что никакой знак не может быть представлен в абсолютном смысловом вакууме и что любое значение от наименее до наиболее значительного он может получать исключительно в том или ином смысловом окружении, т.е. контексте. Данное утверждение предполагает значительное количество очевидных аксиом того или иного рода, которые могут не быть сформулированы, но предполагаться к существованию каждым лингвистом. Из возникающего множества вопросов может быть выбрана проблема, связанная с явлением значения, - ввиду самого термина, имеющего определенное распространение в лингвистике.

А.Ф. Лосев уверен, что понятия знака и значения являют-

ся в достаточной степени изученными, и очевидно, что любой знак имеет значение. Вместе с тем требует рассмотрения вопрос о том, что есть значение. Нет сомнений в том, что знак является знаком исключительно в тех случаях, когда характеризуется тем или иным значением. Равнозначно существует возможность утверждать, что означаемое также характеризуется определенным значением. Само действие обозначения также может рассматриваться в качестве носителя значения, связанного с соответствующими знаком, значением и обозначаемым.

Из вышесказанного следует, что тот или иной знак обозначает что-либо, в силу чего может характеризоваться определенной совокупностью значений. Исследователю представляется небезынтересным случай наличия у знака бесконечного количества значений. При первичном рассмотрении понятие бесконечного знака может казаться невозможным или бессмысленным; вместе с тем очевидным образом не требует доказательств тот факт, что существует неограниченное количество контекстов, характеризующих любой текст. В подавляющем большинстве случаев существует возможность формирования дополнительных контекстов, не кодифицированных посредством словарей или грамматических пособий. А.Ф. Лосеву представляется в указанном контексте правомерным использование известного термина «символ»; в то же время, по его мнению, необходимо отвлечься от существующего в науке множества толкований приведенного термина, поскольку в противном случае возникала бы необходимость в самостоятельном формировании терминологической базы каждым отдельным исследователем. Описанное выше утверждение исходит из представления о том, что неперечислимость совокупности значений языкового знака необходимо обозначить определенным термином, и в случае, если термин «символ» не является подходящим, он может быть заменен иным. В текущем же состоянии терминологической базы можно считать допустимым использование термина «символ», из чего непосредственно вытекает аксиома символа: всякий знак может иметь бесконечное количество значений, т. е. быть символом. Сообразно, отталкиваясь от концепции, представленной А.Ф. Лосевым, можно характеризовать языковой символ как

языковой знак, имеющий бесконечное количество значений.

Подводя итог, необходимо отметить, что в концепции А.Ф. Лосева рассматривается существенно важный вопрос о природе значения языкового знака. Фактически А.Ф. Лосев делает акцент на прагматическом аспекте языковой единицы, постулируя тем самым контекстную обусловленность ее значения. При этом в рамках концепции не предпринимается попытка представить контекст в качестве средства дискретизации значения языкового символа, т.е. механизма, позволяющего выделить из непрерывной совокупности значений одно конкретное, характеризующее символ в определенный текстовый момент; напротив, само утверждение о непрерывности суммы значений языкового символа обосновывается ученым через привлечение понятия множественности контекстов. Таким образом, ставится вопрос не о дискретизации совокупности значений символа, т.е. выделении конкретно-дискретного значения такового, но о дискретизации совокупности контекстов, характеризующих определенный текст; вследствие подобной операции значение символа либо будет непосредственным образом продиктовано выбранным контекстом, либо возникнет на пересечении нескольких равноправных контекстов, что позволяет по-новому взглянуть на вопросы «машинного понимания» текстов на естественном языке.

Говоря о значении языкового знака и символа, А.Ф. Лосев, тем не менее, никак не обозначает в разработанной им аксиоматике языкового знака, каким именно является это значение. Исходя из аргументации, приводимой ученым, представляется вероятным, что при формулировании упомянутых аксиом понятия языкового знака и символа ассоциировались с понятием слова, и под значением таковых подразумевалось лексическое значение.

По мнению известного отечественного лексиколога Ю.Д. Апресяна [2], семантика языкового знака отражает наивное понятие о вещи, свойстве, событии и т.д. Необходимо отметить, что наивная картина мира, отражающая материальный и духовный опыт народа, может расходиться с чисто логической (научной) картиной, которая является общей для людей, принадлежащих к различным языковым сообществам. Сообразно, при описании лексического значения слова пред-

ставляется необходимым осуществление анализа наивной картины мира и отражения ее в системе научных толкований. Кроме того, картины мира, извлекаемые из значений слов различных языков, могут в деталях отличаться друг от друга, в то время как научная картина мира не зависит от языка. Описанные положения, по мнению ученого, имеют принципиальный смысл в вопросах описания лексических значений слов.

Ю.Д. Апресян также обращает внимание на необходимость расчленения семантики языкового знака. По его свидетельству, семантику знака принято рассматривать на двух уровнях – денотативном (референционном) и сигнификативном. Денотат знака при этом определяется как класс обозначаемых им фактов, в то время как сигнификат может быть представлен в качестве общих признаков всех фактов, относящихся к этому классу. Соответственно, при денотативном тождестве возможны различия сигнификатов; не предпринимая попыток перечисления всех возможных типов сигнификативных различий, Ю.Д. Апресян отмечает лишь один из таковых – различия в логических акцентах (активная и пассивная форма глагола, различные типы лексических конверсивов).

Говоря о синтактике, исследователь акцентирует внимание на различии между лексическим значением и сочетаемостью (дистрибуцией) слова. Лексическая сочетаемость определяется как информация о том, каким должно быть слово B или класс слов Bn, с которым(и) синтаксически связано слово A; семантическая сочетаемость – как информация о том, какими семантическими признаками должно обладать синтаксически связанное с A слово B [2, с. 61]. Ю.Д. Апресян также поясняет, что при возникновении трудностей в разграничении лексического значения и семантической сочетаемости допустим ряд решений: истолкование семантической информации как особенности семантики слова, описание семантической особенности слова как особенности его сочетаемости, неединственность семантических описаний. В то же время в ряде случаев вопрос о границе между синтактикой и семантикой остается, по мнению ученого, открытым.

Комментируя вопросы, имеющие непосредственное отношение к прагматике знака, Ю.Д. Апресян считает необходимым начать с семантических ассоциаций, или конно-

таций – элементов прагматики, отражающих культурные представления и традиции, которые связаны с тем или иным словом, господствующую в обществе практику использования тех или иных предметов и иные экстралингвистические факторы. Данные признаки, по мнению исследователя, не входят в семантику слова непосредственно, однако способны представлять для нее определенный интерес, т.к. на их основе слово метафоризуется, включается в сравнения, участвует в различных языковых процессах. Вследствие этого, отмечает Ю.Д. Апресян, признак, который в одном лексическом значении является ассоциативным и прагматическим, в другом значении выступает в качестве существенного и семантического. Другим важным аспектом прагматики является, по мнению ученого, оценка говорящим вероятности ситуации, имплицитная ссылка на наблюдателя и т.п. компоненты, образующие в своей совокупности модальную рамку значения. На основании изложенного выше Ю.Д. Апресян делает заключение: под лексическим значением слова понимается семантика знака и часть его прагматики, которая включается в модальную рамку толкования.

Для исследователя вполне очевиден тот факт, что значение языкового знака, в том числе – лексическое значение слова, не может быть рассмотрено в статике и в отрыве от связей, участником которых становится знак в процессе его действительного употребления. Обозначая часть прагматики знака термином «модальная рамка», Ю.Д. Апресян подчеркивает, что данный аспект является преимущественно экстралингвистическим по своей сущности и состоит главным образом из определенных психологических интенций говорящего. Применяя к концепции Ю.Д. Апресяна понятие контекста, мы со всей очевидностью придем к заключению, что в качестве такового будут выступать «культурные представления и традиции, которые связаны с тем или иным словом, господствующую в обществе практику использования тех или иных предметов и иные экстралингвистические факторы» [2, с. 67] – иными словами, те обстоятельства, которые сопутствовали генерации текста или высказывания или сопровождают процесс его понимания. Дальнейшее развитие описанной концепции позволяет утверждать, что одни лишь экстралингвистические

факторы могут порождать не менее четырех контекстов: 1) социокультурный контекст генерации высказывания, 2) социокультурный контекст понимания высказывания, 3) модальная рамка говорящего, 4) модальная рамка понимающего. Дальнейшее ветвление указанных контекстов фактически способно приблизить нас к ситуации их бесконечности, о чем было сказано выше применительно к идеям А.Ф. Лосева.

В целом, говоря о трех аспектах языкового знака – синтактике, семантике и прагматике, – необходимо обозначить взаимосвязь, существующую между указанными аспектами, тем самым определив ответ на вопрос о степени зависимости значения языкового знака от его сочетаемости с другими знаками в высказывании (для собственного удобства назовем такое окружение знака «физическим контекстом», подчеркивая, что речь идет исключительно о языковых единицах, физически находящихся поблизости от рассматриваемого знака) и от его практического употребления, от дополнительных значений, возникающих в ситуации использования знака (что мы, в свою очередь, определим как «идеальный контекст», обозначая тем самым принципиальную его экстралингвистичность).

В диссертационном исследовании Б.А. Плотникова «Исследование лексического значения путем квантитативного анализа дистрибутивных свойств слова» [81] отмечается, что особую актуальность, а также теоретическую и практическую значимость может приобретать разработка чисто лингвистических критериев анализа лексических значений – в частности, одним из таковых автору представляется дистрибутивно-статистический метод, особенности которого состоят в изучении закономерностей распределения языковых единиц в тексте или речи относительно друг друга, иными словами – их сочетаемости. По мнению исследователя, наличие связей между качеством элементов и их количественным распределением предоставляет возможность исследования внутренних свойств языковых единиц без непосредственного обращения к их содержанию; изучение лексической сочетаемости слов может быть применено в качестве наиболее ценного средства описания лексических значений.

Б.А. Плотников поясняет, что «в то время как лексическое значение слова в целом является фактически неисчерпаемым

по своему содержанию, при своем употреблении в тексте оно реализует тот или иной вариант своего значения, который уточняется контекстом, или сочетаемостью данного слова с другими словами» [81, с. 7]. По мнению ученого, наиболее полное представление о значении слова может быть получено посредством фиксирования сколь возможно большего количества различных вариантов значений, каждый из которых, в свою очередь, выявляется в новом контексте; контекст же употребления слова, продолжает исследователь, образуется из дистрибутивных свойств данного слова – совокупности сочетаний с другими словами в связном тексте либо их совместная встречаемость в словарном тексте. Таким образом, делает вывод автор, посредством количественных методов обработки данных об употреблении слова в тексте может быть измерена значимость тех или иных вариантов лексического значения, определен семантический объем слова и т.д. – иными словами, качественные и количественные характеристики слова непосредственным образом связаны.

Исходя из сказанного выше, представляется допустимым признать, что существует взаимосвязь между синтактикой и семантикой слова, т.е. зависимость значения от физического контекста.

В свою очередь, американские исследователи Р. Шенк, Л. Бирнбаум и Дж. Мей в совместной статье «К интеграции семантики и прагматики» [114] считают необходимым ставить вопрос о тесной взаимосвязи семантики с прагматическим аспектом языкового знака. По мнению авторов, семантические знания применяются в процессе обработки языковых данных аналогично прагматическим; «семантика – неотъемлемая часть прагматики, то есть нашего общего знания о мире и об использовании языка» [114, с. 33]. Понимание может быть достигнуто объединенным применением семантических и прагматических знаний, из чего непосредственным образом вытекает утверждение авторов о тесной и неразрывной связи лексикона с прочими знаниями, эрудицией, кругозором (утверждение, фактически аналогичное размышлениям Ю.Д. Апресяна на соответствующую тему). Исследователи, таким образом, полагают, что семантика неразрывно связана с прагматикой и не может изучаться независимо от нее. Интеграль-

ное единство семантики и прагматики детерминируется в работе как «концептуальный уровень».

Авторы отмечают существование стандартного возражения против рассматриваемой концепции, заключающегося в том, что осуществляется апелляция к практической невозможности действительного указания всех знаний о мире, которыми обладает человек; по их мнению, подобное возражение не может обесценивать описываемую теорию, т.к. в данном случае речь идет скорее не об указании всех знаний, но о способности вызывать эти знания из памяти. Цель теории американских исследователей, по их собственным словам, состоит в демонстрации того, «как некоторая совокупность знаний о мире используется при обработке языковых данных в предположении, что говорящий / слушающий эти знания имеет» [114, с. 35].

Интегральная концепция может, по мнению авторов, иметь некоторые небезынтересные гипотетические следствия. В частности, процесс понимания буквальных высказываний может в значительной степени быть уподоблен процессу понимания метафорических высказываний; фактически данный комплекс представлений отчасти свидетельствует о том, что язык всегда в некоторой степени метафоричен. Опираясь на изложенные соображения, возможно говорить также и о существовании связи между семантикой и прагматикой, т.е. о зависимости конкретно-дискретного лексического значения от идеального контекста высказывания.

Таким образом, будет вполне естественным прийти к заключению о том, что семантика языкового знака принципиально может быть рассмотрена в отрыве от сопутствующих ей синтаксического и прагматического аспектов, однако в целях максимально полного понимания значения может представляться целесообразным привлечение указанных аспектов в их взаимосвязи с семантикой, иначе – рассмотрение языкового знака как в физическом, так и в идеальном контексте.

Вопросы зависимости значения от структуры слова рассмотрены в работах известного отечественного исследователя Е.С. Кубряковой [49], уделявшей существенное внимание вопросам семантики сложных и производных слов в их отличии от простых; по мнению исследователя, каждое употребление

слова вызывается необходимостью объективировать, выявить, реализовать его значения как ту часть сведений о мире или информации о нем, получение которой является результатом познавательной и общественной деятельности человека и которая соотнесена с конкретным языковым знаком.

Е.С. Кубрякова указывает, что необходимость описания семантики слова в терминах семантической или смысловой структуры была впервые обозначена в работах В.В. Виноградова; следует отметить, что понятие смысловой структуры было введено в целях отражения системы значений одного слова как целостного объединения его лексико-семантических вариантов. Важной характеристикой смысловой структуры слова выступает, по мнению ученого, связь значений; в процессе исследования семантики производного слова выдвигается утверждение о том, что данная характеристика может быть организована в соответствии с особой формальной схемой.

В цитируемом исследовании отмечается, что простые слова характеризуются референтностью исключительно по отношению к т.н. миру действительности, в то время как производные слова обращены и к миру действительности, и к «миру слов». Ассоциации между словами могут иметь, по мнению Е.С. Кубряковой, разнообразный характер, отражая не только экстралингвистические, но и чисто лингвистические связи между однокоренными словами либо словами, построенными по одной словообразовательной модели. Отмечено, что пренебрежение подобными связями являлось негативной чертой генеративной грамматики.

Говоря о понятии семантической обусловленности производного слова, исследователь замечает, что сама возможность объяснения индивидуального значения того или иного слова через посредство другого ставит вопрос о необходимости системного описания подобных фактов, отражающих регулярно повторяющиеся семантические связи между однородными образованиями. Материальная база для поддержания семантической ассоциации между мотивирующим и мотивированным словами создается за счет формальной повторяемости одного и того же корня, основы, совокупности основ слова, однако это не позволяет со всей уверенностью говорить о характере подобной связи. Изучение всех возможных типов со-

относительности мотивированных и мотивирующих единиц позволяло, по мнению Е.С. Кубряковой, существенно продвинуть разработку теории мотивации. В свою очередь, анализ процесса установления связи между определенным предметом объективной реальности и некоторыми иными мог бы явиться средством понимания природы и сущности словообразовательного значения.

Е.С. Кубрякова отмечает, что значения, повторяющиеся в семантических дефинициях однотипных слов, могут рассматриваться в качестве лежащих в основе некоторых единых формул мотивации, действующих в пределах определенных словообразовательных моделей. В каждой подобной модели формальные средства служат для указания на использование предмета или тип его связи с другими объектами. Этим подчеркивается не только регламентированность и моделируемость семантики слова, но и системность характера семантики лексических единиц, представляющих некоторые «свои» серии форм, что позволяет, по мнению ученого, ставить вопрос о предсказуемости значений слов, в частности – производных.

Характерной чертой семантики производного слова является, с точки зрения исследователя, внутренняя расчлененность передаваемого содержания, сложный составной характер; напротив, семантическая структура простого слова существует как обозначение некоторой совокупности признаков, в которой ни одному из них нельзя приписать автономной сущности. Подобная совокупность целостна и интегративна; лексическое значение и значение семантического разряда здесь нераздельны [49, с. 25]. Несмотря на возможность выделения в составе простых и производных слов одних и тех же семантических признаков, форма их представления также оказывается фундаментально различной. Указанное различие лингвистически маркировано способом оформления одного и того же содержания.

В соответствии с позицией Е.С. Кубряковой, если по объекту наименования производные и сложные слова имеют черты определенного сходства с простыми словами, а по формальным параметрам разделяют характеристики класса слов в целом, то по способу отражения действительности могут

представляться изоморфными синтаксическим структурам. Семантику подобных единиц возможно считать композиционной функцией семантики.

Комментируя представленные положения, отметим, что дальнейшее развитие идей эвристического понимания производных слов, равно как и сам изложенный выше комплекс представлений о характере и сущности лексического значения простых и производных слов, может оказать существенное влияние на исследования в области когитологии.

1.2. ЛЕКСИЧЕСКАЯ ПОЛИСЕМИЯ И ЕЕ РАЗНОВИДНОСТИ

Рассмотрев базовые концепции, имеющие непосредственное отношение к вопросу о сущности лексического значения, считаем целесообразным перейти к изложению опорных представлений о лексической многозначности. В частности, Д.Н. Шмелев [115], комментируя вопросы, связанные с семантической структурой слова, отмечает, что факт обусловленности контекстом всех значений многозначных слов не означает ее одинаковой существенности для них. Традиционное деление значений слов на первичные и вторичные, если оно основано на действительном соотношении этих значений, корректно отражает неравноправное положение этих значений в семантической системе; в то же время нуждается в уточнении общепринятое определение видов лексических значений, ставящее вопрос разграничения главных и частных значений в зависимость от контекста. Данная дефиниция, по мнению Д.Н. Шмелева, не до конца объясняет сущность этих разновидностей лексических значений; кроме того, из него остается неясным, следует ли признавать частные значения закрепленными за словом либо же фактами речи. Исследователь также указывает, что между лексическими значениями слов устанавливается определенное соотношение – все значения, кроме основного, не имеют сколь-либо существенной парадигматической закрепленности, однако вполне очевидна их семантическая закрепленность, – и поясняет, что все значения слова реально не ограничены в возможностях собственного совмещения. В силу вышеизложенного ученый предлагает понимать под ос-

новным то значение слова, которое наиболее обусловлено парадигматически и наименее – синтагматически [115, с. 214]. Д.Н. Шмелев отмечает также факт изменчивости семантической структуры слов, выражающейся в изменении соотношений между лексическими значениями: к примеру, в процессе исторического развития языка первичные значения могут быть вытеснены переносными. Семантическое развитие слов может, по мнению исследователя, осуществляться как от конкретного к абстрактному, так и наоборот. Исследование семантических изменений, происходящих в лексике, подтверждает, что в семантической структуре слова объективно могут быть выделены отдельные элементы, в своей совокупности представляющие собственно лексическое значение. Представляет интерес также и выяснение того, какие из упомянутых элементов оказываются устойчивыми в диахронии, а также того, посредством каких семантических признаков различные значения многозначных слов оказываются объединенными в семантической структуре одного и того же слова [115, с. 219]. По мнению ученого, лексико-семантические трансформации играют исключительно важную роль в организации синхронической системы лексики. Фактором относительной устойчивости словарного состава языка и фактором изменения значений слов является также и функциональный момент, т.е. изменение формы и внутреннего устройства предметов, характера отдельных действий и т.д.; исследователь предполагает, что функциональный элемент является, в частности, неотъемлемой составляющей семантической организации названий предметов, созданных человеком; все они функционально ориентированы в семантическом отношении. От того, допускает ли смысловое содержание слова его функциональную интерпретацию, зависит, по Д.Н. Шмелеву, возможность его употребления в тех или иных семантических условиях.

По мнению ученого, требует рассмотрения также эмоциональный аспект употребления слова. Возможен вопрос о связи предметно-логического содержания слова с его экспрессивной значимостью; Д.Н. Шмелев замечает, что необходимо различать три группы слов: 1) те, которым эмоционально-оценочные значения свойственны изначально, 2) те, эмоциональная значимость которых создается посредством

словообразования, 3) те, в собственно лексическом значении которых заключена оценка обозначаемого явления. Таким образом, в составе эмоционально окрашенной лексики находятся слова с различными функциями; кроме того, слова, в значении которых заключена та или иная оценка, часто являются одновременно стилистически окрашенными, и наоборот. Исследователь заключает, что значимым в слове является не только его предметно-понятийное содержание, но и экспрессивно-стилистическая окраска в широком смысле.

Представления, изложенные Д.Н. Шмелевым, акцентируют внимание на дополнительных критериях выделения первичных и вторичных лексических значений. Не отрицая существенной значимости контекста, исследователь отмечает, что вопрос о первичности значения не может быть решен исключительно за счет анализа его контекстуальной обусловленности; иным важным аспектом ученому представляется парадигматическая обусловленность того или иного значения, что возвращает нас к необходимости учета внутриязыковых механизмов самоорганизации при формировании представлений о структуре языкового (и, в частности, лексического) значения.

Действительно, было бы не вполне корректным признавать исключительность роли внеязыковых факторов при определении или формировании конкретно-дискретных значений языковых знаков. Частичный переход вопроса о частных (в терминологии Д.Н. Шмелева) значениях слов в сферу внутренних механизмов языка как самоорганизующейся системы имеет существенное значение для автоматизированного моделирования процессов понимания, поскольку фактически обосновывает возможность успешного машинного понимания частных значений за счет внутренних ресурсов текста и языка без необходимости использования традиционно немашинных понятий из сферы эмоционально-психологических особенностей говорящего и понимающего. Однако необходимо заметить, что указанный переход может быть лишь частичным, поскольку сам Д.Н. Шмелев указывает в дальнейшем на важность эмоционального аспекта употребления слова.

В свою очередь, в уже упоминавшемся исследовании «Лексическая семантика» Ю.Д. Апресян, рассматривая вопросы о подобии явлений лексической многозначности и словообра-

зования, а также синонимии, указывает на необходимость выделения различных типов неоднозначности с целью формулирования результирующего определения лексической многозначности. В частности, исследователь выделяет следующие типы неоднозначности:

1) Синтаксическая и лексическая неоднозначность. Согласно Ю.Д. Апресяну, в ряде многозначных высказываний существенно необходимо отделение т.н. синтаксической омонимии (т.е. случая, в котором двусмысленность толкования высказывания обусловлена исключительно возможностью приписать фразе 2 и более синтаксические структуры) от собственно лексической многозначности либо ее сочетания с синтаксической неоднозначностью.

2) Языковая и речевая многозначность. Исследователь поясняет, что многозначность может быть провоцирована как собственно внутриязыковыми факторами (устройством грамматического строя либо характером словаря), так и внеязыковыми причинами, определяемыми Ю.Д. Апресяном как «энциклопедические факторы» (т.е. апеллирующие к общей эрудиции, кругозору и т.п.). Особым видом речевой многозначности Ю.Д. Апресян полагает также вопросительные предложения, которые могут быть представлены либо как средство запроса не известной говорящему информации, либо как средство проверки наличия соответствующей информации в распоряжении адресата, а также предложения наподобие Этот текст не может быть усовершенствован, которые не содержат никакой явно выраженной информации о причинах высказанного суждения.

3) Метонимически и метафорически мотивированная многозначность. В связи с разделением данных типов неоднозначности Ю.Д. Апресян отмечает, что обычное определение многозначности не может быть применено в отношении тех метафорически мотивированных значений, «словарное толкование которых не обнаруживает даже частичного сходства со словарным толкованием исходных значений» [2, с. 178] – т.е. случаев метафорического переноса, основанного на семантических ассоциациях, или коннотациях. Комментируя существование подобных значений, исследователь обращается к необходимости введения т.н. прагматической, или конно-

тативной, зоны словарной статьи, содержащей семантические ассоциации слова в том или ином значении.

4) Многозначность слова и диффузность значений. Ученый считает целесообразным обратить внимание на тот факт, что «концепция многозначности покоится на представлении о дискретной организации лексических значений», в то время как «фактически, при ориентации на реальное разнообразие текстов, обнаруживается неясная, размытая промежуточная область» [2, с. 179]. В данном контексте Ю.Д. Апресян ссылается на Д.Н. Шмелева и иных исследователей, придерживающихся представлений о диффузности (неопределенности, двусмысленности) лексических значений многозначного слова в его действительном употреблении.

5) Топологические типы многозначности. Исходя из более ранних исследований, Ю.Д. Апресян считает необходимым выделить три основных типа лексической многозначности: а) радиальная полисемия, в случае которой все значения слова мотивированы некоторым центральным значением; б) цепочечная полисемия, в случае которой каждое последующее значение слова мотивируется предыдущим; в) радиально-цепочечная полисемия как случай смешения двух предыдущих типов. В связи с разделением топологических типов многозначности Ю.Д. Апресян полагает полезным рассмотрение различия между непосредственной и опосредованной многозначностью, т.е. обнаруживаемой на первом и на последующих шагах семантического описания.

6) Многозначность и омонимия. Исследователь представляет обозрение некоторых критериев омонимии, считавшихся объективными на момент их разработки, а также аргументирует целесообразность отказа от данных критериев ввиду их противоречивости. Для четкого разграничения полисемии и омонимии ученый предлагает «формализовать понятие семантического сходства-несходства значений — единственный фактор, непосредственно отражающий существо этих двух явлений» [2, с. 184], в результате чего выдвигается тезис об относительности понятий полисемии и омонимии и о возможности выделения типов полисемии, в различной степени удаленных от омонимии.

7) Многозначность и моносемия. Рассматривая вопрос о

соотношении многозначности и моносемии, Ю.Д. Апресян отмечает возможность существования т.н. исключительно-дизъюнктивной (полисемичной) и включительно-дизъюнктивной (моносемичной) организации значений, т.е., говоря логическим языком, A = B XOR C (1) и A = B OR C (2) соответственно. Очевидно, что выражение (1) не допускает одновременной реализации B и C, в то время как для выражения (2) данный вариант нормативен, хотя, как комментирует ученый, исключительная дизъюнкция принципиально допускает двусмысленность в тексте, в то время как в случае включительной дизъюнкции это невозможно.

Исходя из представленных выше соображений, Ю.Д. Апресян приходит к заключению, что лексическая многозначность должна быть определена через понятие сходства значений. Поясним, что сходство значений определяется ученым как существование таких уровней семантического описания, на которых их толкования либо коннотации имеют нетривиальную общую часть, выполняющую в толковании одну и ту же роль относительно других семантических компонентов. В итоге Ю.Д. Апресян дает следующее определение лексической многозначности:

«Слово A называется многозначным, если для любых двух его значений a_i и a_j найдутся такие значения a_1, a_2, ..., a_k, a_l, что a_i сходно с a_1, a_l – с a_2 и т.д., a_k с a_l и a_l — с a_j» [2, с. 187].

Определения, данные Ю.Д. Апресяном, а также ряд идей, высказанных исследователем в процессе рассмотрения различных типов неоднозначности, могут представляться существенно ценным материалом при теоретическом исследовании вопросов автоматизированного моделирования процессов понимания многозначных слов и высказываний. Определенный интерес вызывает тот факт, что само определение многозначности осуществляется в терминах теории множеств; это позволяет с уверенностью говорить о возможности успешного математического моделирования явления полисемии и структуры значений полисемичного слова. Необходимо согласиться с мыслью исследователя о необходимости использования прагматической зоны словарной статьи (или компонента записи в БД в случае машинного словаря); самообучающиеся системы, оснащенные эвристическими алгоритмами, со всей

очевидностью успешно могли бы устанавливать временные внутритекстовые связи между словами и фразами на основании соотнесения контекста и данных из прагматической зоны словарных статей.

Представляет интерес также осуществленное Ю.Д. Апресяном разграничение типов неоднозначности в целом: отсутствие четко определенных рамок изучаемого явления может привести к исследованию смежных с ним областей, вследствие чего достоверность получаемых данных может закономерно подвергаться сомнению. Разработчик системы машинного понимания текстов на естественном языке, в свою очередь, не может не принимать во внимание факт существования различных типов неоднозначности, поскольку каждому из подобных типов должен быть поставлен в соответствие тот или иной алгоритм извлечения смысла. В связи с этим закономерен поднятый Ю.Д. Апресяном вопрос о формализации дефиниций языковых объектов с целью их однозначного разграничения и выявления их сущностных отличий, которые позволили бы автоматизированному анализатору применять к тексту определенный алгоритмический набор и адаптировать состав данного набора в зависимости от выявленных характеристик конкретного речевого произведения.

1.3. СОВРЕМЕННЫЕ ПРОДУКТЫ И РЕШЕНИЯ В СФЕРЕ АНАЛИЗА ТЕКСТОВ

RCO – Russian Context Optimizer

Официальный информационный ресурс RCO в сети Интернет [144] отмечает, что в ходе проводимых специалистами данной компании исследований был накоплен обширный опыт разработки и внедрения разнообразных компонентов информационно-поисковых и аналитических систем с применением последних достижений компьютерной лингвистики и искусственного интеллекта. Основной задачей компании является решение тех прикладных задач, которые требуют высокоточной компьютерной обработки данных на естественных языках. В продуктах разработчика применяется широкий спектр оригинальных алгоритмов и технологий интеллектуальной обработки текстов на естественном языке, за счет чего

выпускаемые им решения способны успешно составлять содержательный портрет текста, извлекать именованные объекты, вычленять связи и факты из массивов неструктурированных данных, анализировать тональность текста, выявлять инородные включения и заимствования, обнаруживать дублирующиеся тексты и их фрагменты.

Под маркой RCO выпускается серия продуктов, которые предназначены для поддержки и аналитического обеспечения различных экспертных систем, использующих средства поиска и анализа текстовой информации – в частности, к числу таковых могут быть отнесены информационно-поисковые и аналитические системы, электронные архивы и системы управления документооборотом. Для выполнения поставленных задач в решениях разработчика задействуются инновационные авторские технологии обработки текста, математические и лингвистические алгоритмы, которые могут быть применены для успешного решения тех или иных прикладных задач: от контекстного поиска документов с учетом многообразия словоформ, синонимов и опечаток вплоть до создания полноценных систем поддержки принятия экспертных решений на основе анализа информационных массивов. Рассматриваемые продукты могут успешно интегрироваться и взаимодействовать с системными решениями и средствами управления базами данных от корпораций Oracle и Microsoft. Представляется необходимым упомянуть следующие основные программные компоненты производства RCO, входящие в состав инструментариев разработчика и аналитика:

1. RCO Morphology SDK – библиотека, предназначенная для нужд разработчиков информационно-поисковых систем; она обеспечивает нормализацию формы и выведение всех возможных грамматических форм для любой лексической единицы русского языка.

2. RCO Morphology Professional SDK – библиотека, которая позволяет интегрировать средства работы с русской морфологией в системы автоматического анализа текстов и поиска данных. Рассматриваемый продукт поддерживает широкий спектр возможностей грамматического анализа русскоязычной лексики.

3. RCO Fact Extractor SDK – комплексный набор инстру-

ментов для разработки информационно-поисковых и аналитических систем, процесс функционирования которых требует осуществления лингвистического анализа русскоязычных текстов. Основой данного пакета является библиотека RCO FX Ru, которая способна к осуществлению полного синтактико-семантического разбора исследуемого текста; кроме нее, в состав пакета входят исходные коды приложений на языке программирования C++, которые показывают на примерах возможности использования библиотеки RCO FX Ru для решения полного спектра стандартных задач, встающих перед разработчиками прикладных информационных систем.

4. RCO KAOT – информационно-аналитическая система, обеспечивающая работу в локальной сети на базе операционных систем Microsoft Windows и серверных решений Microsoft Internet Information Server (IIS), которая реализует на практике обширный комплекс функций интеллектуального анализа данных и поиска определенной текстовой информации через пользовательский веб-интерфейс. Рассматриваемый продукт может являться основным конструктивным элементом автоматизированного рабочего места аналитика; в его состав входят инновационные технологии обработки естественноязыковых текстов, а также лингвистические и математические алгоритмы, которые могут быть применены для решения широкого спектра задач.

5. RCO Fact Extractor Desktop – персональное приложение для операционных систем Microsoft Windows, предназначение которого состоит в аналитической обработке русскоязычных текстов и выявлении разнообразных фактов, имеющих непосредственное отношение к задаваемым для исследования объектам – к примеру, определенным лицам либо предприятиям и организациям. Данное аналитическое приложение способно решать задачи, относящиеся к тематической области компьютерной разведки, т.е. требующие осуществления поиска информации с высокой точностью – так, одним из примеров может являться автоматический подбор материала для составления досье на целевой объект, либо же мониторинг определенных сторон его активности на основании публикаций в средствах массовой информации.

6. RCO Text Categorization Engine – библиотека для нужд

разработчиков информационно-поисковых систем, которая обеспечивает определение принадлежности текста к заданному множеству категорий.

NooLab Conext

НооЛаб - новосибирская инновационная компания, образованная в 1998 году [151]. Исследователи компании изучают технологии и решения по семантическому анализу текстов на естественном языке для задач поиска документов в Сети, аналитики, совершения экспертных выводов и машинного перевода, решают задачи автоматического понимания текстов по содержанию, ведут исследовательские работы по созданию интеллектуальных систем нового поколения, а также разрабатывают специализированные базы знаний, средства управления знаниями и человеко-машинные комплексы.

Одним из основных проектов компании является ConExT - система автоматизированного извлечения знаний из текстов на естественном языке. Данная система нацелена на решение ряда практических задач, в числе которых – помощь пользователю в получении знаний из больших массивов текстовых документов без необходимости подробного ознакомления с их содержанием. На вход системы подаются тексты на естественном языке, содержание которых подвергается реконструированию; в результате данного исследования выделяются имеющиеся в тексте знания, которые могут быть либо представлены пользователю в виде рефератов и схем, либо направлены в базу знаний. Разработчик предлагает собственную модель обработки, эффективность которой была подтверждена экспериментально; ее основой является метод выделения содержательно-значащих единиц, которые отделяются от собственно грамматической формы и переводятся на уровень моделирования содержания. Реализация описываемого метода осуществляется за счет привлечения модельных представлений двух типов: «1) правила, традиции и феномены выражения мыслей в текстах на русском языке; 2) инвариантный к языку ограниченный набор категорий, использующихся при выражении подавляющего большинства знаний о мире» [151]. Одним из основных структурных элементов системы является язык программирования tExp, разработанный для

обеспечения ее тонкой настройки и проведения структурного анализа неформализованных текстов на естественном языке.

Проект ВААЛ

Работа над системой ВААЛ ведется отечественными разработчиками с 1992 г. [141]; посредством ее применения оператор может прогнозировать подсознательный эффект, производимый тем или иным текстом на слушателей, осуществлять анализ текстов с целью выявить характер и направленность указанного эффекта, генерировать текстовые сообщения для оказания определенного воздействия на аудиторию, определять и оценивать психологические и личностные качества автора того или иного текста, проводить углубленное исследование содержимого и т.д.

В число основных возможностей рассматриваемой системы входят:

- оценка психологического и эмоционального воздействия, оказываемого фонетической структурой текстов и отдельных лексических единиц на подсознание слушателя,
- генерация слов и текстовых сообщений, обладающих определенными фоносемантическими характеристиками,
- указание специфических характеристик эффекта, которого необходимо достичь тем или иным сообщением, и корректировка требуемых параметров текстов для реализации желаемого воздействия,
- оценка комплекса синестезийных характеристик текстовых сообщений и лексических единиц,
- проведение словарного анализа текстов,
- осуществление анализа содержимого текста и его категоризация,
- определение специфических тем в пределах анализируемого текста и автоматическая его классификация в соответствии с выделенной тематикой,
- проведение эмоционально-лексического исследования текстов,
- проведение вторичного анализа данных посредством технологий визуализации, факторного и корреляционного анализа.

Система реализована в виде набора DLL-библиотек, подключаемых к текстовому редактору Microsoft Office Word; данный способ позволяет упростить освоение системы и ее последующее использование.

Cognitive Technologies

Компания Cognitive Technologies, созданная на базе научного коллектива Института системного анализа Российской Академии наук, развивает традиции академических научных исследований в передовых областях когнитивных наук [142]. Научные достижения и разработки коллектива в области информатики и искусственного интеллекта обеспечивают создание инновационных технологий, которые затем внедряются в новые продукты или проектные решения компании. Сильной стороной внедряемых компанией технологий является использование в них уникальных наукоемких элементов собственных исследований. Научные исследования в Cognitive Technologies традиционно ведутся в двух основных направлениях: теория создания баз данных и сложно-структурированных информационных систем, исследования в области искусственного интеллекта.

Наиболее актуальными направлениями исследований, осуществляемых Cognitive Technologies в области искусственного интеллекта и автоматизированного анализа текстов, являются обработка и последующее распознавание изображений (включая как печатные тексты, так и рукописные образцы), выделение и распознавание разнообразных объектов в видеопотоке (в том числе фрагментов текстовой информации), восприятие и синтез устной речи, анализ текстов по морфологическому и синтаксическому составу, рубрицирование потоков текста в автоматическом режиме, поиск нормализованных ключевых слов и словосочетаний, а также машинный перевод на иностранные языки и обратно.

Заслуживает упоминания такой продукт рассматриваемого производителя, как специализированный программный пакет синтаксического разбора и машинного перевода текстов Cognitive Dwarf. Как следует из его описания, основными структурными элементами пакета являются синтаксический анализатор для русского и английского языков и система ав-

томатического перевода в обоих направлениях. Аналитик имеет возможность применять данный инструмент для решения ряда задач, в том числе следующих:

- Статистическое исследование встречаемости определенных слов и конструкций в целевом языке. В тех случаях, когда востребовано преодоление неоднозначностей, возникающих в процессе синтаксического разбора текста и при выборе необходимых эквивалентов для его перевода, равно как и при статистическом моделировании языка в интересах распознавания акустической речевой информации, соответствующие закономерности могут иметь определяющее значение.
- Сопоставление изучаемых текстовых образцов по частоте встречаемости определенных слов и конструкций. Подобный функционал необходим при классифицировании документов, а также востребован в стилеметрии.
- Формирование корпусов ассоциированных разноязычных текстов.
- Построение специализированного набора инструментов для нужд АСОТ и его сопряжение с сопутствующими системами различного рода – от текстовых процессоров до СУБД.
- Создание методики машинного понимания текстов, а также извлечения из них фактов и данных с их последующей систематизацией.
- Любые иные задачи, связанные с синтаксическим разбором или автоматизированным переводом текстов.

По мнению специалистов Cognitive Technologies, необходимым условием успешной автоматизированной обработки текстов является решение ряда специфических задач повышенной сложности. К одной из таковых, в частности, они относят собственно разработку соответствующего инструментария. Если имеющаяся совокупность методик и аппаратно-программных средств не способна обеспечить надлежащий уровень проведения необходимых научных и практических исследований, то как теоретико-аналитическая, так и опытно-экспериментальная работа может существенно осложняться. Пакет Cognitive Dwarf с данной точки зрения обладает рядом преимуществ, поскольку способен обеспечить последователь-

ное развитие аналитического аппарата и его эффективное взаимодействие с другими техническими средствами обработки текстовой информации.

Технологии Text Mining

Как отмечает Д. Ландэ в обзорной статье о технологиях извлечения сведений из естественноязыковых текстов, суммарный объем необработанных и неструктурированных данных может достигать 90% от общего количества информации, с которой приходится работать пользователям [152]. Поиск необходимых сведений в обширном потоке данных составляет довольно трудоемкую задачу, решение которой представляется возможным исключительно с использованием современных технологий работы с информацией – в том числе электронной, в силу роста ее влияния и значимости в современном социальном контексте. Проблема избытка информации осложняется интенсивным развитием телекоммуникационных сетей, а также разрастанием корпоративных и ведомственных БД и цифровых архивах. В результате возникает «потребность в высокоэффективных технологиях автоматизированного анализа и классификации информации, представленной на естественном языке» [152].

Основанием для технологий извлечения знаний и фактов из текстов на естественных языках, которые часто обобщенно именуются Text Mining, служат средства статистического анализа и лингвистического исследования текстовых образцов вкупе с возможностями искусственного интеллекта. Их задача состоит в смысловой и семантической приоритезации содержимого предъявляемых текстов, техническом обеспечении эффективной навигации по документу и интеллектуального поиска данных. В результате использования подобных систем пользователи получают возможность извлекать из текстов наиболее важные сведения, факты и знания, не затрачивая временные и вычислительные ресурсы на обработку и структуризацию документов.

Д. Ландэ отмечает, что в настоящее время сложилась специфическая методология, описывающая базовые компоненты и технологии Text Mining. В частности, к таковым могут быть отнесены «суммаризация, выделение феноменов, понятий,

кластеризация, классификация, ответ на запросы, тематическое индексирование и поиск по ключевым словам» [152]. Кроме того, инструментарий Text Mining может быть дополнен и иными сопутствующими средствами – например, таксономическими либо тезаурусными средствами.

Классификация рассматривается Д. Ландэ как наиболее актуальная и часто решаемая задача в рамках концепции Text Mining. Под классификацией может пониматься группировка определенных элементов (как правило, записей в базах данных) на основании определенных критериев, распределение их по заданным категориям. При этом задача классификации может непосредственно соотноситься с распознаванием и самообучением информационно-аналитических систем; необходимо заметить, что сведения об объектах и всех характеристиках, им сопоставленных, требует ускорения и оптимизации процессов обработки информации.

Отчасти родственной задаче классификации является кластеризация – «выделение компактных подгрупп объектов с близкими свойствами» [152]. При этом на аналитическую систему возлагается обязательство по самостоятельному изысканию необходимых признаков, а также разнесению объектов и их атрибутов по группам. Как отмечает Д. Ландэ, можно различать несколько специфических разновидностей кластеризации, в том числе две основные: а) иерархическая и б) бинарная (двоичная). В первом случае кластеры организуются в виде иерархически организованного дерева, устанавливающего структуру взаимоотношений между группами документов, в то время как во втором случае речь идет об объединении исследуемых текстов по признаку взаимного подобия их свойств и содержимого. В основе кластеризации, как правило, лежит построение корпуса взаимонаправленных ссылок, позволяющих устанавливать связь между документами на основании наличия в них определенных ключевых слов и частоты их встречаемости. Чем больше тот или иной массив документов, тем более эффективной может быть данная процедура; кроме того, решение данной задачи может быть полезно для облегчения и упрощения тех или иных поисково-аналитических операций с исследуемыми текстами, а также для выявления близких по содержанию,

дублирующихся, либо, напротив, уникальных документов.

Также с процедурой кластеризации соотнесена задача поиска сходных параметров и атрибутов изучаемых документов. Не имея сведений о том, на основании каких именно признаков может быть установлена взаимосвязь документов, система проводит соответствующую операцию эвристически. В отличие от кластеризации, данная процедура работает с корпусом атрибутов.

Д. Ландэ указывает, что на данный момент на рынке имеется довольно много продуктов и решений от разных производителей, в рамках которых используются технологии Text Mining. В частности, исследователь считает необходимым особо отметить такие разработки, как Intelligent Miner for Text (IBM), TextAnalyst и WebAnalyst (Megaputer Intelligence), Text Miner (SAS), SemioMap (Semio Corp.), Oracle Text (Oracle), Knowledge Server (Autonomy), Galaktika-ZOOM (корпорация «Галактика») и InfoStream (Информационный центр «ЭЛВИСТИ»). Представим краткое обозрение данных систем.

Решение от IBM Intelligent Miner for Text является совокупностью автономных инструментов, каждый из которых способен выполнять определенные операции по обработке текстов. Данная система имеет репутацию одного из лучших продуктов для выполнения расширенного и углубленного анализа текстов [152]. В ее состав входят средство автоматической атрибуции языка документа, компонент категоризации исследуемого образца, инструмент группирования документов на основании сходства их стилевых или формальных характеристик, утилита определения темарематических особенностей текста и поиска определенных ключевых слов по словарю, а также элемент реферирования и аннотирования изучаемых текстовых образцов. Кроме того, в состав решения входят и некоторые другие базовые составляющие, призванные обеспечивать информационно-аналитическое и техническое сопровождение; с их помощью могут быть решены задачи индексирования веб-страниц и поиска по базам данных, изыскания требуемой информации на конкретных сетевых ресурсах либо в пределах локальной вычислительной сети организации, а также администрирования продукта и графического оформления средств поиска.

Продукт Megaputer Intelligence TextAnalyst, в свою очередь, обеспечивает аналитика инструментами построения и исследования семантической сети обрабатываемого текста, его резюмирования и реферирования, осуществления полнотекстового поиска, а также категорирования и группировки текстов по заданным показателям. В частности, формирование и анализ семантической сети обеспечивают эффективную навигацию по тексту вкупе с первичным пониманием содержимого на уровне ключевых понятий. Выделение значимых лексических единиц и анализ частоты их встречаемости позволяют добиться успешного составления резюме текста. Технологии Text Mining реализованы в рассматриваемой системе как обособленный комплекс алгоритмов и математического обеспечения, доступный для встраивания в решения посторонних разработчиков.

Вариацией TextAnalyst является продукт WebAnalyst того же поставщика. Данное интеллектуальное решение организовано по принципу «клиент-сервер» и предназначено для совершенствования и оптимизации процессов анализа веб-информации в корпоративной вычислительной среде. В то время как серверный компонент предназначен для централизованного вынесения экспертных решений и обеспечения информационного поиска, клиентские модули вносят в базу данных сведения о пользователях и об информации, их интересующей, обрабатывают полученный комплекс данных и формируют содержимое на базе индивидуального подхода к запросам посетителей. В итоге технологии Text Mining актуализируются в интересах поисковой оптимизации сетевых ресурсов, адресного веб-маркетинга и сколь возможно более полного удовлетворения запросов гостей и участников сетевого проекта.

Разработка компании SAS Institute Text Miner характеризуется универсальностью применения и позволяет выполнять разнообразные сопоставительно-аналитические операции. Она обеспечивает логическую обработку текста, предоставляя аналитику возможность объединения необработанной информации со структурированными сведениями определенного характера. Небезынтересно также отметить, что возможности Text Miner могут быть применены в интересах

установления степени истинности текстовых образцов; производитель указывает, что факт наличия ложных сведений в исследуемом документе может быть выявлен посредством поиска и выделения тех стилистических модификаций, которые неизбежно возникают при попытке искажения либо сокрытия сведений в письменной речи. При этом продукт не выполняет специфических процедур, направленных на машинное понимание текста, ограничиваясь статистическим анализом тенденций и аномалий. В базе данных Text Miner присутствуют эталонные образцы текстов, специфические характеристики которых также используются в процессе сопоставительного анализа. Компания предлагает продукт к применению на предприятиях и в организациях, на адреса которых поступают значительные объемы электропочтовых сообщений, а также рекомендует применять его для нужд правоохранительных органов наравне с полиграфами.

Система анализа от компании Entrieva SemioMap считается первым решением класса Text Mining, которое было реализовано в виде центрального серверного модуля и сопряженных с ним клиентских агентов [152]. В ходе работы системы осуществляются три основных вида операций – индексирование исследуемого содержимого и внесение полученной разметки в базу данных, понятийная кластеризация и формирование лексической сети документа на основании статистических сведений о встречаемости тех или иных слов и их дистрибутивной совместимости, а также визуализация полученных сведений и установленных связей средствами графического интерфейса. Последняя операция обеспечивает также оперативное открытие документа для чтения и успешное перемещение как по иерархической структуре текстов, так и по самому документу. В основе SemioMap лежит т.н. лексический экстрактор – средство статистического исследования, позволяющее выделять из текста слова, фразы, сверхфразовые единства и устанавливать частоту их встречаемости вкупе с возможностями их сочетаемости друг с другом в пределах одного и того же анализируемого текстового образца.

Средства Text Mining внедрены также и в системы управления базами данных от корпорации Oracle. Данный программный комплекс известен как Oracle Text. Его назначение

состоит в обработке поисковых запросов, имеющих непосредственное отношение к неструктурированным данным, причем пользователь одновременно располагает полным спектром возможностей по взаимодействию с реляционной БД – например, допускается применение команд на языке SQL непосредственно в процессе работы с текстовым образцом. При помощи технологий Text Mining обеспечивается совершенствование поиска и повышение количества обнаруживаемых по запросу документов: к изысканию подключаются различные морфологические вариации требуемых слов, синонимы из семантического словаря (тезауруса), а также полные или частичные омографы и омофоны. Данные вспомогательные процедуры способствуют более успешному решению поисковых задач, эффективно обнаруживая необходимые документы даже в случае наличия ошибок, лексических замен и опечаток. Кроме того, система способна определять основную тематику текста и осуществлять его реферирование.

Система Knowledge Server, производимая компанией Autonomy, проводит автоматическую классификацию текстов и обеспечивает построение структуры взаимонаправленных ссылок посредством применения специализированных шаблонов, анализа контекста, а также выделения знаний и фактов [152]. В основе системы лежат алгоритмы статистического анализа текстовых образцов, в работе которых используется сложный математический аппарат. Особого упоминания заслуживают т.н. концепция адаптивного вероятностного моделирования, посредством которой осуществляется выявление в текстах тех или иных шаблонов вкупе с поиском и определением ключевых концепций (что позволяет, к примеру, эффективно обнаруживать схожие по смыслу тексты, записи и документы), а также механизм динамического рассуждения, построенный на основе нейросетей. С их помощью реализуются базовые функции концепт-анализа, создания и обучения программных агентов, а также собственно поиска. Запросы могут формулироваться на естественном языке.

Решение Galaktika-ZOOM является разработкой российской компании «Галактика». Данная система нацелена на морфологический поиск информации в соответствии с заданными ключевыми терминами, а также построение корпусов

текстов и индексов. Объемы обрабатываемой информации при этом могут быть весьма существенными, что составляет одну из отличительных черт рассматриваемого продукта: он ориентирован на работу со средствами массовой информации, с методической и нормативной документацией, сетевыми информационными ресурсами, электронной корреспонденцией и т.д. В состав системы входит также совокупность специализированных инструментов, позволяющих аналитику исследовать объективные смысловые связи и строить многомерную модель конкретной проблемы в потоке данных [152]. Кроме того, система способна отслеживать, учитывать и прогнозировать динамику развития рассматриваемой проблемы.

Продукт информационного центра «ЭЛВИСТИ» – специализированная технология InfoStream – представляет собой совокупность аппаратных и программных средств, объединенных в рамках трех базовых компонентов: центра сбора и обработки информации, центра организации интерактивного доступа к базам данных и центра контент-мониторинга. В основе технологии обработки содержимого, реализованной в рассматриваемом продукте, находится средство полнотекстового поиска необходимых данных под наименованием InfoReS. Описываемая технология, как отмечает Д. Ландэ, предоставляет возможности по формированию полнотекстовых баз данных и проведению поиска информации, позволяет организовывать каналы передачи данных по тематическому принципу, обеспечивает классификацию и категоризацию данных в автоматическом режиме, систематизирует сведения и преобразует их в резюме и рефераты, строит таблицы, отражающие связь между теми или иными понятиями, а также использует средства визуализации для представления «веса» тех или иных понятий в обрабатываемом тексте вкупе с частотой и интенсивностью их встречаемости по временной шкале. Работоспособность системы не зависит от формата входных данных [152].

1.4. ФОРМАЛЬНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ В АСОТ

Несмотря на активное употребление термина «формально-семантический анализ» в современных исследованиях, не существует сколь-либо четкого определения настоящего термина. В докладе И.В. Азаровой и Е.А. Овчинниковой «Пропозициональное и формально-семантическое описание фактов при обработке текстов на русском языке» отмечено, что «можно говорить о процедуре семантического анализа, состоящей из нескольких этапов. На каждом этапе создается определенное семантическое описание анализируемых единиц, которое подвергается дальнейшей обработке на следующем этапе» [136]; в соответствии с данными представлениями исследователи предлагают различать 2 базовых этапа семантического анализа – а) выделение пропозиций, которые ставятся в соответствие фразам в пределах анализируемого текста и б) получение формально-семантического описания фактов на основании получаемых пропозициональных структур; описанные этапы определяются докладчиками как пропозициональный и формально-семантический компоненты автоматического анализа текста. Авторы поясняют, что «пропозициональная структура, формализованная в виде структуры признаков, может быть описана на языке формальной семантики, то есть, в виде формулы» [136]. Иными словами, в цитируемом исследовании предлагается понимать под формально-семантическим анализом описание семантики языковых знаков посредством формул, соответствующих определенному стандарту описания.

Близкой к идее формально-семантического анализа представляется концепция формально-концептуального анализа (Formal Concept Analysis), описанная в работе «Deriving Concept Hierarchies from Text by Smooth Formal Concept Analysis» коллектива авторов университета Карлсруэ, Германия [148]. Базовым понятием концепции, выстраиваемой авторами, является т.н. «ontology», определяемое как «формальное описание процесса концептуализации» («an ontology is a formal specification of a conceptualization»); в его состав входят идентификаторы концептов, словарный запас языка и т.д. Го-

воря о лексической реализации концептов, авторы отмечают, что концепт может быть выражен посредством различных высказываний (синонимов), либо одно высказывание может соответствовать различным концептам (что, по мнению исследователей, соответствует полисемии).

Определяя формально-концептуальный анализ, авторы комментируют, что главным образом он применяется для извлечения и обработки эксплицитной информации, которая впоследствии структурируется в т.н. формальные абстракции концептов (formal abstractions of concepts). Исследователи указывают, что центральным понятием формально-концептуального анализа является формальный контекст, определяемый как триада (G,M,I), где G – множество объектов, M – множество атрибутов, I – двоичное отношение между данными множествами, называемое сочетаемостью (incidence) контекста. Приведенная дефиниция позволяет авторам определить понятие формального концепта как диаду (A,B), где $A \subseteq G, B \subseteq M, A' = B \wedge A = B'$. Иными словами, (A,B) является формальным концептом тогда и только тогда, когда множество атрибутов объектов из A равно множеству B, и, в свою очередь, A является множеством всех объектов, объединенных множеством атрибутов из B. Соответственно, A предлагается определять как внешний аспект концепта, B – как внутренний. Концепты в пределах контекста ранжируются авторами на основании «субконцептуальных и сверхконцептуальных отношений» [148].

По мнению исследователей, основным вопросом, возникающим при попытке применения изложенных выше представлений к тексту, является вопрос о том, что именно надлежит полагать объектами и атрибутами. Наиболее очевидной авторам представляется мысль об извлечении из текста взаимосвязей глаголов и их дополнений; дополнения могут быть ассоциированы с множеством объектов, в то время как соответствующие им отглагольные прилагательные – с множеством атрибутов (к примеру, to rent an apartment – слово apartment будет отнесено к объектам, и ему будет соответствовать атрибут rentable).

Авторы отмечают, что существуют работы, в которых ontology рассматривается в качестве семиотической знаковой

системы, вследствие чего возможно сопоставление нескольких ontology на синтаксическом и семантическом уровне, выявление и сопоставление иерархии их концептов. Семантическое сопоставление, в частности, требует от авторов введения понятия semantic cotopy, понимаемого как множество всех субконцептов и суперконцептов и описываемого формулой

$$SC(c_i, \geq C) := \{c_j \mid c_i \geq Cc_j \vee c_j \geq Cc_i\},$$

где c_i и c_j принадлежат множеству концептов C.

Реализация описанных представлений позволяет исследователям произвести формальную кластеризацию слов в тексте по признаку их контекстуальной сочетаемости, выстраивая таким образом искомую ими иерархию концептов.

Понятие иерархии становится ключевым также в иных подобных исследованиях, в частности – в работе «A Formal Text Representation Model Based on Lexical Chaining» [150] группы ученых университета Билефельда, Германия. Авторы рассматривают вопросы, связанные с разработкой формальной модели представления текста, привлекая два различных подхода – изучение лексических последовательностей (lexical chaining) и квантитативный структурный анализ. Формальная модель лексических последовательностей выстраивается в настоящем исследовании на основании т.н. логической структуры документа и иерархии элементов содержимого (hierarchy of content objects).

Следует констатировать, что в работе А.Н. Баранова «Введение в прикладную лингвистику» [9] представлен комментарий, достоверно описывающий современные системы машинного понимания текстов: «Понимание высказываний (системой автоматизированного анализа. – Н.Г.) предполагает два основных этапа – анализ и интерпретация... Особенности построения систем поддержки диалога с ЭВМ таковы, что часто граница между синтаксическим и семантическим анализом почти отсутствует». Действительно, описанные выше исследования немецких авторов позволяют сделать вывод о том, что современные исследования в области формализации семантики нередко фактически сводятся к дистрибутивному анализу, т.е. к выявлению характера сочетаемости слова и определению физических контекстов, в которых оно может

употребляться; не отрицая ценности данного подхода, необходимо отметить, что исследование физического контекста не может рассматриваться как полноценное средство выявления и описания семантики языкового знака.

Замечания о соотношении синтаксического и семантического анализа представлены, в частности, в статье М. Селфридж «Интегральная обработка обеспечивает надежное понимание» [87]. Автор поясняет, что построение надежного интерфейса на естественном языке между машиной и пользователем требует определения двух разновидностей отношений: между синтаксисом и семантикой, а также между пониманием языка и обработкой памяти. Не касаясь второго отношения как внутримашинного, обратимся к вопросу о синтаксисе и семантике в понимании исследователя.

Резюмируя умозаключения Р. Шенка и Л. Бирнбаума, М. Селфридж формулирует следующие основные вопросы: 1) обрабатывается ли синтаксическая структура раньше семантической, или их анализ производится единовременно? и 2) разделены ли процессы обработки синтаксиса и семантики, или их обработка составляет единый процесс?

Автор отмечает, что существуют две различные точки зрения на эти вопросы, условно называемые «сепаратистской» («разъединительной») и «интегральной». Согласно первой позиции, «синтаксический анализ высказывания осуществляется до любого вида семантического анализа, а на его выходе мы получаем синтаксическое описание высказывания. Это описание затем подается на вход семантического анализа» [87, с. 169]; в свою очередь, интегральная концепция гласит, что данные типы анализа надлежит проводить синхронно. Ответ же на второй поставленный вопрос с точки зрения «разъединительной» концепции подразумевает отличие механизма, строящего синтаксическое описание, от механизма, конструирующего семантическое описание (т.е., комментирует автор, эти механизмы работают по различным алгоритмам); в свою очередь, интегральные представления предполагают использование единого механизма, выполняющего как семантический, так и синтаксический анализ. Шенк и Бирнбаум придерживаются второй концепции, формулируя т.н. гипотезу об интегральной обработке. По мнению М. Селфридж, указанная гипотеза не

утверждает тождественности синтаксического и семантического знания, так что при разработке авторской концепции понимающей машины исследователь считает необходимым обрабатывать синтаксис и семантику единовременно и посредством единого механизма, но при этом разграничивать их не на основании процессов обработки, но на основании знаний – иными словами, поясняет автор, «различие между синтаксисом и семантикой кроется в специфике знания, представляемого в соответствующих структурах, а не в порядке их применения и не в процессах обработки» [87, с. 170].

В ряде работ сам по себе термин «формально-семантический анализ» не употребляется, однако представляется вполне очевидным, что соответствующая проблематика находится в фокусе внимания их авторов. В диссертационном исследовании А.М. Шахмайкина «Некоторые принципы формализации семантики естественных языков (для целей семантической типологии)» [107] отмечается, что формализация не может быть сведена исключительно к использованию логико-математических методов, но также и предполагает однозначное и последовательное использование при анализе всего категориального аппарата семантики – т.е. процедура формализованного анализа состоит, по мнению ученого, «из установления инвентаря семантических признаков и отношений между ними» [107, с. 1]. Семантическая структура языкового знака представляется автором в виде формулы $\Sigma = < D, Dm, S, Sm >$, где $\sum$ – языковой знак, D – денотат, т.е. класс обозначаемых знаком явлений, Dm – денотативное значение (совокупность признаков денотата), S – сигнификат, образ, посредством которого кодифицируется денотат, и Sm – сигнификативное значение, т.е. совокупность сигнификативных признаков знака. Далее исследователь утверждает, что единообразное и эксплицитное описание семантической подсистемы естественных языков может быть обеспечено исключительно посредством формального языка-посредника, в силу чего насущно необходимой может представляться разработка семантического метаязыка, позволяющего формально описывать семантику языковых знаков; результатом подобного описания будет являться совокупность формул, образующих некоторую систему. По мнению ученого, подобный метаязык надлежит

строить отдельно для каждой семантической подсистемы – категорий времени, пространства, модальности и т.д. Структура семантического метаязыка представляется автору «в виде необходимых и достаточных для описания той или иной конкретной подсистемы элементарных неопределяемых и далее неделимых семантических единиц (множителей) и отношений между ними» [107, с. 11]; указанные отношения должны быть определены на базе формально-логических категорий. В качестве примера А.М. Шахмайкин приводит результаты анализа семантики грамматического времени русского языка, описывающие ряд временных значений:

$$1.\ (e_1 > e_2 \wedge e_2 \in T);\ 2.\ (e_1 \parallel e_2 \wedge e_2 \in T);\ 3.\ (e_1 > e_2 \wedge e_2 \in P);$$

$$4.\ (e_1 \parallel e_2 \wedge e_2 \in P);\ 5.\ (e_1 > e_2 \wedge e_2 \in F);\ 6.(e_1 \parallel e_2 \wedge e_2 \in F),$$

где е с индексами – события, > - знак предшествования, ‖ – знак одновременности, Λ – знак конъюнкции, ∈ – знак принадлежности, Т – настоящее время, Р – прошедшее время, F – будущее время.

Подводя итоги изложения основных представлений о формально-семантическом анализе, считаем необходимым суммировать вышесказанное в целях выработки единого определения данного термина. С этой целью отметим основные элементы, которые могли бы составить итоговое определение.

1. Представляется вполне очевидным, что дефиниция должна содержать указание на процесс формализации. Исходя из определений, данных в работах [136] и [148], будем понимать формализацию как процесс создания некоторого формального описания, т.е. описания чего-либо посредством формул и приближенных к ним выражений, таких, как псевдоалгоритмический язык и т.п. Фактически речь идет о разработке формальной модели изучаемого явления.

2. Кроме того, дефиниция подразумевает, что выполняется анализ семантики, т.е. некоторой совокупности значений, приписываемой языковому знаку, а также (возможно) структуры указанных значений. Идеалом данного анализа, безусловно, будет являться полное отделение семантики знака от его синтактики и прагматики, что не всегда достижимо, но может рассматриваться как конечная цель.

3. Рассматриваемый нами анализ может производиться как автоматически, так и при участии человека. Заметим, что в данном случае возможны некоторые процессуальные различия: человек может выполнять первичный формально-семантический анализ в интересах создания формальной модели семантики, перемещаясь в направлении от семантики к средствам ее выявления, в то время как автомат вероятнее всего будет использовать данную модель и выполнять обратную работу, двигаясь от средств выявления семантики к ней самой.

4. Исходя из предыдущей составляющей, необходимо прокомментировать понятие средств выявления семантики. Представляется вероятным, что в фокусе внимания как человеческого исследователя, так и автоматизированного анализатора будут находиться некоторые внешние, эксплицитные параметры языковых знаков или, шире, текстов; мы полагаем целесообразным придерживаться идеи о том, что внешние характеристики языкового знака позволяют с определенной вероятностью делать выводы о его имплицитных свойствах. Таким образом, мы можем сформулировать итоговую дефиницию:

Формально-семантический анализ – это процедура выявления и интерпретации эксплицитно выраженных характеристик языковых знаков, которые позволяют составить представление о ядерных элементах и структуре их семантики, и описания полученных данных посредством формул, пригодных для использования в построении математических моделей процессов понимания значений языковых знаков как в пределах физического и идеального контекста, так и вне его.

1.5. ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ АСОТ

Важность взаимодействия лингвистики и компьютерной науки была отмечена, в частности, Б.Ю. Городецким во вступительной статье к сборнику «Новое в зарубежной лингвистике: Компьютерная лингвистика» [28]. Исследователь поясняет в указанной статье, что существует ряд причин, обусловливающих потребность в лингвистическом обеспечении компьютеризации; в качестве основных подобных причин автор указывает прежде всего на существенные объемы информации,

накапливаемой и создаваемой в современном обществе, которая подлежит обработке при помощи ЭВМ; данная информация преимущественно организована средствами естественного языка, в силу чего ее действительное освоение возможно, по мнению ученого, лишь при условии автоматической смысловой обработки текстов без их предварительной подготовки человеком. Кроме того, продолжает Б.Ю. Городецкий, потребность в лингвистическом обеспечении обусловлена также нацеленностью современной науки на создание человеко-машинных систем решения интеллектуальных задач. Важным аспектом проблемы автор полагает и «необходимость учитывать лингвистическую семантику при построении языков представления знаний вне системы обработки текстов в ее полном виде» [28, с. 6]. В завершение исследователь обозначает социальную значимость лингвистических проблем компьютеризации, которая связана с формированием принципиально новых форм человеческой деятельности, и подводит итог формулированием следующего определения:

«Итак, лингвистическое обеспечение автоматизированных систем – это совокупность средств, позволяющих осуществлять компьютеризацию языковой деятельности, сопровождающей (в явной или скрытой форме) практически любую интеллектуальную деятельность человека» [28, с. 7].

Б.Ю. Городецкий поясняет, что внешние критерии, служащие ориентиром для создателей автоматизированных систем обработки текстов (АСОТ), в целом соответствуют единой формуле – «оптимизация взаимодействия человека и машины»; указанную задачу решает компьютерная лингвистика и ее наиболее существенная для исследователя составляющая – вычислительная семантика. Прикладные данные, получаемые в описываемой области знания, – семантические алгоритмы и семантические метаязыки, используемые при построении моделей языка и речевой деятельности. Исследователь отмечает безусловную сложность естественного языка для компьютерного моделирования, акцентируя внимание на необходимости создания специальных научных теорий, которые обеспечили бы данный процесс.

Рассмотрение т.н. «проблемы знаний» позволяет Б.Ю. Городецкому непосредственно увязать рассматриваемые им

проблемы с теорией искусственного интеллекта, а также с упомянутым выше понятием семантического метаязыка. По мнению автора, наиболее актуальной проблемой лингвистики является задача представления знаний, равнозначная задаче разработки семантических метаязыков; кроме того, «...семантический метаязык – неотъемлемый инструмент исследования языка в его когнитивном и коммуникативном аспектах. Он есть необходимая составная часть метода моделирования, применяемого к языку» [28, с. 28]. Исследователь указывает, что в класс семантических метаязыков могут также быть включены информационные языки, что позволяет говорить об использовании в информационной работе всего спектра теоретических и методических средств прикладной семантики.

Понятие метаязыка возводится Б.Ю. Городецким в ранг одного из ключевых в сфере моделирования языковых процессов, так как, по мнению ученого, процессы анализа текстов преимущественно сводятся к перезаписи информации на тех или иных – структурных либо алгоритмических – метаязыках.

Прежде чем переходить к более подробному рассмотрению собственно лингвистического обеспечения АСОТ, определимся с составом термина «автоматизированные системы обработки текстов» и обозначим, какие именно программные продукты следует отнести в данную группу, т.к. рассматриваемый термин неочевиден. Согласно уже упоминавшейся работе А.Н. Баранова «Введение в прикладную лингвистику» во взаимодействии с автоматизированными системами рассматриваются и решаются следующие вопросы оптимизации функционирования языка как средства передачи информации:

1. Перевод как прикладная лингвистическая дисциплина;
2. Оптимизация общения с ЭВМ: обработка естественного языка;
3. Теория и практика информационно-поисковых систем.

Таким образом, практические потребности, связанные с ростом объемов информации, провоцировали заинтересованность в развитии систем машинного, или автоматического, перевода. В процессе машинного перевода возможно применение ряда стратегий:

- стратегия «прямого перевода» - замены слов или словосочетаний на входном языке на их словарные эквивален-

ты в выходном языке. В подобных системах не выполняется моделирование функционирования языковой системы в целом, в редких случаях может выполняться анализ контекста; кроме того, прямой перевод всегда привязан к конкретной паре языков. Модификацией описанной стратегии является система с «трансфером», позволяющая не только выполнять пословный перевод, но и приспосабливать выход к структуре предложения языка-цели; в то время как в исходной стратегии нет различий между анализом и синтезом, в трансферных системах их различение может быть произведено.

- стратегия «языка-посредника», особенность которой состоит в том, что между структурами входного и выходного языков находятся один или несколько промежуточных языков, на которые по соответствующим правилам осуществляется перенос выражений языка-источника. В рамках данной стратегии осуществляется принципиальное разделение процессов анализа и синтеза; в качестве языков-посредников применяются синтаксические, семантические, семантико-синтаксические метаязыки, языки глубинной семантики, приближающиеся к концептуальному представлению языка в категориях теории знаний. Системы, основанные на знаниях, могут рассматриваться в качестве составной части систем искусственного интеллекта. Последовательное развитие данной идеи, продолжает А.Н. Баранов, привело к формированию замысла стратегии «универсального семантического языка», которая на данный момент не может быть реализована; в качестве компенсации указанной проблемы разрабатываются промежуточные стратегии, такие, как «сужение проблемной области» - ориентация на узкотематические сферы текстов и использование тематических словарей.

В рассматриваемой работе описываются также действительно существующие и активно используемые системы машинного перевода: GAT, CETA / GETA, TAUM, ЭТАП, ФРАП и т.д. Автор комментирует, что большинство подобных систем может быть отнесено к идейно устаревшим, в то время как потенциал вспомогательного компьютерного инструментария

остается лишь частично реализованным; перспективными, по мнению исследователя, являются также системы перевода на основе технологий искусственного интеллекта.

Развитие систем общения между человеком и ЭВМ на естественном языке, и, более широко, систем понимания естественноязыковых текстов, обусловлено, с точки зрения А.Н. Баранова, экспоненциальным ростом количества конечных пользователей ЭВМ, полномасштабное обучение которых программированию является принципиально невозможным; решение обозначенной проблемы производилось как посредством адаптации языков программирования и операционных систем к пользователю, так и через создание систем, которые позволяли бы взаимодействовать с ЭВМ на естественном языке либо его модифицированном варианте. Указанные системы включают ряд базовых блоков: диалоговый компонент, блок понимания, блок порождения высказываний.

В рамках рассматриваемой группы автор выделяет следующие типологические разновидности:

- вопросно-ответные системы как средство оптимизации взаимодействия с поисковыми системами, в частности – система ПОЭТ, программа BASEBALL; данные системы являются пассивными коммуникантами, рассматривающими вопросы пользователя как разновидность запроса информации.
- диалоговые системы решения задач, являющиеся, в отличие от предыдущих, активными коммуникантами, т.е. способными к самостоятельному формулированию вопросов в ситуации недостаточности информации.
- системы обработки связных текстов, функции которых состоят в понимании текста и представлении ответов на вопросы о его содержании; понимание, поясняет автор, сводится к процессу извлечения информации из текста, определяемому конкретной коммуникативной задачей. В качестве примеров могут быть приведены системы RESEARCHER, TAILOR, BORIS и подобные им.

Процесс развития информационных технологий, по мнению А.Н. Баранова, обусловил также создание информационно-поисковых систем, структура и организация которых определяются проблемной ситуацией:

1) имеется множество документов,

2) имеется коммуникативное задание, выраженное в запросе,

3) требуется найти тексты, соответствующие запросу и удовлетворяющие коммуникативную интенцию пользователя.

Центральными для описываемых систем являются понятия запроса и документа; степень соответствия запросу определяется как релевантность документа. Документ в информационно-поисковой системе характеризуется также поисковым образом – формализованным описанием на информационно-поисковом языке, - которому сопоставляется введенный пользователем запрос. Исследователь подразделяет подобные системы на документальные и фактографические ИПС, в которых могут находить применение различные типы формальных языков: иерархические языки-классификаторы, фасетные языки, а также языки дескрипторного типа, использующие легко формализуемые понятия поисковых координат и матрицы признаков (в качестве примера продукта последнего типа приводится система УНИТЕРМ). Новым этапом развития подобных языков автору представляется технология информационно-поискового тезауруса – дескрипторного словаря, описывающего допустимые при индексировании ключевые слова. В завершение обзора А.Н. Баранов обращается к рассмотрению критериев выдачи документов – вычисляемых и логических.

Таким образом, опираясь на представленные выше идеи автора, мы полагаем допустимым понимать под автоматизированными системами обработки текстов аппаратно-программные комплексы следующих базовых типов:

1) системы машинного перевода,

2) системы машинного понимания текстов на естественном языке,

3) информационно-поисковые системы.

Итак, как было отмечено выше, представляется допустимым утверждать, что вопросы лингвистического обеспечения АСОТ непосредственным образом связаны с разработками в сфере искусственного интеллекта, а также с проблемами формирования семантических метаязыков; полагаем необходимым осуществить более подробное рассмотре-

ние указанных разработок, проблем и теоретических выкладок.

Во вводной статье Р.Р. Мдивани к аналитическому обзору С.М. Шевенко «Лингвистика и проблемы машинного «понимания» естественного языка» [63] отмечается, что современная прикладная лингвистика непосредственным образом связана с исследованием проблем искусственного интеллекта; исследования в области создания систем понимания естественного языка осуществляются посредством теоретического и экспериментального изучения, а также моделирования процессов понимания текстов на естественном языке человеком. По мнению исследователя, конечной целью подобных исследований является создание программных продуктов, способных «понимать» тексты на естественном языке с целью их перевода, реферирования, перефразирования, ответов на вопросы и т.д. Так как в процессе понимания человек в существенной степени опирается на имеющиеся у него знания о действительности, в автоматизированных системах понимания естественного языка нашли активное применение экстралингвистические сведения, образующие тезаурус. Автор заключает, что в области искусственного интеллекта в силу вышеуказанных причин на современном этапе исследований центральным становится вопрос о природе и представлении знания.

Согласно А.Н. Баранову [9], для изучения и представления знаний в теории знания используются различные структуры знаний – фреймы, сценарии и планы. Определение фрейма впервые было дано М. Минским [69]; в соответствии с изложенными исследователем представлениями, «фрейм – это структура данных, предназначенная для представления стереотипной ситуации». А.Н. Баранов комментирует, что фрейм может пониматься как концептуальная структура, предназначенная для декларативного представления знаний о некоторой типизированной тематически единой ситуации, содержащей некоторое количество связанных между собой слотов. Формально фрейм может быть представлен в виде таблицы, предиката, кортежа и т.д. Автор считает необходимым отметить, что понятие фрейма является омонимичным; онтологически данная структура является частью когнитивной системы человека, в то время как инструментально она обеспечивает

декларативное представление знаний. С позиций содержания фрейм представляется исследователю близким категории толкования.

В свою очередь, сценарий «представляет собой концептуальную структуру для процедурного представления знаний о стереотипной ситуации или стереотипном поведении» [9, с. 17]. В трудах Р. Шенка и Р. Абельсона [132] сценарий понимается как простейшая понятийная структура, используемая для представления информации о стереотипных эпизодах, причинная цепочка, представляющая обычную последовательность событий. А.Н. Баранов поясняет, что сценарий может быть представлен как в виде алгоритма, так и в виде сети, отображающей связи между некоторыми ситуациями.

Результатом применения сценария к конкретной проблемной ситуации становится, по мнению исследователя, план, применяемый для процедурного представления знаний о возможных действиях, которые ведут к достижению определенной цели. В общем случае план включает некоторую последовательность процедур, переводящих начальное состояние системы в конечное; в системах ИИ, отмечает автор, план возникает в результате работы соответствующего модуля – модуля планирования. В отличие от сценария, план связан с конкретной ситуацией и характеризуется категорией выполнимости.

Кроме того, А.Н. Баранов обозначает связь между категориями фрейма и сцены. Сцена понимается исследователем как концептуальная структура для декларативного представления ситуаций и их компонентов, актуализованных в речевом акте и выделенных теми или иными языковыми средствами; в системах искусственного интеллекта сцены могут быть применены при распознавании образов и при анализе проблемных ситуаций.

Говоря об указанных выше структурах, Р. Шенк, Л. Бирнбаум и Дж. Мей [114] отмечают возможность их применения в интересах обработки языковых данных и построения концептуальных представлений, позволяющих выводить правдоподобные умозаключения. В данном случае уместно вернуться к аналитическому обзору С.М. Шевенко [111] и пояснить, что в указанном обозрении представлено различие между двумя базовыми типами понимания: т.н. «слабым» и «сильным» по-

ниманием; в то время как слабое, или языковое, понимание «состоит в извлечении из каждого высказывания содержащейся в нем языковой информации» [111, с. 15], сильное понимание определяется как умение построения гипотез о содержании последующих предложений в тексте, основанного на построении цепочек умозаключений. В процессе сильного понимания используются, по мнению автора, также энциклопедические сведения и аксиомы действительности. Представляется вполне очевидным, что предельно адекватное понимание текста фактически недостижимо без использования эвристических механизмов, основанных на описанных выше структурах представления знаний; иными словами, «полноценное понимание рассматривается как сочетание семантического и логического компонентов» [111, с. 16].

Говоря о практических возможностях использования фреймов при автоматическом анализе текстов, С.М. Шевенко замечает, что одной из базовых проблем, с которой сталкиваются разработчики систем машинного понимания, является проблема структурирования знаний о мире. Осуществить практическую реализацию теоретических разработок удается, по мнению автора, исключительно в системах, которые ориентированы на ограниченные предметные области, иначе – на определенные подъязыки. В узких предметных областях наиболее эффективным оказывается именно применение фреймов, позволяющих произвести настройку на события определенного рода. В частности, поясняет исследователь, потребность в экстралингвистической информации существенно велика в системах автоматизированного перевода, которые ориентированы на типологически различные языки – в частности, при анализе текстов на изолирующих языках, где формальные грамматические показатели минимизированы, движение по заранее заданной ситуационной рамке представляется автору единственно возможным эффективным подходом. С.М. Шевенко комментирует, что, хотя не существует общетеоретических данных о принципах отбора знаний для их представления в машине, вполне очевиден рост эффективности анализа по мере увеличения степени специализации априорных данных. При обработке текстов узкой тематики семантические классы слов, к примеру,

уже могут быть выведены с помощью формальной процедуры.

В качестве примеров систем, использующих фреймовое представление знаний, автор приводит диалоговые информационно-поисковые системы GUS, ДИЛОС, специализированные системы обработки рефератов научных публикаций. Кроме того, продолжает исследователь, на понятии фрейма основана система «понимания» естественного языка, предложенная С. Риджером; семантическое представление текстов осуществляется в данной системе с помощью графа, или семантической сети, отражающей значения и связи между ними. Актуальной автору представляется идея о расширении классического понятия фрейма и его экстраполяции на различные языковые явления – в качестве примера приводится концепция грамматического фрейма.

Тем не менее, исследователь считает необходимым заметить, что возможности фреймов при моделировании пресуппозиции являются ограниченными, поскольку они не моделируют способность человека понимать непредвиденные ситуации, извлекать из контекста окказиональные пресуппозиции – причиной этого является тот факт, что фрейм моделирует ассоциации по смежности, но не по сходству; кроме того, недостатком фреймов является их чрезмерная жесткость. В то же время фреймовый подход все же вносит, по мнению исследователя, определенную гибкость в машинный анализ, что и обусловливает его активное применение в современных теориях.

Действительно, представляется вполне очевидным, что современные исследователи уделяют существенное внимание теории фреймов. Согласно аналитическому обзору «Категории искусственного интеллекта в лингвистической семантике: фреймы и сценарии» [10], успешность применения фреймов в системах искусственного интеллекта при моделировании процессов понимания текстов на естественных языках явилось причиной активного теоретического и практического использования данной концепции в рамках лингвистической семантики. Основы указанной успешности автор обзора видит как в психологических, так и в лингвистических аспектах теории фреймов; в частности, отмечено, что «притягательность» понятия фрейма в лингвистической семантике связана как с близостью данного концепта к процедуре семантическо-

го анализа, так и с активной разработкой «фреймоподобных» концепций в истории лингвистики (см., в частности, статью Ч. Филлмора «Фреймы и семантика понимания» [123]: концепция «фундаментальных реляций» Дж. Милля, теория семантического поля И. Трира и др.). Основным содержанием описываемого обзора является последовательное изложение комплекса представлений о понятии фрейма в лингвистической теории, условно подразделяемого на 4 основных направления:

1) фреймвсоотношениистрадиционновыделяемымиязыковыми выражениями (слово, предложение, высказывание);
2) фрейм и типы информации (знаний), содержащихся в высказывании;
3) фрейм и языковая способность человека;
4) фрейм в изучении структуры текста.

Необходимо отметить, что, по мнению автора анализируемого обзора, фрейм может рассматриваться в качестве элемента метаязыка лингвистической семантики.

При изучении проблемы соотношения фрейма и слова исследователь обращается к упомянутой выше статье Ч. Филлмора; по мнению американского ученого, задача семантики фреймов в сфере лексики состоит в выявлении фрейма, стоящего за значением слова. Оппозиционирование фреймов может имплицитно содействовать пониманию некоторой информации, никаким образом не содержащейся в высказывании. А.Н. Баранов также считает необходимым привести предположение, высказанное в статье Т. Хофманна «Семантика фреймов и представление содержания» [125], согласно которому фреймы могут не входить непосредственно в значение слова, функцией же лексем может быть лишь активация необходимых для понимания высказывания фреймов. В статье отмечается, что последовательность интерпретации высказывания определяется несколькими базовыми факторами: контекстом (идеальным. – Н.Г.), фреймом (или сценарием), физическим окружением, а также памятью. По мнению Т. Хофманна, активации фрейма может предшествовать процедура выбора конкретного, наиболее соответствующего ситуации, фрейма среди многих, близких ему по содержанию, в связи с чем вводится т.н. принцип максимального пересече-

ния. Исследователь считает важным отметить, что успешное использование понятия фрейма как нового элемента метаязыка лингвистической семантики требует установления четких границ между ним и другими, уже существующими категориями.

Говоря о соотношении фрейма с предложением и / или высказыванием, автор комментирует, что данная тема не является существенно разработанной – семантика предложения видится исследователям в рассматриваемой области через посредство лексических фреймов. В частности, вновь упоминаются исследования Ч. Филлмора в сфере понимания высказываний, содержащих отрицание; кроме того, представляется подробное изложение идей Ф. Кифера, представленных в его статье «Как объяснить феномен ситуационного значения?» [128]. Исследователь подразделяет потенциальные значения предложений на буквальные и небуквальные, каковым сопутствует соответственно нейтральный и специфический контекст; буквальное значение моделируется автором в виде функции $LM(t) = F(B(E), c(t))$ (т.е. буквальное значение может быть определено по функции F, аргументами которой являются логическая форма языкового выражения B(E) и нейтральный контекст c(t)), небуквальное – в виде функции $NM(t) = MET(LM(t), c^1(t))$ (т.е. небуквальное значение зависит от буквального, а также от специфического контекста $c^1(t)$). Ситуационное значение предложения представляется исследователем в виде формулы $SM(t) = H(LM(t), Script_1)$, за счет чего оно ставится в зависимость от буквального значения и соответствующего предложению сценария. В целом предлагаемый Ф. Кифером подход предполагает замену традиционной бинарной оппозиции «семантика – прагматика» на четырехчленное противопоставление «логическая форма – буквальное значение – значение высказывания – коммуникативный смысл».

Обращаясь к вопросу о соотношении фрейма и типов знаний, необходимых для понимания естественноязыковых сообщений, А.Н. Баранов поясняет, что одним из основных вопросов, возникающих при описании языка, является вопрос о том, что в слове относится непосредственно к знанию языка и что – к знанию о мире, иначе – в какой степени при опи-

сании значений слов необходимо отражать лингвистическую и энциклопедическую информацию. В трудах, представленных в обзоре, фрейм рассматривается либо как инструмент, позволяющий привлекать для понимания естественноязыковых сообщений необходимую для этого часть экстралингвистической информации (Т. Хофманн), либо как структура, идентичная семантическому представлению содержания языковой единицы, что уравнивает лингвистическую информацию и энциклопедические знания (Р. Хадсон, Р. Шенк), либо как принципиально нелингвистическая категория, непригодная для лингвистического описания языковых структур (Й. Уилкс).

Взаимосвязь фрейма и языковой способности человека раскрывается исследователями через его соотнесение с пониманием, рассуждением и процессом коммуникации. В частности, Ч. Филлмор находит целесообразным различать «понимающую» семантику и семантику «истинности», поясняя, что основой семантики понимания является т.н. «фрейм интерпретации», а базовой целью – определение связи между языковыми текстами и возможностью их полного понимания в конкретной ситуации. В вопросе о естественноязыковом рассуждении автор обзора отмечает статью Х. Ыйма и М. Салувеера «Фреймы в лингвистических описаниях» [129]; по мнению исследователей, использование фреймов и иных подобных коммуникативных категорий изменяет объект традиционного лингвистического анализа на «коммуникативную компетенцию» носителя языка, включающую в себя лингвистическую, когнитивную и интерактивную компетенции. Соответственно, фрейм представляется авторами статьи в качестве инструмента изучения коммуникативной компетенции; в рамках фрейма выделяется три базовых компонента – SETTING (условия, окружение, физический и идеальный контекст), EVENT (событие), CONSEQ (последствия действий или процессов). ; структуризация фрейма иллюстрируется исследователями на примере слова *lend* (A – агент, B – реципиент, C – объект, t - время).

LEND
SETTING
(a) До t_i A владел C ;
(b) B сообщает A свое желание иметь C в течение некоторого $t_i – t_j$;
(c) B обязуется вернуть C A после t_j ;
EVENT
(d) A дает C B в t_i ;
CONSEQ
(e) B владеет C после t_i ;
(f) B обязан вернуть C A после t_j ;
(g) DEFAULT : B возвращает C A после t_i .

Между компонентами фрейма и собственно словом *lend* существуют определенные взаимоотношения. К примеру, поясняют авторы, пропозиции (d) – (g) связаны с lend отношением импликации (A → B & !A → !B), пропозиции (a) – (c) связаны с данным словом отношением пресуппозиции (A → B & !A → B) или утверждения (A → B & !A → B OR !A → !B); исходя из представленных соотношений, исследователи определяют операции над фреймами, которые могут быть использованы при моделировании естественноязыковых рассуждений:

$$1)\ FRAME_i \xrightarrow{i} EVENT(F)_i\ ;\ 2)\ FRAME_i \xrightarrow{i} CONSEQ(F)_i\ ;$$

$$3)\ FRAME_i \xrightarrow{a} SETTING(F)\ ;\ 4)\ FRAME_i \xrightarrow{pr} SETTING(F)_i\ .$$

Соотношение фрейма и коммуникации представлено автором обзора на материале статьи Д. Таннен «Фреймы и схемы в коммуникативном взаимодействии» [133]. В указанном исследовании отмечено, что, к примеру, коммуникативное поведение врача может быть связано с двумя базовыми интерактивными фреймами – «фреймом осмотра» и «фреймом консультирования». Конфликт между данными фреймами, по мнению автора, может являться причиной неадекватных ответов врача, если во время осмотра ему задаются вопросы, уместные в рамках фрейма консультирования, либо наоборот. Таким образом, задача семантики фреймов в исследовании коммуникации может состоять в выявлении интерактивных фреймов, определяющих структуру коммуникативных ожиданий участников общения, резюмируют авторы обзора.

Завершая рассмотрение лингвистического аспекта теории фреймов вопросом о роли фрейма в изучении структуры текста, исследователи обращаются к соответствующим идеям Й. Уилкса и Ч. Филлмора; в частности, Й. Уилкс, анализируя вопрос о степени адекватности представления структуры текста при помощи фреймов и подобных им когнитивных категорий, приходит к выводу, что последние несущественно отличны от «синтаксических грамматик» текста, ориентированных на порождающую модель и характеризуемых правилами типа TEXT – SETTING – CHARACTERS – PLOT – CONCLUSION. По мнению ученого, алгоритмы, составляющие основу текстового фрейма, недостаточны для адекватного поведения (и, соответственно, понимания) в той или иной ситуации; с помощью фрейма может быть понят конкретный текст, но это не является неоспоримым свидетельством адекватного представления фреймом его семантической структуры. Фактически, комментирует Й. Уилкс, фреймы такого рода являются «неприменимыми планами» [135].

Возвращаясь к обзору С.М. Шевенко [111], обозначим ряд проблем, вызывающих обоснованный скептицизм специалистов в отношении потенциальных возможностей систем машинного понимания текстов.

Автор отмечает, что машина способна лишь имитировать понимание за счет определения и анализа значимости слов, никаким образом не переходя на уровень смысла (хотя даже такая операция требует привлечения контекста); помимо этого, данных лексической и синтаксической семантики оказывается, по мнению исследователя, недостаточно – для полноценного понимания необходима также и прагматика, т.е. пресуппозиция. При этом описание семантики производится при помощи дискретной логики, в то время как системе нередко требуется обрабатывать размытые понятия (для описания которых автор предлагает обратиться к аппарату нечетких множеств).

Кроме того, С.М. Шевенко обозначает важную проблему отсутствия «при построении систем «понимания» фундаментальных лингвистических принципов системного характера» [111, с. 53]. В связи с этим автор обзора цитирует общую оценку систем понимания, которая была дана Дж. Вейценбаумом;

по мнению последнего, разработчики систем машинного понимания принципиально заблуждаются относительно природы обрабатываемых ими задач, будучи не в состоянии осознать, что для достижения поставленных целей им необходимы более сильные общие принципы, нежели утверждение о возможности представления тех или иных законов в виде алгоритмов и программ. Вследствие этого разработчики подобных систем полагают, что перспективы ИИ не были реализованы в их машинах исключительно по причине ограниченности возможностей данных машин. Соответственно, заключает С.М. Шевенко, большинство работ в рассматриваемой автором обзора области характеризуется применением исключительно структурных методов исследования вне единого системного подхода к рассматриваемым явлениям. К примеру, аппарат логического вывода – один из базовых компонентов имитации умозаключений в искусственном интеллекте – способен к эффективному функционированию лишь в условиях наличия полного набора аксиом действительности, что является принципиально некорректным подходом к организации базы данных. Поэтому слабость современных систем понимания, резюмирует автор, состоит в отсутствии системно-лингвистического фундамента, результатом чего становится излишне активное использование ad hoc-правил.

Подводя итог сказанному, отметим, что последние разработки в области лингвистического обеспечения автоматизированных систем обработки текстов преимущественно связаны с развитием технологий искусственного интеллекта и совершенствованием возможностей такового. Вполне очевидно, что ИИ может находить применение во всех разновидностях систем обработки естественноязыковых текстов в силу своей способности к эвристическому анализу и моделированию умозаключений; соответственно, базовой проблемой построения автоматизированных систем обработки текстов на основе ИИ является вопрос о представлении значений, знаний и смыслов. Большинство исследователей склонны полагать, что наиболее эффективным и эргономичным решением данной проблемы может являться применение когнитивных структур – фреймов, сценариев, полей и т.д. Хотя использование фреймов интуитивно понятно и может быть согласовано

с психологическими представлениями о человеческом мышлении, полноценное моделирование психических и языковых процессов при помощи данных структур все еще не может быть осуществлено, что обусловливает определенный скептицизм ряда ученых по отношению к перспективам применения фреймов в АСОТ.

Необходимо также отметить, что вопрос о полноценном машинном моделировании тех или иных психических либо языковых процессов со всей очевидностью может быть поставлен лишь после того, как будет достигнуто полное понимание самих указанных процессов в психологии, в лингвистике и на стыке данных наук. На данный момент, вероятно, о полном понимании говорить не приходится – исследование психических и языковых процессов нередко имеет характер саморефлексии, что традиционно не гарантирует истинности полученного результата.

Тем не менее, надлежит указать, что достижения современной науки позволяют создавать довольно успешные образцы «понимающих» машин, способных к восприятию поступающей извне речевой информации и к ее эффективной обработке – примером тому может служить СК IBM Watson. Ряд исследователей (Дударь, Шуклин 2000; Шуклин 2001; Шевелев 2006) указывают на потенциально высокую эффективность нейронных сетей, которые могут выступать в качестве первоосновы систем искусственного интеллекта нового поколения; не исключено, что передовые разработки в указанном направлении позволят повысить производительность и качество работы систем автоматизированной обработки текстов, приблизив их возможности к способностям человеческого мозга. Механизмы нечеткой логики, лежащие в основе упомянутых выше нейросетей, теоретически способны обеспечить «человекоподобное» мышление, систематизацию данных, совершение умозаключений и формирование выводов, что вкупе со способностью машины к самообучению и накоплению произвольно объемного количества данных может обеспечить решение ряда существенных проблем, с которыми в настоящее время сталкиваются разработчики систем автоматизированного анализа естественноязыковых текстов. Тем не менее, необходимо отметить, что перечисленные перспективные

разработки и прогресс достижений вычислительной техники никоим образом не снимают вопросов о необходимости оптимизации АСОТ, рассматриваемых в настоящем исследовании.

1.6. ВЫВОДЫ

Изучение вопроса о природе значения языкового знака позволило акцентировать внимание на прагматическом аспекте языковой единицы, постулируя тем самым контекстную обусловленность ее значения, а также подтвердить необходимость учета взаимосвязи, существующей между семантикой, синтактикой и прагматикой знака, и, следовательно, зависимости значения языкового знака от его сочетаемости с другими знаками в высказывании и от его практического употребления, от дополнительных значений, возникающих в ситуации использования знака. Хотя семантика языкового знака принципиально может изучаться в отрыве от сопутствующих ей синтаксического и прагматического аспектов, в целях максимально полного понимания значения было признано целесообразным рассматривать языковой знак как в физическом, так и в идеальном контексте.

Рассмотрение типологии лексической многозначности позволило утверждать, что последняя должна быть определена через понятие сходства значений. В качестве основного определения принимается формализованная дефиниция, предложенная Ю.Д. Апресяном – «слово А называется многозначным, если для любых двух его значений a_i и a_j найдутся такие значения a_1, a_2, ..., a_k, a_l, что a_i сходно с a_1, a_l – с a_2 и т.д., a_k с a_l и a_l –с a_j». Поскольку определение многозначности осуществляется в терминах теории множеств, допустимо с уверенностью говорить о возможности успешного математического моделирования явления полисемии и структуры значений полисемичного слова.

Установлено, что, несмотря на активное употребление термина «формально-семантический анализ» в современных исследованиях, не существует достаточно четкого определения настоящего термина. Рассмотрение основных представлений о формально-семантическом анализе позволило суммировать разрозненные дефиниции в целях выработки единого опре-

деления и предложить таковое в следующей формулировке: формально-семантический анализ – это процедура выявления и интерпретации эксплицитно выраженных характеристик языковых знаков, которые позволяют составить представление о ядерных элементах и структуре их семантики, и описания полученных данных посредством формул, пригодных для использования в построении математических моделей процессов понимания значений языковых знаков как в пределах физического и идеального контекста, так и вне его.

Рассмотрение современных разработок в области лингвистического обеспечения АСОТ позволило заключить, что они ориентированы преимущественно на достижения в области искусственного интеллекта, нейронных сетей, исследования в сфере теории фреймов и на построение семантических метаязыков. В класс последних включены информационные языки, что позволяет говорить об использовании в информационной работе всего спектра теоретических и методических средств, актуальных для решения задач прикладной семантики.

К основным типам АСОТ относятся такие базовые их разновидности, как поисковые системы и каталогизаторы, средства понимания сообщений на естественных языках, а также машинные переводчики. Для всех указанных типов актуальны вопросы повышения их основных показателей и характеристик: точности, скорости, производительности функционирования. При этом требуют решения также и вопросы о разработке достаточно надежного лингвистического обеспечения подобных систем, поскольку нередко их проектирование и построение становится задачей специалистов в иных областях знания, не имеющих непосредственного отношения к языковедению и коммуникативистике.

В результате был сделан вывод о том, что существует потребность в реализации комплексного подхода к автоматизированной обработке текстов, в формировании надежного лингвистического обеспечения для соответствующих алгоритмов, а также в улучшении качества и производительности основных типов АСОТ; решение актуальных проблем автоматизированного анализа естественноязыковых текстов видится в построении оптимизационных алгоритмов, учитывающих вышеупомянутые потребности.

2. ВОЗМОЖНОСТИ ОПТИМИЗАЦИИ АСОТ ПОСРЕДСТВОМ ФУНКЦИОНАЛЬНО-СТИЛИСТИЧЕСКОЙ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

2.1. ФУНКЦИОНАЛЬНЫЙ СТИЛЬ КАК ФАКТОР ОПТИМИЗАЦИИ АВТОМАТИЗИРОВАННОГО АНАЛИЗА ТЕКСТА

Согласно изложенным выше базовым представлениям о процедуре текстового анализа и степени адекватности его лингвистического обеспечения, в настоящее время существует потребность в оптимизации систем автоматизированного анализа естественноязыковых текстов. Действительно, в связи с активным развитием сети Internet и постоянным ростом объемов информации, которая подлежит обработке, наблюдается повышенный интерес к возможностям разработки новых и совершенствования существующих средств анализа естественноязыковых текстов; при этом первоочередную значимость приобретают оптимизированные эвристические алгоритмы, позволяющие сколь возможно быстро обрабатывать значительные объемы сведений и не требующие при этом организации и ведения громоздких баз данных (БД).

Как было отмечено нами ранее, сфера автоматизированного анализа текстов на естественном языке включает три основных раздела: машинный перевод, автоматическое «понимание» естественного языка в целях выработки оптимального человеко-машинного интерфейса, а также теория и практика информационно-поисковых систем. Вполне очевидно, что обозначенные сферы достаточно разноплановы, и задача оптимизирования тех или иных систем автоматизированного анализа текстов представляет определенные сложности в поиске унифицированных алгоритмов.

Отметим, что нам представляется целесообразным формировать именно унифицированный алгоритм оптимизации, который мог бы в равной степени удовлетворять потребности

автоматизированных систем, разработанных для выполнения задач в каждом из трех упомянутых выше разделов, поскольку разработка программных модулей, построенных на единых принципах и подходах к оптимизации систем автоматизированного анализа текстов на естественных языках, позволяет успешно реализовать тот системный подход к организации машинных анализаторов, о насущной потребности в актуализации которого говорилось в представленном выше исследовании С.М. Шевенко. В связи с этим приоритетной задачей в процессе построения средств оптимизации автоматизированного анализа представляется поиск тех параметров естественноязыковых текстов, которые в равной или близкой степени значимы как для машинных переводчиков, так и для средств «понимания» и информационно-поисковых систем, а также характеризуются определенным эвристическим потенциалом.

На наш взгляд, тот факт, что вопрос интерпретации высказываний на естественном языке самым непосредственным образом увязан с вопросами семантики языковых знаков, является вполне очевидным. Действительно, корректное интерпретирование текста едва ли может быть осуществлено, если некоторые его составляющие – лексические единицы, сверхфразовые единства, предложения, высказывания, - не поняты или истолкованы неверно; отметим, что одним из базовых проблемных факторов, создающих сложности при интерпретации языковых знаков, является фактор их полисемантичности – вряд ли вызывает какие-либо сомнения то утверждение, что вопрос выбора того или иного значения из некоторого множества таковых может представлять определенные сложности для участников коммуникации и является трудоемкой задачей для автоматизированного анализатора.

Совершение корректного выбора в пользу того или иного значения полисемантичного языкового знака может быть обусловлено приданием определенного веса тому или иному члену множества значений. Основным фактором, который позволяет осуществить обозначенный выбор, является физический и / или идеальный контекст, о чем было сказано выше; в современных исследованиях, посвященных вопросам применения контекста в процессе автоматизированного анализа текстов, негласный приоритет отдается контексту физиче-

скому, т.е. непосредственному окружению языкового знака. Закономерным итогом указанной тенденции стало преимущественное использование дистрибутивного анализа в целях совершения выбора того или иного члена из множества потенциальных значений полисемантичного языкового знака.

Отметим, однако, что в рамках комплексного подхода к оптимизации процедур автоматизированного анализа естественным образом подлежат учету все факторы, оказывающие непосредственное воздействие на интерпретацию текста, что предопределяет необходимость изучения и применения в интересах указанной оптимизации как физического, так и идеального контекста. С данной позиции формально-семантический анализ как метод в полном соответствии с предложенным нами выше определением может рассматриваться как эффективное средство оптимизации, поскольку позволяет формировать представление о значении как в пределах физического и идеального контекста, так и вне его.

Под идеальным контекстом мы предлагаем понимать ситуацию общения в наиболее широком понимании данного термина, т.е. множество условий и факторов (по преимуществу экстралингвистического характера), в рамках или под воздействием которых осуществляется коммуникация, равно как и тех фоновых знаний, которые учитываются коммуникантами в процессе общения; с целью упростить дальнейшее изложение условимся полагать синонимичными понятия «идеальный контекст», «коммуникативная ситуация» («ситуация общения») и «речевая ситуация».

Действительно, полноценная интерпретация сообщений, посылаемых адресантом текста, будь то текст письменный или же устный, представляется принципиально нерешаемой задачей при отсутствии у реципиента совокупности представлений об идеальном контексте высказывания, сколь возможно более адекватной аналогичным представлениям отправителя сообщения. В случае дистантной письменной коммуникации данный фактор представляется особенно значимым в силу постоянного устаревания коммуникативной ситуации и фоновых знаний, что вызывает потребность, к примеру, в пространном комментировании литературных произведений, с момента создания которых прошло 50-100 лет и более.

Соответственно, в интересах наиболее эффективной интерпретации текстов на естественных языках в процессе их автоматического анализа разрабатываемое лингвистическое обеспечение должно учитывать как физический, так и идеальный контекст исследуемого высказывания или текста. Не касаясь неплохо разработанных в языкознании вопросов анализа физического контекста, мы считаем необходимым обратиться к рассмотрению идеального контекста, его компонентов, степени критичности в вопросах выбора единичных значений из множества таковых, а также аналитического потенциала вышеуказанных компонентов с точки зрения оптимизации процедур автоматизированного анализа текстов.

Краткий анализ существующих воззрений на структуру и компоненты идеального контекста позволяет установить, что не существует единого подхода к выделению вышеуказанных компонентов. В частности, А.В. Ленец при рассмотрении речевой ситуации (РС) лжи считает необходимым выделить такие составляющие РС, как «участники общения, место и время общения, предмет общения, цель общения, обратная связь между участниками общения» [139], а Г.Г. Матвеева предлагает включать в состав РС «условия общения, предметный ряд, время и место коммуникации, самих коммуникантов, их отношение друг к другу и др. признаки» [140], при этом отмечая, что восходящие к К. Бюлеру классификации неизменно содержат три базовых элемента – отправитель, получатель и предмет речи, – однако в деталях могут быть существенно различны в том числе в рамках одной и той же работы. Г.Г. Матвеева также находит уместным активное использование классификации, предложенной Д. Хаймсом [102]; согласно указанным воззрениям, могут быть выделены семь компонентов, или факторов, речи, как то: отправитель, получатель, форма сообщения (речевой жанр), канал связи (или вид речевой деятельности - устный/письменный), код (язык и норма словоупотребления), тема (функциональный стиль, подъязык) и обстановка. В работе Г.Г. Матвеевой также отмечается, что именно в соответствии с данной классификацией Р. Якобсоном [121] были выделены функции, или роли, тех или иных составляющих речевой ситуации. Мы также находим предпочтительным обратиться к рассмотрению элементов РС в том

составе, который предложен Д. Хаймсом, с позиции получателя сообщения (т.е. собственно интерпретатора).

1) Отправитель (адресант). Представляется вполне очевидным, что в задачи интерпретатора входит активное использование сведений, позволяющих производить некоторые умозаключения об отправителе сообщения; в то же время нам не известен опыт привлечения данных об авторе исследуемого текста в интересах автоматизированного анализа. Изучение информации об адресанте непосредственным образом увязано с определением и актуализацией фоновых знаний, которые наряду с физическим окружением (местом, временем и предметами объективной реальности) составляют содержание также и последнего компонента классификации Д. Хаймса («обстановка»).

В то же время знания об отправителе сообщения по преимуществу не могут быть извлечены эвристически и требуют формирования и поддержки БД, что не позволяет обратиться к исследованию данного фактора в целях выработки оптимизированных алгоритмов автоматического анализа текстовых сообщений.

2) Форма сообщения (речевой жанр). В рамках излагаемой концепции под речевым жанром предлагается понимать конкретные формы текстовых сообщений, существующих в пределах одного функционального стиля речи (как то: лекция, доклад, реферат, статья, аннотация в пределах научного стиля) или того или иного типа творческой деятельности (литературные жанры). Данные формы обладают, как следует из их наименования, определенной совокупностью исчислимых формальных признаков и, соответственно, характеризуются некоторым эвристическим потенциалом; однако выбор речевого жанра, очевидно, не оказывает сколь-либо существенного воздействия на значение тех или иных языковых единиц – а, следовательно, данный фактор не может рассматриваться как источник оптимизации интерпретирующих алгоритмов. Действительно, сложно представить, чтобы один и тот же языковой знак мог характеризоваться различными значениями в пределах лекции и в пределах реферата; отличать же упомянутую выше лекцию от романа или от газетной статьи представляется неуместным.

3) Канал связи (вид речевой деятельности). Данный элемент, как следует из рассматриваемой нами концепции, определяет, каким именно образом и в какой среде передается то или иное текстовое сообщение. Отметим, что определение канала, по которому передается активный сигнал, осуществляется любым интерпретатором, будь то человек или автоматизированный анализатор, самым естественным образом, поскольку восприятие сведений по устному и письменному каналам производится посредством принципиально различных рецепторов. В сущности, вопросы собственно машинного *восприятия* устных высказываний, безусловно, представляют определенную сложность, равно как и интерес для исследователя, однако указанные проблемы не имеют непосредственной связи с интерпретацией содержания анализируемого текста и в силу этого остаются вне пределов настоящего исследования. Нет также и оснований утверждать, что канал связи сам по себе оказывает какое-либо воздействие на значения языковых знаков; соответственно, данный фактор не может служить источником оптимизации рассматриваемых нами аналитических алгоритмов.

4) Код (собственно язык и норма словоупотребления). При анализе компонентов речевой ситуации принято полагать, что коммуниканты разделяют общий код, на котором осуществляется их речевое взаимодействие; в противном случае коммуникация принципиально невозможна. Следует, однако, заметить, что при человеко-машинном взаимодействии, либо при взаимодействии двух человеческих коммуникантов посредством машины, коммуникация посредством общего языка в настоящее время места не имеет вообще (либо подобные случаи составляют определенную редкость), поскольку развитие операционных систем и прикладного программного обеспечения позволило устранить необходимость непосредственного использования машинных команд человеком. Соответственно, каждая машина, задействованная в процессе автоматического анализа текстов на естественном языке, в сущности является не только получателем сообщения, но также одновременно и переводчиком, осуществляя конверсию сообщений из одного кода в другой и обратно.

Рассматривая вопросы корректности указанной конвер-

сии, а также зависимости значений языковых единиц и содержания текста в целом от кода, на котором осуществляется коммуникативное взаимодействие, представляется возможным прийти к заключению, что в данном случае ведущим интерпретационным фактором со всей очевидностью является собственно язык источника, т.е. код, на котором изъясняется отправитель. В настоящее время системы автоматизированного анализа либо одноязычны (т.е. изначально построены в расчете на исследование сообщений только на одном языке), либо снабжены эффективными алгоритмами определения языка отправителя; подобные алгоритмы успешно апробированы в ежедневной практике работы информационно-поисковых систем Интернета, и какие-либо дополнительные исследования в этой области представляются нам нецелесообразными.

5) Тема (функциональный стиль, подъязык). Согласно Р. Якобсону, речевая ситуация характеризуется референциальностью, т.е. отнесенностью к некоторой предметной области [121]. В пределах некоторой области знания или социального взаимодействия может существовать специфическая разновидность языка, отличная от иных его модификаций и характеризующаяся некоторыми формальными и содержательными признаками; иными словами, продуцируемый текст формируется в рамках некоторого функционального стиля речи.

В настоящее время не вызывает сомнений тот факт, что существует определенная взаимосвязь между значениями языковых единиц и функционально-стилевой принадлежностью текста, в состав которого они входят. В частности, многие современные электронные словари-переводчики снабжены набором специализированных словарей, которые содержат отдельные элементы из совокупности значений тех или иных слов, актуализируемые в рамках различных стилей речи; следует, однако, заметить, что в изученных нами программных продуктах выбор специализированных словарей необходимо осуществлять вручную, в то время как данный процесс потенциально может быть автоматизирован. На наш взгляд, автоматическое определение функционально-стилевой принадлежности текстов может быть равнозначно ценно для всех трех типов автоматических анализаторов. В то же время функци-

ональный стиль речи, как было отмечено выше, обладает некоторой характерной исключительно для него совокупностью формальных и содержательных параметров, которые, с одной стороны, исчисляемы, а с другой стороны, в перспективе могут обладать эвристическим потенциалом.

В силу указанных выше причин мы находим возможным обратиться к анализу формально-содержательных признаков функциональных стилей речи, а также предпринять попытку изыскания тех их параметров, которые отвечали бы ряду базовых требований, как то:

- реализация комплексного подхода к разработке и оптимизации аналитических алгоритмов;
- возможность эффективного детектирования функциональных стилей и автоматического их различения;
- принципиальная исчисляемость указанных параметров, т.е. наличие возможности успешно оценивать их без участия человека;
- наличие эвристического и оптимизационного потенциала, позволяющего реализовать высокопроизводительные алгоритмы, не требующие организации и ведения объемных БД.

Необходимо отметить, что определенным эвристическим потенциалом характеризуется не только функциональный стиль текста, но и те сведения о ситуации, которые могут быть извлечены непосредственно из текстового сообщения. Фактически данные сведения представляют собой компонент фоновых знаний, эксплицитно выраженный в исследуемом тексте; исследования, упомянутые нами выше, продемонстрировали, что наличествующие в тексте сведения могут быть достаточны для успешного формирования некоторых представлений о контексте и даже для понимания неявных ссылок и аллюзий. Разработки в данной области также были довольно успешными: выше нами описан ряд автоматизированных анализаторов, разработанных и построенных американскими исследователями в 1970-80 гг.; опыт практического использования данных систем, равно как и изучение доступных информационных материалов о современных средствах извлечения и анализа фактической информации, позволяют утверждать, что в данной сфере научных изысканий достигнут

достаточно высокий уровень качества анализа, и какие-либо дополнительные разработки, в том числе оптимизационные, не являются насущно востребованными.

Очевидно, что, переходя в сферу стилистики, необходимо в первую очередь определить ряд базовых концепций и терминологических установок, которыми мы будем руководствоваться в дальнейшем изложении материала. Естественным представляется оговорить в настоящем параграфе в первую очередь само понятие функционального стиля, равно как и избрать действующую классификацию ФС, каковой мы станем придерживаться.

Вслед за М.Н. Кожиной [43] мы принимаем в качестве основной дефиницию функционального стиля речи как своеобразного характера речи той или иной социальной ее разновидности, соответствующей определенной сфере общественной деятельности и соотносительной с ней форме сознания, создаваемого особенностями функционирования в этой сфере языковых средств и специфической речевой организацией. При этом считаем важным отметить, что «стиль — это свойства, признаки содержательного (смыслового) плана, обусловленные экстралингвистически, получающие реализацию через своеобразное построение (систему) материальных элементов — своеобразные «комбинации» языковых единиц, особое построение речи (текста), выступающее в качестве плана выражения» [43, с. 86]. Согласно цитируемому исследованию, таким образом манифестируется тот факт, что форма и содержание в широком понимании данных терминов характеризуются системностью и единством; специфика мышления коммуниканта при этом выражается в особенностях собственно речи, а не посредством некоторых внеконтекстуальных факторов.

Также мы считаем необходимым согласиться с мнением М.Н. Кожиной о том, что термин «функциональный стиль» может равноуспешно применяться для описания как стиля языка (т.е. инварианта), так и стиля речи (т.е. его конкретной реализации). Наши теоретические рассуждения, касающиеся типических признаков различных функциональных стилей, безусловно, будут основаны на анализе абстрагированного инварианта, однако собственно материал, с которым мы бу-

дем работать (и который предъявляется автоматизированным системам обработки текстов в ситуациях их реального использования), может рассматриваться исключительно как совокупность конкретных реализаций (т.е. стилей речи); тем не менее, термин «функциональный стиль» сможет быть применен нами в обоих случаях.

Касаясь вопроса о классификации функциональных стилей, отметим, что его решение непосредственным образом зависит от оснований классификации. Известно разделение стилей в непосредственном соответствии с функциями, которые они выполняют, предложенное В.В. Виноградовым [22]; согласно указанной концепции, надлежит выделять функцию общения (разговорный стиль), функцию сообщения (деловой и научный стили), а также функцию воздействия (художественный и публицистический стили). Необходимо признать, что рассматриваемый комплекс представлений не может быть признан успешным с точки зрения собственно классификации, т.к. не дает оснований для отделения делового стиля от научного и художественного – от публицистического; тем не менее, заметим, что идеи В.В. Виноградова представляются нам довольно успешной попыткой типизации функциональных стилей, позволяющей объединять их в группы в соответствии с определенными характерными особенностями.

По мнению М.Н. Кожиной, надежным основанием для классификации в данном случае могут послужить разнообразные формы общественного сознания, «тем более, что каждой из них соответствует определенный вид деятельности и социально значимая, типизированная и традиционная сфера общения со сложившимися принципами употребления языковых средств, образующих своеобразную стилистико-речевую организацию» [43, с. 123]. Рассматриваемые формы общеизвестны: наука, право, политика, религия, искусство, а также обыденное (бытовое) сознание; таким образом, на указанном основании могут быть выделены 6 функциональных стилей – научный, официально-деловой, публицистический, художественный, религиозный и разговорно-обиходный; данная классификация характеризуется обоснованностью и логичностью и в силу этого может быть принята нами в качестве основной.

Особого рассмотрения требует также вопрос о статусе художественного стиля, поскольку ряд исследователей (В.Д. Бондалетов, О.А. Крылова, Л.Ю. Максимов и др.) не усматривает оснований для его выделения в особый функциональный стиль. По мнению упомянутых ученых, язык художественной литературы характеризуется открытостью, полистилистичностью, отсутствием каких-либо специфических маркеров, а также исполняет особую функцию (эстетическую). Мы, однако, согласны с позицией, в соответствии с которой художественную речь надлежит рассматривать в качестве одного из стилей русского языка.

Действительно, исполнение художественным текстом эстетической функции не отменяет выполнения им коммуникативной функции, а, следовательно, не формирует достаточно надежных оснований для отнесения его к неким принципиально особым явлениям, не имеющим ничего общего с функциональной стилистикой. «Выведение художественной речи за пределы функциональных стилей обедняет наше представление о функциях языка, создает ложное мнение о том, что эстетическая функция ему не присуща. Между тем использование языка в эстетической сфере — одно из высших достижений литературного языка, и от этого ни литературный язык не перестает быть самим собой, используясь в художественном произведении, ни язык художественной литературы не перестает быть проявлением литературного языка», - отмечает, в частности, М.Н. Кожина [43, с. 141]. Более того, эстетическая функция может рассматриваться именно как специфический показатель, объединяющий художественные тексты и выступающих в качестве стилеобразующего признака.

Что же касается полистилистичности, то мы полагаем целесообразным поддержать мнение о том, что художественная речь задействует исключительно отдельные элементы иных функциональных стилей, не приобретая при этом их особенных отличительных черт и не обслуживая в полном смысле слова потребностей соответствующих им форм общественного сознания. Упомянутые элементы при этом используются опять же в эстетической функции.

Таким образом, вопрос об определении базовой терминологии и основополагающих концепций, которых мы будем

придерживаться в наших теоретических рассуждениях, можно считать разрешенным.

Прежде чем продолжить изложение материала, следует прокомментировать, что при построении оптимизационного алгоритма на основе детектирования функционального стиля текста существует потребность в предварительном определении собственно основного назначения подобного алгоритма и наиболее общих принципов его работы. Представляется очевидно необходимым организовать алгоритм таким образом, чтобы автоматическое определение функционального стиля оказывало сколь возможно более эффективное воздействие на процесс истолкования значений языковых единиц. Обращение к словарю (Современный толковый словарь русского языка / Под ред. С.А. Кузнецова. - М., 2004) и краткий обзор представленных в нем статей позволяют утверждать, что в пределах совокупности значений полисемантичного слова с позиций функционального стиля наиболее часто проводится значимое различие между понятиями «книжное» и «разговорное»; эмпирический материал, равно как и интуитивные представления о взаимозависимости между значением слова и функциональным стилем, также заставляет в первую очередь задумываться об отличении книжных стилей речи от разговорного. К примеру, может не быть существенных различий в истолковании слова *линейка* (скажем, в значении ‘канцелярская принадлежность, служащая для измерения длины отрезков’) в пределах различных книжных стилей, однако значение, приписываемое указанному слову в современной разговорной речи, может существенно отличаться от соответствующих «книжных» истолкований (в частности, в последние несколько лет указанное слово может относиться к компьютерной игровой программе «Lineage»). В силу обозначенных причин может представиться целесообразным сфокусировать потенциальный алгоритм оптимизации, равно как и собственно процесс разработки его лингвистического обеспечения, в первую очередь на изыскании маркеров, позволяющих осуществить отличение книжных стилей от разговорного.

Тем не менее, существует и ряд факторов, которые препятствуют подобному ориентированию аналитических алгоритмов. С одной стороны, к таковым необходимо причислить

отсутствие сколь-либо надежных источников, которые могли бы поставлять достоверные образцы современного разговорного стиля речи, имеющие материальную реализацию в письменной форме; с другой стороны, абсолютное большинство современных автоматизированных систем обработки текстов работает в сферах, которые не соприкасаются с разговорным дискурсом в эталонном понимании данного термина.

Комментируя изложенные в предыдущем абзаце установки, считаем необходимым заметить, что мы придерживаемся представлений о разговорном стиле речи, согласно которым образцы указанного стиля существуют исключительно в контексте прямого аудиовизуального общения между коммуникантами. В тех случаях, когда разговорный текст получает письменное отражение (чаты, форумы, блог-записи в Интернете, личная переписка посредством почты, пейджеров, служб передачи коротких сообщений), он неизбежно подвергается предварительной обработке, что может приводить к утрате им своих специфических свойств и особенностей. Это, в свою очередь, означает, что задача определения разговорного дискурса стоит исключительно перед узким кругом анализаторов, имеющих дело с распознаванием и пониманием звучащей речи, в то время как иные автоматизированные системы взаимодействуют главным образом с книжными стилями.

В силу этого мы полагаем необходимым отложить вопрос о распознавании образцов разговорного стиля вплоть до момента широкого распространения качественных и надежных анализаторов аудиоинформации и сосредоточиться на исследовании, разработке и реализации процедур различения тех или иных не-разговорных стилей либо их групп.

Под упомянутыми группами мы предлагаем понимать объединение нескольких различных функциональных стилей на основании глобальных общих признаков. К примеру, книжные стили могут быть подразделены на формальные (т.е. предъявляющие строгие требования к внешним и внутренним характеристикам текстов) и свободные (т.е. допускающие слабо формализованное или неформальное изложение информации). Подобное разграничение возвращает нас к типизации функциональных стилей на основании исполняемых ими функций: сообщения (что предполагает доминирование ин-

формативного аспекта, сколь возможно более точное донесение сведений и последующую формализацию) и воздействия (что, в свою очередь, акцентирует внимание на эмоционально-эстетическом аспекте сообщения и неограниченном выборе языковых средств, позволяющих сколь возможно более эффективно осуществить требуемое воздействие на читателя).

Частичное возвращение к идеям В.В. Виноградова требует рассмотрения статуса религиозного стиля, который не был выделен ученым и, соответственно, не отнесен ни к первой, ни ко второй, ни какой-либо иной группе. Комментарий о религиозном стиле речи может быть необходим также и при рассмотрении потенциальной востребованности тех или иных функций аналитического алгоритма; полезность его детектирования не является очевидной.

Поясняя последнюю мысль, отметим, что разработка упомянутого алгоритма основывается в том числе и на соблюдении требований максимальной полезности, разумности и достаточности. Из этого следует, что автоматический анализатор должен в первую очередь различать те функциональные стили, определение которых актуально и востребовано, поскольку при создании современного программного обеспечения не принято расходовать системные ресурсы на исполнение неиспользуемых или редко вызываемых задач; данный подход уже был нами реализован при рассмотрении вопросов, связанных с детектированием разговорно-обиходного стиля.

Нам представляется, что религиозный стиль, равно как и разговорный, занимает в сфере деятельности АСОТ специфическую узкую нишу, имеющую непосредственное отношение исключительно к работе профильных организаций. В настоящее время нам практически не известны случаи порождения новых текстов, выдержанных в соответствии с нормами религиозного стиля; иными словами, корпус текстов религиозного стиля является в известной степени фиксированным и лишенным притока новых образцов, а задача его исследования может вообще не вставать перед современными автоматизированными системами обработки текстов. В силу вышеизложенного мы находим целесообразным ориентировать аналитический алгоритм на работу с четырьмя книжными стилями речи и двумя их группами («формальные» и «свободные»),

различение которых является наиболее востребованной задачей для современных АСОТ.

В дополнение к вышеизложенному мы также полагаем необходимым заметить, что в перспективе автоматическое детектирование функционального стиля может предоставлять возможность выхода на определение дискурсивных характеристик исследуемого текста; данный вопрос, тем не менее, находим целесообразным оставить вне пределов настоящего исследования.

Суммируя сказанное выше, мы считаем возможным предпринять попытку изыскания тех исчисляемых параметров текстов на естественных языках, которые позволили бы составить представление о его функционально-стилевой принадлежности, равно как и попытку исследования возможностей автоматического определения функционального стиля текста на основании анализа указанных параметров. В соответствии с приведенными выше рассуждениями функциональный стиль может рассматриваться как тот единственный компонент речевой ситуации, который, с одной стороны, исчисляем, а с другой стороны, характеризуется оптимизационным и эвристическим потенциалом; определение функционально-стилевой принадлежности текста, в свою очередь, позволит реализовать комплексный подход к его исследованию посредством действительного применения формально-семантического анализа и учета роли идеального контекста в расстановке приоритетов тех или иных потенциальных значений языковых единиц в процессе их истолкования. Зависимость описанного процесса приоритезации от функционального стиля речи полагаем очевидной.

2.2. ФОРМАЛЬНЫЕ МАРКЕРЫ ФУНКЦИОНАЛЬНО-СТИЛЕВОЙ ПРИНАДЛЕЖНОСТИ (МФСП) ТЕКСТА

Исходя из изложенных выше сведений и рассуждений, на текущем этапе построения лингвистического обеспечения оптимизационного алгоритма на основе детектирования функционального стиля необходимо обратиться к изысканию формальных маркеров функционально-стилевой принадлежности (МФСП) текста. Определение функционального стиля,

как отмечено нами выше, не является принципиально новой задачей; известен ряд исследований соответствующей тематики, предлагающих разнообразные алгоритмические решения для реализации детектирующего механизма. В частности, отметим работы П.И. Браславского [138]; в соответствии с концепцией, предлагаемой этим ученым, может быть выполнено формальное детектирование текста, принадлежащего к научному стилю речи. П.И. Браславский предлагает оценивать такие параметры текста, как средняя длина слова, доля именных частей речи, глагольных форм, личных местоимений; иными словами, в рамках концепции предлагается традиционный подход – исчисление тех параметров текста, которые дают представление о его форме, но не о его содержании. В иных исследованиях (к примеру, в концепции Шевелева О.Г. [109]) происходит дальнейшее углубление в исследование и применение исключительно количественных параметров и критериев, таких, как гипергеометрический критерий Фишера, критерий хи-квадрат, понятие частоты рассогласования при кластеризации текста, мера Кульбаха и др. Нам представляется, что данные воззрения, равно как и подобные им, не позволяют в полной мере реализовать комплексный подход к анализу текстов, о насущной необходимости которого было сказано в рассмотренных нами в первой главе аналитических обзорах и исследованиях. Потребность в реализации комплексного подхода к анализу текста заставляет нас осуществлять поиск двух типов параметров:

1) материальные маркеры (маркеры формы), дающие представление о количественных параметрах текста, равно как и о его внешней организации (длина слова, доля слов определенного типа, наличие или отсутствие заголовков, приложений, таблиц, графиков, списков использованной литературы и т.п.), иначе – о том, что принято называть *планом выражения*;

2) информационные маркеры (маркеры содержания), дающие представление о качественных параметрах текста, равно как и о его информационной составляющей (присутствие определенной специфической семантики, степень однозначности / многозначности текста, характер передаваемой информации и ее тематическая отнесенность и т.д.), иначе – о том, что принято называть *планом содержания.*

При этом оба типа маркеров – как материальные, так и информационные – должны являться формальными, т.е. имеющими внешнее выражение, поддающимися исчислению и формализации, доступными для машинного исследования. В случае с маркерами формы нет необходимости отдельно обозначать их формальность – это следует собственно из содержания термина, - однако в случае информационных маркеров мы полагаем необходимым особо подчеркнуть потребность в формализации.

Комментируя сказанное, поясним, что до настоящего времени возможности применения комплексного подхода к анализу текста ограничивались простотой машинного исследования формы и кажущейся принципиальной нерешаемостью задачи соответствующего анализа содержания. Действительно, не обладая сознанием, вычислительное устройство способно работать исключительно с материальной стороной документа – оценивать те или иные внешние его качества (средние арифметические, доли, присутствие или отсутствие каких-либо объектов и т.п.); в то же время для перехода на уровень содержания требуются обычно разум, интеллект, человеческое сознание. В то же время, подобно тому, как в рамках философии существует представление о взаимовоздействии формы и содержания любого компонента объективной реальности, в пределах лингвистики правомерно существование положения о наличии взаимной связи между формой и содержанием текста. Иными словами, коль скоро формальные параметры текста могут определять его содержание, содержательные характеристики исследуемого текста равноуспешно могут проявляться в тех или иных его материальных особенностях, вплоть до провоцирования их возникновения.

Итак, осуществляя отбор маркеров, необходимо в первую очередь учитывать сферы наиболее критического различия между строго формальными книжными стилями (научным и официально-деловым) и свободными книжными стилями (художественным и публицистическим). Разграничивая формализованные и свободные книжные стили, мы подчеркиваем степень важности внешнего оформления в случае официально-делового и научного стилей, наличие определенных требований, предъявляемых к текстам указанных функциональных

стилей (базовым из которых является, на наш взгляд, требование предельно возможной однозначности текстов, что обусловливает применение строго дефинированных терминов).

Говоря о типических особенностях представленных выше групп функциональных стилей, мы считаем необходимым отметить и рассмотреть в первую очередь те их черты, которые определяются традиционной для них речевой ситуацией в целом.

1. Для строго формальных книжных стилей, т.е. официально-делового и научного, априори характерны:

1.1. Стремление к однозначности, недопустимость тропов, фигур речи, ораторских приемов.

1.2. Частичная реализация пункта 1.1. за счет активного использования терминов – слов, как правило, однозначных либо строго определенных, являющихся обыкновенно международными, т.е. заимствованных из языков межнационального общения или классических языков науки (древнегреческий, латинский, английский, французский).

2. Для свободных книжных стилей, т.е. художественного и публицистического, априори характерны:

2.1. Допустимость многозначности и склонность к ней, активное применение изобразительно-выразительных средств языка, фигур речи, приемов убеждения и т.п.

2.2. Отсутствие стремления к экстенсивному употреблению терминов, включение в текст разговорных элементов (попытки установления непосредственного контакта с читателем, использование цитат или прямой речи персонажей). В связи с этим необходимо упомянуть об экономии языковых средств, которое традиционно имеет два основных проявления: а) эллиптичность, предельное сокращение количества произносимых звуков в целях ускорения передачи сообщений, и б) тенденция к отрицанию «сложной» речи, содержащей термины и малознакомые слова, потенциальная многозначность за счет активной замены «сложных» слов и их компонентов знакомыми «простыми» речевыми единицами.

Если суммировать предложенные выше типические черты различных функциональных стилей и их групп, то становится очевидным, что на основании упомянутых черт могут быть выделены два базовых параметра, которые обладают доста-

точным потенциалом для реализации на их основе оптимизированного алгоритма стилевого разграничения.

С одной стороны, это количественный показатель, описывающий количество звуков и, соответственно, букв в употребляемых лексических единицах, или, как его обыкновенно именуют, средняя длина слова. Применение данного параметра, как было отмечено нами выше, не является новым в вопросах анализа стилевой принадлежности текста, однако мы находим возможным выдвинуть предположение о том, что, исходя из типических характеристик функциональных стилей, параметр средней длины слова может являться единственным количественным параметром текста, достаточным для определения его функционально-стилевой принадлежности. Этот параметр может быть избран для реализации оптимизированного алгоритма также и потому, что в сравнении с другими предложенными, к примеру, П.И. Браславским маркерами он требует наименьшего количества необходимых вычислений, прост в программной реализации и не использует каких-либо БД, тем более объемных (в отличие от, скажем, определения доли именных частей речи, для которого требуется некоторая БД, описывающая собственно то, что необходимо считать именными частями речи в процессе анализа).

С другой стороны, это качественный показатель, описывающий степень полисемантичности текста. Отметим, что нами не обнаружено опыта исследования каких-либо формальных параметров текстов, которые позволяли бы составить представление о плане содержания текста и, соответственно, судить о потенциальной степени его полисемантичности; в то же время нам представляется, что подобный показатель может представлять определенную ценность для процесса выявления функционально-стилевой отнесенности текста. Необходимо, однако, выявить такой формальный показатель, который мог бы обладать оптимизационным потенциалом.

Поясняя данную мысль, заметим, что наиболее очевидным и интуитивно понятным способом определить потенциальную полисемантичность текста является приписывание каждому слову атрибута, характеризующего количество его значений, описанных в словаре, т.е., говоря математическим языком, меру неопределенности, которую вносит слово в текст. Одна-

ко данный способ принципиально неприемлем при построении оптимизированного алгоритма, поскольку требует наличия и сопровождения огромной БД, содержащей записи для всех словарных единиц, равно как и процедур поиска слов в БД и поочередной пометки каждого слова. Подобные вычисления могут быть довольно нецелесообразны с точки зрения расхода машинного времени.

Таким образом, сигнатурный метод оценки полисемантичности текста, хотя и обеспечивает достаточно точный результат, не подходит для поставленных нами задач. Требуется эвристический алгоритм, имеющий дело с приближенной оценкой потенциальной полисемантичности анализируемого текста и могущий предъявлять результат с некоторой погрешностью, не оказывающей, однако, существенного влияния на качество детектирования функционального стиля.

Очевидно, что, говоря о формальных признаках потенциальной полисемантичности слова, за счет которой может осуществляться оценка многозначности текста, необходимо обратиться либо к материальному выражению (внешнему облику) того или иного слова, либо к сочетаниям слов. Последний вариант решения проблемы в сущности представляет собой анализ сочетаемости, т.е. дистрибутивный анализ, изучение физического контекста, что не рассматривается нами в рамках настоящего исследования; соответственно, самым естественным образом остается необходимым обратиться к внешней форме слова, самого по себе.

Исходя из изложенных выше типических черт тех или иных функциональных стилей, в особенности – характеристик, связанных с типом предпочитаемых слов (оппозиции «простые слова – сложные слова», «обыденные слова – термины», откуда могут быть выведены также подоппозиции: «короткие слова – длинные слова» и «собственные слова – международные слова (т.е. заимствованные)»), мы считаем допустимым выдвинуть предположение в следующей формулировке: *существует возможная взаимосвязь между длиной слова, фактом его заимствованности / незаимствованности из других языков и количеством его потенциальных значений.*

Рассуждая в терминах философии, можно согласиться с подобным утверждением. Действительно, представляется

интуитивно очевидным, что никакой язык в момент своего становления не станет прибегать к длинным звукосочетаниям до тех пор, пока не исчерпаны возможности создания более коротких слов; в свою очередь, слова, сформированные в рамках собственно языка, существуют в нем численно дольше, нежели заимствованные, и, соответственно, приобретают большую степень полисемантичности. Необходимо, однако, рассмотреть эмпирический материал с целью обосновать выдвинутое утверждение.

Существует ряд воззрений, согласно которым возможна математическая проверка зависимости качественных характеристик слова от его количественных показателей, в частности, от его длины. В частности, в работе [85] при расчетах параметра Херста для определения целостности текста указано, что по закону распределения частот слов Эсту-Ципфа-Мандельброта частота употребляемости слова в языке вообще и в тексте в частности зависит от его длины. Не следует рассматривать данное утверждение как безусловно истинное, поскольку в зависимости от тех или иных факторов данная зависимость может нарушаться (нет сомнений, что в научном стиле речи слово «данный» встречается чаще, чем слово «я»), однако в целом такая зависимость может иметь место.

Помимо указанной, потенциально допустимы также и иные зависимости качественных характеристик слова от количественных – к примеру, упомянутая нами выше возможность того, что длина слова и частота его употребления могут коррелировать со степенью полисемантизации (т.е. количеством потенциальных значений, предложенных в словаре), а также характеризовать слово по признаку заимствованности. С целью выявления подобных зависимостей нами было проведено статистическое исследование словарного материала для русского и английского языков (Современный толковый словарь русского языка под редакцией С.А. Кузнецова и Merriam-Webster Collegiate Dictionary соответственно). Слова короткие (1-4 буквы), средней длины (5-8 букв) и длинные (9-12 букв) отбирались из словаря случайным выбором, по пять слов каждой группы на каждую букву, общим количеством 426 в случае русского языка и 385 – для английского языка. Отобранный материал анализировался по признаку заимствован-

ности и степени полисемантизации; факт заимствованности, равно как и степень полисемантизации, были определены по словарю.

Комментируя понятие степени полисемантизации, отметим, что можно выделить слова неполисемантизованные (однозначные), а также слабо полисемантизованные (два значения), среднего уровня (3-5 значений) и сильно полисемантизованные (6 и более значений). Слабая полисемантизованность характеризуется незначительным расширением семантики слова, откуда, собственно, и происходит предложенный термин «слабая»; порождение большего количества коннотаций свидетельствует соответственно о большей потенциальной полисемантичности слова. Рассуждение о том, что по мере возрастания массовой доли сильно полисемантизованных слов возрастает мера неопределенности при истолковании текста, представляется очевидным. Полученные данные систематизированы нами и отражены в таблицах.

Таблица 1. Статистические данные для русского языка

Группа	**Р**	**З**	**О**	**СП**	**СрП**	**СиП**	**РО**
Короткие	103	41	14	23	67	40	2
Средние	75	66	26	32	62	21	4
Длинные	55	86	73	51	15	2	21
Всего	233	193	113	106	144	63	27
Группа	**РСП**	**РСрП**	**РСиП**	**ЗО**	**ЗСП**	**ЗСрП**	**ЗСиП**
Короткие	14	52	35	12	10	14	5
Средние	16	35	20	22	16	27	1
Длинные	25	7	2	52	26	8	0
Всего	55	94	57	86	52	49	6

Таблица 2. Статистические данные для английского языка

Группа	**А**	**З**	**О**	**СП**	**СрП**	**СиП**	**АО**
Короткие	92	33	14	16	48	47	2
Средние	40	90	36	39	46	9	3
Длинные	21	109	60	52	17	1	4
Всего	153	232	110	107	111	57	9

Группа	АСП	АСрП	АСиП	ЗО	ЗСП	ЗСрП	ЗСиП
Короткие	8	40	42	12	8	8	5
Средние	13	19	5	33	26	27	4
Длинные	15	2	0	56	37	15	1
Всего	36	61	47	101	71	50	10

Примечание. В обеих таблицах условные обозначения количества: Р – русских слов и А – английских слов (т.е. незаимствованных), З – заимствованных, О – однозначных, СП – многозначных слабо полисемантизованных, СрП – многозначных среднего уровня полисемантизации, СиП – многозначных сильно полисемантизованных, РО – русских однозначных и АО – английских однозначных, РСП – русских многозначных слабо полисемантизованных и АСП – английских многозначных слабо полисемантизованных, РСрП – русских многозначных среднего уровня полисемантизации и АСрП - английских многозначных среднего уровня полисемантизации, РСиП - русских многозначных сильно полисемантизованных и АСиП - английских многозначных сильно полисемантизованных, ЗО - заимствованных однозначных, ЗСП – заимствованных многозначных слабо полисемантизованных, ЗСрП – заимствованных многозначных среднего уровня полисемантизации, ЗСиП – заимствованных многозначных сильно полисемантизованных.

Приведенные таблицы позволяют сделать ряд наблюдений и основанных на них выводов, а именно следующие.

1. В обоих случаях (как для русского, так и для английского языка) по мере увеличения длины слова от коротких к длинным последовательно убывает количество незаимствованных лексических единиц и возрастает количество заимствованных слов. Это позволяет предполагать, что чем короче слово, тем выше вероятность того, что оно не является заимствованным, и наоборот.

2. В обоих случаях по мере увеличения длины слова последовательно возрастает количество однозначных и слабо полисемантизованных лексических единиц, в то время как количество слов среднего и высокого уровня полисемантизации последовательно убывает. Это позволяет утверждать, что чем короче слово, тем выше вероятность того, что оно имеет более двух значений, и наоборот.

3. В обоих случаях тенденции, обозначенные в пункте 2, справедливы безотносительно к факту заимствованности лек-

сической единицы, за исключением заимствованных слов среднего уровня полисемантизации, где возрастание количества от коротких слов к средним сменяется его убыванием от средних слов к длинным.

4. В обоих случаях суммарное количество незаимствованных слов, являющихся однозначными или слабо полисемантизованными, существенно ниже суммарного количества заимствованных слов того же уровня полисемантизации, в то время как суммарное количество незаимствованных лексических единиц среднего и высокого уровня полисемантизации превышает суммарное количество заимствованных слов того же уровня полисемантизации. Это позволяет полагать, что средний и высокий уровень полисемантизации является преимущественным свойством незаимствованных слов, и наоборот.

Интеграция представленных выше наблюдений и выводов приводит нас к общему заключению: по мере увеличения длины слова возрастает вероятность того, что оно является заимствованным и обладает либо единственным денотативным значением, либо одним денотативным и одним коннотативным значением, т.е. представляет собой либо однозначную, либо слабо полисемантизованную лексическую единицу. Справедливо также обратное утверждение: по мере увеличения длины слова убывает вероятность того, что оно является незаимствованным и имеет более двух значений, т.е. представляет собой лексическую единицу среднего или высокого уровня полисемантизации. При этом описанные явления не могут быть отнесены к особенностям исключительно русского языка, т.к. аналогичные тенденции наблюдаются и во взятом нами для сравнения английском языке; иными словами, обнаруженные закономерности могут иметь и универсальный характер.

Таким образом, на основании статистических данных и их непосредственной обработки мы находим допустимым подтвердить возможность наличия непосредственной взаимосвязи между длиной слова, фактом его заимствованности и степенью его полисемантичности. Соответственно, большая степень полисемантичности, т.е. развития дополнительных коннотаций, свидетельствует о более активном использовании слова носителями языка, откуда, собственно, и следует, что чем короче слово, тем чаще оно употребляется в речи, т.е.

закон Эсту-Ципфа-Мандельброта подтверждается. Соотнося полученные результаты с соображениями, высказанными нами выше, мы можем прийти к окончательному выводу, который сформулирован ниже.

Итак, согласно проведенным нами исследованиям и теоретическим изысканиям, наиболее характерными особенностями свободных стилей речи могут быть признаны элементы эллиптичности и склонность к экстенсивному увеличению количества возможных коннотаций, которые выражаются во взаимосвязанных процессах сокращения длины слова, активного использования незаимствованных слов и увеличения степени полисемантизации лексических единиц. В свою очередь, типическими характеристиками формальных стилей могут быть признаны применение развернутых конструкций, отсутствие стремления к экономии языковых средств и относительно слабая склонность к порождению дополнительных коннотаций (т.е. к полисемантизации), что выражается в увеличении длины слова, более или менее выраженной тенденции к использованию заимствованных слов и уменьшению степени полисемантизации лексических единиц. Отсюда непосредственно следует, что отделение свободных книжных стилей от формальных может быть осуществлено посредством выявления и исследования двух взаимосвязанных маркеров: средней длины слова и степени потенциальной полисемантичности текста.

Теперь, когда нами установлена взаимосвязь длины слова и степени его полисемантизации, а также обоснована возможность применения данных характеристик в качестве маркеров функционально-стилевой принадлежности текста, нам надлежит обратиться к вопросу формализации второго из перечисленных маркеров. Действительно, в то время как первый параметр – средняя длина слова – является формальным по своей сущности и по определению, второй параметр представляет собой компонент плана содержания, и возможности его формализации неочевидны.

Как было сказано нами выше, в наши задачи на данном этапе исследования входит изыскание внешних признаков потенциальной полисемантичности текста. Выяснив на основании теоретических рассуждений и практических исследований, что низкая степень полисемантичности текста мо-

жет свидетельствовать о высокой доле в нем заимствованных слов, в том числе терминологического характера, и наоборот, в процессе решения указанной задачи мы можем свести его к изысканию внешних признаков заимствованных и незаимствованных слов соответственно, соотнося получаемые данные с вопросами потенциальной полисемантичности текста.

Из истории всякого языка известно, что формирование его письменности связано с попытками формирования собственного алфавита либо адаптации постороннего для нужд конкретного языка. Не углубляясь в предмет, заметим, однако, что закономерным следствием развития и приспособления письменной речи в целом и алфавита в частности к устной реализации языка является возникновение в письменной речи букв и буквосочетаний, а также иных графических знаков и их комбинаций, которые а) характеризуют специфические для данного конкретного языка звуки, сочетания звуков и т.п. и б) характеризуют звуки и их сочетания, в данном конкретном языке отсутствующие.

Поскольку в рамках настоящего исследования мы работаем главным образом со словарным материалом, нам представляется целесообразным осуществлять анализ в данном направлении именно с опорой на лексикографические источники. Наиболее очевидным и просто реализуемым приемом, который позволил бы оценить роль характерных и не характерных для языка графических символов в оценивании лексической многозначности, нам представляется изучение словарных групп, т.е. совокупностей слов, расположенных в определенной последовательности согласно первой букве их графического выражения. Так как выше нами уже была доказана возможность взаимосвязи между графическим обликом слова и фактом его заимствованности, а также со степенью полисемантичности, мы не будем повторно обосновывать данную мысль, однако сформулируем общее предположение в следующем виде.

Функциональный стиль речи, к которому принадлежит текст, может быть определен посредством оценки и интерпретации данных о массовой доле в указанном тексте слов, принадлежащих к определенным словарным группам. Данное предположение базируется на умозаключении о том, что, коль

скоро те или иные звукосочетания, выраженные определенными графическими элементами, могут быть свойственны или не свойственны исследуемому языку, их наличие или отсутствие может являться косвенным свидетельством заимствованности или незаимствованности того или иного слова, а, следовательно, и степени его полисемантичности. Потенциальная же полисемантичность слова как мера неопределенности его толкования экстраполируется и на текст в целом, поскольку мера неопределенности системы (т.е. энтропия) по определению обладает свойством аддитивности.

Подобное исследование самым естественным образом позволит составлять представление о степени потенциальной полисемантичности текста, т.е. о плане его содержания, посредством формального анализа – изучения плана выражения текста.

2.3. ИССЛЕДОВАНИЕ ПОТЕНЦИАЛЬНОЙ ПОЛИСЕМАНТИЧНОСТИ СЛОВ, ПРИНАДЛЕЖАЩИХ К РАЗЛИЧНЫМ СЛОВАРНЫМ ГРУППАМ

В соответствии с представленными выше теоретическими рассуждениями представляется целесообразным осуществить практическое исследование потенциальной полисемантичности слов, принадлежащих к различным словарным группам, на материале словарей русского языка. В процессе подобного исследования необходимо решить следующие непосредственные задачи:

- расчет общего количества слов, которые относятся к той или иной словарной группе,
- расчет количества многозначных слов, с разделением слабо полисемантизованных, слов среднего уровня полисемантизации и сильно полисемантизованных лексических единиц,
- расчет массовой доли многозначных слов в той или иной словарной группе, позволяющий судить о потенциальной полисемантичности слов из указанной группы.

Результаты выполнения обозначенных задач позволят осуществить отбор словарных групп, характеризующихся значи-

мо высокой / низкой массовой долей неоднозначных слов, значимо высокой / низкой массовой долей слов среднего уровня полисемантизации, значимо высокой / низкой массовой долей сильно полисемантизованных слов. Полагаем необходимым отметить, что показатель массовой доли слов определенного типа должен быть именно значимо высоким или низким, т.е. существенно отличаться от среднестатистического показателя, характерного для основной массы слов.

В соответствии с представленными выше принципами и в целях выполнения указанных выше задач нами было проведено соответствующее исследование словарного материала. Расчеты производились для следующих справочных изданий:

1) Ожегов С.И. Словарь русского языка / Под общ. ред. проф. Л.И. Скворцова. – 25-е изд., испр. и доп. – М., 2008. (далее - «словарь Ожегова»)

2) Современный толковый словарь русского языка / Под ред. С.А. Кузнецова. - М., 2004. (далее - «словарь Кузнецова»)

3) Малый академический словарь русского языка в 4-х тт. / РАН, Ин-т лингвистических исследований; Под ред. А. П. Евгеньевой. – 4-е изд., стер. — М., 1999. (далее – «МАС»)

Для каждой словарной группы каждого из словарей производились: расчет общего количества входов (заголовочных слов) словарных статей; расчет количества слов, имеющих более одного значения, а также их суммарной массовой доли по формуле М / С * 100%, где М – количество неоднозначных слов, С – общее количество слов в группе; расчет количества слов, имеющих более трех значений, а также их суммарной массовой доли по формуле СрП / С * 100%, где СрП – количество слов среднего уровня полисемантизации; расчет количества слов, имеющих более шести значений, а также их суммарной массовой доли по формуле СиП / С * 100%, где СиП – количество сильно полисемантизованных слов; расчет суммарной массовой доли слов среднего и высокого уровня полисемантизации по формуле СрП+СиП / С * 100%.

Полученные данные систематизированы нами и представлены в таблицах.

Таблица 3. Результаты расчетов, данные для словаря Ожегова

Буква	Слов	О	М	СП	СрП	СиП	% М	% СрП	% СиП	% СрП +СиП
А	726	575	151	130	21	0	20,80%	2,89%	0,00%	2,89%
Б	1497	1176	321	250	67	4	21,44%	4,48%	0,27%	4,74%
В	2514	1915	599	414	169	16	23,83%	6,72%	0,64%	7,36%
Г	993	769	224	156	63	5	22,56%	6,34%	0,50%	6,85%
Д	1396	1030	366	275	76	15	26,22%	5,44%	1,07%	6,52%
ЕЁ	133	94	39	26	11	2	29,32%	8,27%	1,50%	9,77%
Ж	301	206	95	63	28	4	31,56%	9,30%	1,33%	10,63%
З	1788	1314	474	340	123	11	26,51%	6,88%	0,62%	7,49%
И	935	719	216	172	40	4	23,10%	4,28%	0,43%	4,71%
Й	5	4	1	1	0	0	20,00%	0,00%	0,00%	0,00%
К	2259	1706	553	396	149	8	24,48%	6,60%	0,35%	6,95%
Л	781	592	189	138	46	5	24,20%	5,89%	0,64%	6,53%
М	1404	1104	300	212	84	4	21,37%	5,98%	0,28%	6,27%
Н	2229	1670	559	411	142	6	25,08%	6,37%	0,27%	6,64%
О	2748	1974	774	560	193	21	28,17%	7,02%	0,76%	7,79%
П	6287	4685	1602	1135	429	38	25,48%	6,82%	0,60%	7,43%
Р	2102	1444	658	458	187	13	31,30%	8,90%	0,62%	9,51%
С	3607	2691	916	624	258	34	25,40%	7,15%	0,94%	8,10%
Т	1272	936	336	212	113	11	26,42%	8,88%	0,86%	9,75%
У	1056	752	304	218	80	6	28,79%	7,58%	0,57%	8,14%
Ф	490	358	132	97	32	3	26,94%	6,53%	0,61%	7,14%
Х	441	321	120	82	29	9	27,21%	6,58%	2,04%	8,62%
Ц	196	142	54	38	16	0	27,55%	8,16%	0,00%	8,16%
Ч	488	374	114	77	32	5	23,36%	6,56%	1,02%	7,58%
Ш	509	391	118	94	23	1	23,18%	4,52%	0,20%	4,72%
Щ	72	46	26	17	8	1	36,11%	11,11%	1,39%	12,50%
Э	323	255	68	53	15	0	21,05%	4,64%	0,00%	4,64%
Ю	45	34	11	11	0	0	24,44%	0,00%	0,00%	0,00%
Я	135	90	45	30	14	1	33,33%	10,37%	0,74%	11,11%
Всего	36732	27367	9365	6690	2448	227				

Для словаря Ожегова суммарная массовая доля многозначных слов составила 25,5%; доля слов среднего уровня полисемантизации – 6,66%; высоко полисемантизованных слов – 0,62%; слов, имеющих более трех значений – 7,28%.

Таблица 4. Результаты расчетов, данные для словаря Кузнецова

Буква	Слов	О	М	СП	СрП	СиП	% М	% СрП	% СиП	% СрП +СиП
А	884	593	291	224	67	0	32,92%	7,58%	0,00%	7,58%
Б	1623	1023	600	414	171	15	36,97%	10,54%	0,92%	11,46%
В	2654	1662	992	630	317	45	37,38%	11,94%	1,70%	13,64%
Г	1083	656	427	278	129	20	39,43%	11,91%	1,85%	13,76%
Д	1506	968	538	335	184	19	35,72%	12,22%	1,26%	13,48%
ЕЁ	136	83	53	31	19	3	38,97%	13,97%	2,21%	16,18%
Ж	281	166	115	76	34	5	40,93%	12,10%	1,78%	13,88%
З	1971	1190	781	490	265	26	39,62%	13,44%	1,32%	14,76%
И	1064	664	400	274	119	7	37,59%	11,18%	0,66%	11,84%
Й	6	4	2	2	0	0	33,33%	0,00%	0,00%	0,00%
К	2548	1527	1021	614	361	46	40,07%	14,17%	1,81%	15,97%
Л	841	509	332	200	120	12	39,48%	14,27%	1,43%	15,70%
М	1564	874	690	427	241	22	44,12%	15,41%	1,41%	16,82%
Н	2457	1474	983	636	313	34	40,01%	12,74%	1,38%	14,12%
О	2960	1685	1275	769	455	51	43,07%	15,37%	1,72%	17,09%
П	6903	3846	3057	1795	1106	156	44,29%	16,02%	2,26%	18,28%
Р	2299	1324	975	593	341	41	42,41%	14,83%	1,78%	16,62%
С	4060	2271	1789	1005	654	130	44,06%	16,11%	3,20%	19,31%
Т	1390	769	621	344	232	45	44,68%	16,69%	3,24%	19,93%
У	1160	587	573	318	220	35	49,40%	18,97%	3,02%	21,98%
Ф	574	339	235	146	82	7	40,94%	14,29%	1,22%	15,51%
Х	454	242	212	122	67	23	46,70%	14,76%	5,07%	19,82%
Ц	208	99	109	55	47	7	52,40%	22,60%	3,37%	25,96%
Ч	469	248	221	123	86	12	47,12%	18,34%	2,56%	20,90%
Ш	571	324	247	168	74	5	43,26%	12,96%	0,88%	13,84%
Щ	74	39	35	16	17	2	47,30%	22,97%	2,70%	25,68%
Э	410	261	149	104	42	3	36,34%	10,24%	0,73%	10,98%
Ю	54	33	21	13	8	0	38,89%	14,81%	0,00%	14,81%
Я	145	85	60	15	42	3	41,38%	28,97%	2,07%	31,03%
Всего	40349	23545	16804	10217	5813	774				

Для словаря Кузнецова суммарная массовая доля многозначных слов составила 41,65%; доля слов среднего уровня полисемантизации – 14,41%; высоко полисемантизованных слов – 1,92%; слов, имеющих более трех значений – 16,33%.

Таблица 5. Результаты расчетов, данные для МАС

Буква	Слов	О	М	СП	СрП	СиП	% М	% СрП	% СиП	% СрП +СиП
А	1600	1338	262	222	40	0	16,38%	2,50%	0,00%	2,50%
Б	3122	2577	545	409	128	8	17,46%	4,10%	0,26%	4,36%
В	6049	4586	1463	1095	337	31	24,19%	5,57%	0,51%	6,08%
Г	2274	1821	453	306	137	10	19,92%	6,02%	0,44%	6,46%
Д	3528	2838	690	525	149	16	19,56%	4,22%	0,45%	4,68%
ЕЁ	226	165	61	43	16	2	26,99%	7,08%	0,88%	7,96%
Ж	630	489	141	91	46	4	22,38%	7,30%	0,63%	7,94%
З	4828	3673	1155	846	284	25	23,92%	5,88%	0,52%	6,40%
И	2250	1654	596	455	131	10	26,49%	5,82%	0,44%	6,27%
Й	18	16	2	2	0	0	11,11%	0,00%	0,00%	0,00%
К	5056	3870	1186	800	350	36	23,46%	6,92%	0,71%	7,63%
Л	1768	1357	411	298	100	13	23,25%	5,66%	0,74%	6,39%
М	3473	2661	812	557	242	13	23,38%	6,97%	0,37%	7,34%
Н	5846	4386	1460	1047	379	34	24,97%	6,48%	0,58%	7,06%
О	6646	4794	1852	1347	459	46	27,87%	6,91%	0,69%	7,60%
П	16594	12415	4179	3199	890	90	25,18%	5,36%	0,54%	5,91%
Р	4725	3284	1441	980	418	43	30,50%	8,85%	0,91%	9,76%
С	8611	6317	2294	1481	701	112	26,64%	8,14%	1,30%	9,44%
Т	2969	2188	781	491	252	38	26,31%	8,49%	1,28%	9,77%
У	2582	1686	896	637	238	21	34,70%	9,22%	0,81%	10,03%
Ф	1347	1053	294	214	74	6	21,83%	5,49%	0,45%	5,94%
Х	1008	776	232	141	74	17	23,02%	7,34%	1,69%	9,03%
Ц	491	349	142	91	48	3	28,92%	9,78%	0,61%	10,39%
Ч	1134	861	273	174	87	12	24,07%	7,67%	1,06%	8,73%
Ш	1212	904	308	213	91	4	25,41%	7,51%	0,33%	7,84%
Щ	156	113	43	24	17	2	27,56%	10,90%	1,28%	12,18%
Э	881	686	195	148	43	4	22,13%	4,88%	0,45%	5,33%
Ю	74	50	24	18	6	0	32,43%	8,11%	0,00%	8,11%
Я	193	111	82	54	25	3	42,49%	12,95%	1,55%	14,51%
Всего	89291	67018	22273	15908	5762	603				

Для МАС суммарная массовая доля многозначных слов составила 24,94%; доля слов среднего уровня полисемантизации – 6,45%; высоко полисемантизованных слов – 0,68%; слов, имеющих более трех значений – 7,13%.

Примечание. Для всех трех таблиц условные обозначения: О – количество однозначных слов, М – количество многозначных слов, СП – количество слабо полисемантизованных слов, СрП – количество слов среднего уровня полисемантизации, СиП – количество сильно полисемантизованных слов, %М – массовая доля неоднозначных слов, %СрП – массовая доля слов среднего уровня полисемантизации, %СиП – массовая доля сильно полисемантизованных слов, %СрП+СиП – массовая доля слов, имеющих более трех значений

Представленные данные позволяют заключить, что, хотя словарная база рассмотренных справочных изданий довольно различна, и составителями словарей практикуются неодинаковые подходы к определению полисемантичности слова (что выражается в существенной разнице между значениями массовой доли неоднозначных слов), для всех трех словарей наблюдается значимый разброс долевых показателей. Отобразим эти показатели в таблице:

Таблица 6. Сопоставление основных показателей

Сл. / Показ.	Min %М	Max %М	Δ %М	Min %СрП	Max %СрП	Δ %СрП
Ожегов	20%	36%	16%	0%	11%	11%
Кузнецов	33%	52%	19%	0%	29%	29%
МАС	11%	42%	31%	0%	13%	13%
Сл. / Показ.	**Min %СиП**	**Max %СиП**	**Δ %СиП**	**Min %СрП+СиП**	**Max %СрП+СиП**	**Δ %СрП+СиП**
Ожегов	0%	2%	2%	0%	13%	13%
Кузнецов	0%	5%	5%	0%	31%	31%
МАС	0%	2%	2%	0%	15%	15%

Из таблицы видно, что во всех случаях имеются диапазоны значений с достаточно существенной для их различения дельтой. Небезынтересно также отметить, что во всех трех словарях имеются словарные группы, в пределах которых принципиально отсутствуют слова, имеющие более трех значений.

Проанализируем полученные нами данные, осуществив выборку репрезентативных значений из таблиц 3-5. Для этого

возьмем 5 групп, обладающих наиболее высокими показателями по каждой из четырех колонок со значениями массовых долей (доля многозначных слов, доля слов среднего уровня полисемантизации, доля высоко полисемантизованных слов, доля слов, имеющих более трех значений), и 5 групп, обладающих наименее высокими показателями по обозначенным колонкам, из каждой таблицы. В случае, если к последней группе примыкает элемент с аналогичным целым показателем, он также берется в рассмотрение, но выделяется серой заливкой.

Таблица 7. Выборки из данных для словаря Ожегова

Буква	**% М**	**Буква**	**% СрП**	**Буква**	**% СиП**	**Буква**	**% СрП +СиП**
Выборка 1		Выборка 2		Выборка 3		Выборка 4	
Щ	36%	**Щ**	11%	**Х**	2,04%	**Щ**	13%
Я	33%	**Я**	10%	**ЕЁ**	1,50%	**Я**	11%
Ж	32%	**Ж**	9%	**Щ**	1,39%	**Ж**	11%
Р	31%	**Р**	9%	**Ж**	1,33%	**ЕЁ**	10%
ЕЁ	29%	**Т**	9%	**Д**	1,07%	**Т**	10%
У	29%			**Ч**	1,02%	**Р**	10%
						Б	5%
						Ш	5%
Б	21%	**Б**	4%	**Ц**	0,00%	**И**	5%
М	21%	**И**	4%	**Э**	0,00%	**Э**	5%
Э	21%	**А**	3%	**А**	0,00%	**А**	3%
А	21%	**Ю**	0%	**Ю**	0,00%	**Ю**	0%
Й	20%	**Й**	0%	**Й**	0,00%	**Й**	0%

Таблица 8. Выборки из данных для словаря Кузнецова

Буква	**% М**	**Буква**	**% СрП**	**Буква**	**% СиП**	**Буква**	**% СрП +СиП**
Выборка 1		Выборка 2		Выборка 3		Выборка 4	
Ц	52%	**Я**	29%	**Х**	5%	**Я**	31%
У	49%	**Щ**	23%	**Ц**	3%	**Ц**	26%
Щ	47%	**Ц**	23%	**Т**	3%	**Щ**	26%
Ч	47%	**У**	19%	**С**	3%	**У**	22%

Продолжение табл. 8

Буква	% М	Буква	% СрП	Буква	% СиП	Буква	% СрП +СиП
Х	47%	Ч	18%	У	3%	Ч	21%
				Б	0,92%		
В	37%			Ш	0,88%		
Б	37%	И	11%	Э	0,73%	И	12%
Э	36%	Б	11%	И	0,66%	Б	11%
Д	36%	Э	10%	Ю	0,00%	Э	11%
Й	33%	А	8%	А	0,00%	А	8%
А	33%	Й	0%	Й	0,00%	Й	0%

Таблица 9. Выборки из данных для МАС

Буква	% М	Буква	% СрП	Буква	% СиП	Буква	% СрП +СиП
Выборка 1		Выборка 2		Выборка 3		Выборка 4	
Я	42%	Я	13%	Х	1,69%	Я	15%
У	35%	Щ	11%	Я	1,55%	Щ	12%
Ю	32%	Ц	10%	С	1,30%	Ц	10%
Р	30%	У	9%	Щ	1,28%	У	10%
Ц	29%	Р	9%	Т	1,28%	Т	10%
				Ч	1,06%	Р	10%
		Ф	5%				
		П	5%	М	0,37%		
Г	20%	Э	5%	Ш	0,33%	Э	5%
Д	20%	Д	4%	Б	0,26%	Д	5%
Б	17%	Б	4%	Ю	0,00%	Б	4%
А	16%	А	3%	А	0,00%	А	3%
Й	11%	Й	0%	Й	0,00%	Й	0%

В наших ранних работах [27] уже было представлено первое приближение к вопросам, связанным с изучением массовой доли многозначных слов в словаре Ожегова. Для обозначения числа, выражающего статическое отношение количества полисемантичных слов к общему количеству лексических единиц в словарной группе / словаре в целом, нами был предложен термин «коэффициент степени полисемантизации» (КСП). Очевидно, что рассчитанные нами массовые доли

слов определенного уровня полисемантизации в сущности представляют собой именно вариации КСП для каждой словарной группы в отдельности (т.н. локальные коэффициенты, ЛКСП), для удобства восприятия преобразованные в процентные соотношения.

Анализ выборок, представленных в таблицах 7-9, позволяет выделить пять типов характеристических словарных групп, обладающих повышенными или пониженными показателями по каждой из основных колонок общей таблицы данных:

1) наиболее достоверные словарные группы;
2) достаточно достоверные словарные группы;
3) не вполне достоверные словарные группы;
4) слабо достоверные словарные группы;
5) недостоверные словарные группы.

В данном случае под «достоверностью» мы понимаем степень уверенности в том, что рассматриваемая группа действительно обладает значимо высокими или низкими показателями потенциальной полисемантичности и может обеспечить вероятностное определение последней в процессе исследования текста на естественном языке.

Мы предъявляем к словарным группам следующие критерии отбора:

- наиболее достоверной признается словарная группа, имеющая в выборке четыре вхождения из четырех возможных, все из которых находятся в «чистой зоне» (5 первых / 5 последних пунктов);
- достаточно достоверной признается словарная группа, имеющая в выборке три вхождения из четырех возможных, все из которых находятся в «чистой зоне»;
- не вполне достоверной признается словарная группа, имеющая в выборке четыре вхождения из четырех возможных, не все из которых находятся в «чистой зоне»;
- слабо достоверной признается словарная группа, имеющая в выборке три вхождения из четырех возможных, не все из которых находятся в «чистой зоне»;
- все прочие словарные группы признаются недостоверными.

Отразим распределение групп по типам в таблице:

Таблица 10. Распределение словарных групп

Словарь / Группа	Ожегов		Кузнецов		МАС	
	(+)	(-)	(+)	(-)	(+)	(-)
НД	Ж, Щ	А, Й	У, Ц	А, Й, Э	Я	А, Б, Й
ДД	ЕЁ, Я	Э, Ю	Ч, Щ	И	У, Ц, Щ	Д
НВД	нет	нет	нет	Б	нет	нет
СД	Р	Б	нет	нет	нет	нет

Примечание. Условные обозначения в таблице: НД – наиболее достоверные, ДД – достаточно достоверные, НВД – не вполне достоверные, СД – слабо достоверные, (+) – словарные группы со значимо высокими показателями, (-) – словарные группы со значимо низкими показателями. Все группы, не отмеченные в таблице, являются недостоверными

Таким образом, интегрировав результаты, полученные для всех трех словарей, мы имеем возможность окончательно определить список достоверных словарных групп, на основании которых может производиться вероятностное оценивание потенциальной полисемантичности текста в процессе его автоматического анализа. Представим итоговый анализ собранного нами эмпирического материала:

1) Наиболее достоверной словарной группой со значимо высокими показателями полисемантизации является группа Щ. В словаре Ожегова данная группа характеризуется как наиболее достоверная, в словаре Кузнецова и МАС – как достаточно достоверная.

Наиболее достоверными словарными группами со значимо низкими показателями полисемантизации являются группы А и Й. Во всех изученных словарях эти группы относятся к числу наболее достоверных.

2) Достаточно достоверными словарными группами со значимо высокими показателями полисемантизации являются группы У, Ц, Я. Каждая из этих групп характеризуется как наиболее достоверная и / или достаточно достоверная в двух словарях из трех изученных нами.

Достаточно достоверными словарными группами со значимо низкими показателями полисемантизации являются

группы Б и Э. Группа Б представлена во всех трех частях итоговой таблицы, характеризуясь как слабо достоверная в словаре Ожегова, не вполне достоверная в словаре Кузнецова и наиболее достоверная в МАС; группа Э характеризуется как наиболее достоверная в словаре Кузнецова и достаточно достоверная в словаре Ожегова.

3) Не вполне достоверными словарными группами со значимо высокими показателями полисемантизации являются группы ЕЁ, Ж, Ч. Каждая из этих групп характеризуется как наиболее достоверная и / или достаточно достоверная в одном словаре из трех изученных нами.

Не вполне достоверными словарными группами со значимо низкими показателями полисемантизации являются группы Д, И, Ю. Каждая из этих групп характеризуется как достаточно достоверная в одном из трех изученных нами словарей.

4) Слабо достоверной словарной группой со значимо высокими показателями полисемантизации является группа Р. Данная группа характеризуется как слабо достоверная в словаре Ожегова.

Слабо достоверных словарных групп со значимо низкими показателями полисемантизации не существует.

Таким образом, процедура вероятностной оценки потенциальной полисемантичности текста может быть представлена как совокупность нескольких этапов, состоящих в определении массовой доли слов, которые относятся к каждому из типов достоверных словарных групп, и последующем сравнении полученных результатов с эталонными показателями. Эталонные характеристики могут быть получены посредством анализа заданной выборки текстов, принадлежащих к каждому из рассматриваемых функциональных стилей, и усреднения итоговых данных. Подразделение достоверных словарных групп на типы позволяет решить задачу приоритезации указанных этапов, а также избежать определяющего воздействия концепции построения каждого отдельного словаря: очевидно, что, если в трех независимых друг от друга словарях зафиксирована высокая или низкая доля полисемантичных слов в рамках одной словарной группы, то становится невозможным признать подобный результат зависящим от содержания изучаемых справочных изданий.

Следует, однако, отметить, что вероятностная оценка потенциальной полисемантичности речевого произведения на основании определения массовых долей слов из достоверных словарных групп, хотя и характеризуется определенной логической взаимосвязью с ситуацией общения, но может и не коррелировать напрямую с функциональным стилем исследуемого текста. Преобладание тех или иных словарных групп может быть продиктовано иными факторами, не имеющими непосредственного отношения к функционально-стилевой отнесенности. В силу этого надлежит принимать во внимание необходимость верификации полученных вероятностных результатов; согласно нашему предложению, указанная верификационная роль может быть возложена на процедуру определения средней длины слова.

Тем не менее, результаты осуществленного нами эмпирического исследования словарного материала позволяют уверенно утверждать, что текст, характеризующийся высокой массовой долей слов из словарных групп ЕЁ, Ж, У, Ц, Ч, Щ, Я, имеет относительно высокий уровень потенциальной полисемантичности и подлежит специфической обработке с приданием повышенного веса коннотативным значениям слов и подключением дополнительных тематических словарей. В свою очередь, текст, который характеризуется высокими долевыми показателями словарных групп А, Б, Д, И, Й, Э, Ю, обладает относительно низким уровнем потенциальной полисемантичности и должен быть обработан со смещением логического веса на денотативные элементы лексического значения слов. При этом основополагающим показателем является массовая доля тех слов, словарные группы которых не менее чем в двух обработанных нами словарях имеют наиболее высокие / низкие коэффициенты:

- общей степени полисемантизации,
- среднего уровня многозначности,
- высокого уровня многозначности,
- суммарного количества слов, имеющих более трех значений.

В сущности, соответствие всем обозначенным критериям и является основанием для отнесения группы к типу наиболее достоверных. Однако необходимо заметить, что количество

безусловно надежных групп относительно невелико, что создает риск возникновения существенных погрешностей при оценке. Нейтрализация названного риска достигается за счет привлечения прочих словарных групп меньшего уровня достоверности с соответствующим ранжированием и приоритезацией полученных результатов.

2.4. ВЫВОДЫ

Представлены результаты теоретических и эмпирических исследований, направленных на построение оптимизационных решений для нужд АСОТ на основании надежного и достоверного лингвистического обеспечения.

Краткий анализ существующих воззрений на структуру и компоненты речевой ситуации позволил установить, что не существует единого подхода к выделению вышеуказанных компонентов. Распространенной является модель речевой ситуации Якобсона-Хаймса, в соответствии с которой могут быть выделены семь компонентов, или факторов, речи, как то: отправитель, получатель, форма сообщения, канал связи, код, тема (функциональный стиль, подъязык) и обстановка; данная модель была принята в качестве основной.

В результате рассмотрения данных компонентов был сделан вывод о том, что автоматическое определение функционально-стилевой принадлежности текстов характеризуется равнозначной ценностью для всех типов автоматических анализаторов, а функциональный стиль речи обладает некоторой характерной исключительно для него совокупностью формальных и содержательных параметров, которые, с одной стороны, исчисляемы, а с другой стороны, в перспективе могут обладать эвристическим потенциалом.

В качестве основной была принята дефиниция функционального стиля речи как своеобразного характера речи той или иной социальной ее разновидности, соответствующей определенной сфере общественной деятельности и соотносительной с ней форме сознания, создаваемого особенностями функционирования в этой сфере языковых средств и специфической речевой организацией. В соответствии с данным критерием выделимы 6 функциональных стилей – научный, официально-деловой, публицистический, художественный, религиозный

и разговорно-обиходный; в силу ряда существенных причин, накладывающих ограничения на потенциальный алгоритм оптимизации, а также в соответствии с принципом разумности и достаточности из предложенной совокупности были избраны 4 функциональных стиля, вопрос о разграничении которых является наиболее существенным для современных АСОТ: научный, официально-деловой, публицистический и художественный.

В ходе дальнейшего исследования были выявлены те исчисляемые параметры текстов на естественных языках, которые позволяют составить представление о его функционально-стилевой принадлежности – средняя длина слова и потенциальная полисемантичность, – а также определены возможности автоматического детектирования функционального стиля текста на основании анализа указанных параметров. В частности, установлено, что функциональный стиль речи, к которому принадлежит текст, может быть определен посредством оценки и интерпретации данных о массовой доле в указанном тексте слов, принадлежащих к определенным словарным группам; изучение словарного материала и проведение требуемых расчетов в совокупности с последующим анализом полученных результатов позволили составить список достоверных словарных групп, на основании которых в дальнейшем производилось вероятностное оценивание потенциальной полисемантичности текста в процессе его автоматического анализа. В итоге получены данные, свидетельствующие, что текст, характеризующийся высокой массовой долей слов из словарных групп «Е(Ё)» (29%...39%), «Ж» (22%...41%), «У» (29%...49%), «Ц» (28%...52%), «Ч» (23%...47%), «Щ» (28%...47%), «Я» (33%...43%), имеет относительно высокий уровень потенциальной полисемантичности, а текст, который характеризуется высокими долевыми показателями словарных групп «А» (16%...33%), «Б» (17%...37%), «Д» (20%...36%), «И» (23%...38%), «Й» (11%...33%), «Э» (21%...36%), «Ю» (24%...39%), обладает относительно низким уровнем потенциальной полисемантичности. При этом основополагающим показателем является массовая доля тех слов, словарные группы которых не менее чем в двух обработанных словарях имеют наиболее высокие / низкие коэффициенты общей степени полисемантизации, среднего уровня многозначности, высокого уровня многозначности и суммарного количества слов, имеющих более трех значений.

3. РЕАЛИЗАЦИЯ ОПТИМИЗАЦИОННОГО ПОТЕНЦИАЛА ФУНКЦИОНАЛЬНО-СТИЛИСТИЧЕСКОЙ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

3.1. АЛГОРИТМ ДЕТЕКТИРОВАНИЯ ФУНКЦИОНАЛЬНОГО СТИЛЯ ПРИ ПОМОЩИ МФСП

В предыдущей части нашего исследования мы выявили основные параметры, которые могут лечь в основу унифицированного алгоритма определения функционально-стилевой принадлежности текста. Напомним, что, согласно базовой гипотезе, расчет средней длины слова (СДС) текста и массовой доли в нем слов, принадлежащих к специфическим словарным группам, может оказаться достаточным для выявления его функционального стиля с некоторой достоверностью.

Прежде чем приступать собственно к разработке алгоритма детектирования функционального стиля текста при помощи МФСП (условимся для краткости именовать его в дальнейшем «МФСП-алгоритм»), полагаем необходимым осуществить предварительный анализ параметров текстов, принадлежащих к различным ФС. С одной стороны, исследование характеристик типичных образцов функциональных стилей позволит определить количественный параметр средней длины слова, на который необходимо будет ориентироваться в дальнейшем, а, с другой стороны, предоставит возможность наиболее общим образом оценить перспективы параметра потенциальной полисемантичности текста как детектирующего критерия. В отношении последнего необходимо получить достаточно отличные друг от друга результаты, чтобы определение ФС текста было уверенным.

Таким образом, нам необходимо изучить среднюю длину слова и относительную степень полисемантизованности для текстов различных функциональных стилей, а также исследовать возможность применения тех или иных полученных по-

казателей в целях идентификации стилевой принадлежности текста. Для этих целей нами были взяты 4 текста приближенно равной длины (около 20 000 знаков), относящихся к научному, художественному, публицистическому и официально-деловому стилям речи соответственно: 1) обзорная статья «Модели мандатного контроля и управления доступом»; 2) перевод на русский язык рассказа О. Уайлда «Молодой король»; 3) статья А. Трапезникова «Я в наследстве отцов ощутил...»; 4) Федеральный закон «О рекламе», статьи 1-15.

Для данных текстов были проведены расчеты средней длины слова и относительного уровня полисемантизованности. Средняя длина слова рассчитывалась с помощью меню «Сервис – Статистика» текстового процессора Microsoft Office Word 2003 по формуле СДС = Qзбп / Qс, где СДС – средняя длина слова, Qзбп – количество знаков без пробелов, Qс – количество слов в документе. Определение количества слов, относящихся к той или иной словарной группе, было сведено к определению количества знакосочетаний «пробел плюс первая буква группы», что является достаточным для данного шага исследования. Данные, полученные в результате, систематизированы, обобщены и представлены в следующей таблице:

Таблица 1. Количество слов в каждой из словарных групп для текстов различных функциональных стилей

№	СГ	ХТ	ПТ	НТ	ОТ	15	О	275	200	231	269
1	А	40	93	56	33	16	П	329	360	245	277
2	Б	121	124	111	15	17	Р	116	86	107	277
3	В	315	351	183	202	18	С	356	407	390	256
4	Г	107	92	12	38	19	Т	158	153	99	247
5	Д	116	159	160	119	20	У	74	87	94	61
6	ЕЁ	90	84	43	30	21	Ф	3	15	35	68
7	Ж	35	63	9	2	22	Х	22	34	20	11
8	З	92	87	55	63	23	Ц	10	12	1	7
9	И	451	273	279	321	24	Ч	92	111	77	44
10	Й	0	0	0	0	25	Ш	22	11	5	2
11	К	233	189	188	93	26	Щ	3	1	0	0
12	Л	100	89	56	58	27	Э	31	65	34	10
13	М	137	134	191	86	28	Ю	15	4	0	7
14	Н	325	437	178	197	29	Я	37	39	18	13

Примечание. Условные обозначения в таблице: СГ – словарная группа, ХТ – художественный текст, ПТ – публицистический текст, НТ – научный текст, ОТ – официально-деловой текст

Полученные статистические данные позволяют нам произвести расчет ряда других параметров, таких, как:

1) общее количество слов, относящихся к слабо полисемантизованным группам (КССП);

2) общее количество слов, относящихся к сильно полисемантизованным группам (КССиП);

3) массовая доля слов, относящихся к слабо полисемантизованным группам (% СП), рассчитываемая по формуле % СП = КССП/КС * 100 %, где КС – общее количество слов в тексте;

4) массовая доля слов, относящихся к слабо полисемантизованным группам (% СиП) , рассчитываемая по формуле % СП = КССиП/КС * 100 %, где КС – общее количество слов в тексте;

5) отношение количества слов в сильно полисемантизованных группах к количеству слов в слабо полисемантизованных группах (СиП/СП).

Произведя необходимые расчеты, представим итоговый результат в следующей таблице:

Таблица 2. Количественные характеристики текстов, принадлежащих к различным функциональным стилям

Текст	КС	СДС	КССП	КССиП	%СП	%СиП	СиП/СП
ХТ	4123	5,1	774	341	20,89%	9,20%	0,44
ПТ	3929	5,6	742	397	19,73%	10,56%	0,54
НТ	3215	6,6	662	242	23,04%	8,42%	0,37
ОТ	3046	7,1	553	157	19,71%	5,60%	0,28

Согласно данным в таблице 2, художественный и публицистический тексты характеризуются относительно низким показателем средней длины слова (в диапазоне от 5 до 6 знаков), в то время как научный и официально-деловой тексты имеют относительно высокие показатели СДС – от 6 до 7 знаков; соотношения между массовыми долями сильно полисемантизо-

ванных и слабо полисемантизованных групп, а также между количеством слов, принадлежащих к сильно и слабо полисемантизованным словарным группам, различаются для текстов всех стилей. Считаем необходимым особо отметить следующие факты:

1. Разница между % СП и % СиП видна довольно отчетливо. Для научного и официально-делового текстов она составляет 14,62% и 14,11% соответственно, в то время как для художественного и публицистического – 11,69% и 9,18%.

2. По параметру % СиП наблюдается определенное расхождение. Наибольшая разница существует между публицистическим и официально-деловым стилем: 4,96%. Параметр позволяет уверенно отличить официально-деловой текст как от публицистического, так и от художественного; в свою очередь, художественный стиль заметно отличен от официально-делового, однако разница по параметру % СиП между ним и научным стилем относительно невелика – 0,78%.

3. Соотношение СиП / СП демонстрирует полностью аналогичные закономерности. Наибольшая разница выявлена между публицистическим и официально-деловым стилями; она составляет 0,25.

Суммируя вышеизложенное и данные таблицы 2, мы приходим к предварительному заключению, что МФСП-алгоритм позволит с определенной вероятностью отличать художественные и публицистические тексты с одной стороны от научных и официально-деловых с другой стороны. Различение можно проводить по параметрам СДС, % СП, % СиП, СиП / СП.

В то же время представленные данные не всегда позволяют говорить об уверенном детектировании тех или иных стилей ввиду близости детектирующих показателей. Этот факт свидетельствует о том, что, хотя произведенных расчетов может быть достаточно для разграничения функциональных стилей, в ряде случаев для уверенной идентификации стилевой принадлежности текста может быть необходим более детализированный алгоритм. Представляется допустимой попытка выделения наиболее достоверных идентификаторов стилевой отнесенности на базе уже полученных и представленных нами в таблицах выше.

Действительно, рассмотрим массовую долю сильно и слабо полисемантизованных словарных групп в каждом из текстов.

Таблица 3. Массовые доли сильно и слабо полисемантизованных словарных групп в текстах различных функциональных стилей

СП-группа	**А**	**Б**	**Д**	**И**	**Й**	**Э**	**Ю**
ХТ	1,08%	3,27%	3,13%	12,17%	0,00%	0,84%	0,40%
ПТ	2,47%	3,30%	4,23%	7,26%	0,00%	1,73%	0,11%
НТ	1,95%	3,86%	5,57%	9,71%	0,00%	1,18%	0,00%
ОТ	1,18%	0,53%	4,24%	11,44%	0,00%	0,36%	0,25%
СиП-группа	**ЕЁ**	**Ж**	**У**	**Ц**	**Ч**	**Щ**	**Я**
ХТ	2,43%	0,94%	2,00%	0,27%	2,48%	0,08%	1,00%
ПТ	2,23%	1,68%	2,31%	0,32%	2,95%	0,03%	1,04%
НТ	1,50%	0,31%	3,27%	0,03%	2,68%	0,00%	0,63%
ОТ	1,07%	0,07%	2,17%	0,25%	1,57%	0,00%	0,46%

Можно видеть, что устойчивые закономерности для массовых долей в различных стилях речи демонстрируют словарные группы «ЕЁ», «Ж», с некоторыми оговорками – «Ц», а также «Щ» и «Я». Для этих групп массовые доли в художественном и публицистическом, равно как и в научном и официально-деловом текстах, с одной стороны, сопоставимы, а, с другой стороны, явным образом отличаются от противоположной пары стилей. Соответственно, в случае существования неопределенности основной алгоритм детектирования может быть дополнен вспомогательным алгоритмом вычисления суммарной массовой доли перечисленных словарных групп.

В качестве доказательства рассмотрим данные, полученные по этому алгоритму для рабочих текстов: научный текст – 2,47%; официально-деловой текст – 1,85%; художественный текст – 4,72%; публицистический текст – 5,29%. Предположительно пороговый уровень может находиться в пределах от 3 до 4% .

Таким образом, предварительные расчеты показывают, что по ряду параметров даже при самых приблизительных вычислениях существует принципиальная возможность различения

основных книжных стилей речи. Полученные результаты мы полагаем достаточно успешными для продолжения последующих изысканий в данном направлении.

Данные, отраженные в таблицах 1-3, необходимо и обязательно воспринимать как наиболее приблизительные, поскольку под-алгоритм, по которому мы находили количество слов из тех или иных словарных групп, предельно минималистичен и изначально предназначен лишь для грубой оценки потенциала второго детектирующего критерия. Разработка же программного обеспечения, реализующего МФСП-алгоритм, потребует предварительной обработки текста и введения ряда существенных ограничений, поскольку полноценная и качественная процедура подсчета слов в конечном счете не может быть сведена исключительно к поиску и протоколированию знакосочетаний «пробел + первая буква группы»; такой метод пригоден лишь для вынесения предварительных общих оценок, что и было нами осуществлено. После уточнения алгоритма и введения процедур обработки текста перед анализом мы сможем ожидать получения более точных и характерных показателей за счет существенного снижения погрешности вычислений.

Комментируя сказанное выше, необходимо пояснить, что использованный нами упрощенный под-алгоритм не учитывает ряд вариантов. Так, комбинация «пробел + первая буква слова» не позволяет обнаружить лексические единицы, с которых начинаются абзацы либо фразы, заключенные в кавычки или скобки, а также выявить имена собственные и аббревиатуры (которые за редким исключением не имеют никакого отношения к оценке потенциальной полисемантичности). Кроме того, в подобном алгоритме не учитывается тематика текста, которая способна существенно повлиять на итоги расчетов, равно как и необходимость исключения служебных слов, которые, подобно именам собственным, не имеют семантического аспекта значения. Все эти и некоторые другие проблемы необходимо решить прежде, чем машина приступит непосредственно к выполнению базовых расчетов и интерпретации их результатов.

Таким образом, необходимым структурным элементом МФСП-алгоритма представляется под-алгоритм предвари-

тельной обработки текстов, с разработки которого мы и полагаем необходимым начать.

В соответствии с изложенными выше соображениями текст, поступающий на вход анализатора, необходимо очистить от потенциальных помех и подготовить к формально-семантическому исследованию. Представим требуемую последовательность действий в виде ряда шагов.

1. Удаление знаков препинания, создающих помехи:

1.1. Удаление кавычек (одинарных, двойных, угловых).

1.2. Удаление скобок (круглых, квадратных, фигурных, угловых).

1.3. Удаление черточек и тире.

2. Удаление особых знаков, создающих помехи:

2.1. Замена непечатаемых символов переноса строки (знак абзаца) на стандартный пробел.

3. Удаление имен собственных и аббревиатур:

3.1. Удаление знакосочетаний вида “прописная буква + точка”.

3.2. Удаление слов, начинающихся с прописной буквы, если перед ними не обнаружена точка.

3.3. Удаление слов, набранных исключительно прописными буквами.

4. Удаление служебных слов:

4.1. Удаление предлогов, союзов и частиц в соответствии с предварительно заданным списком.

Алгоритм считаем необходимым реализовать в виде последовательного циклического поиска и обработки определенных знакосочетаний в предъявленном тексте по каждому из шагов. Обработанный материал на выходе будет подаваться на вход собственно МФСП-анализатора; как уже было сказано выше, посредством предварительной обработки мы ожидаем получить более точные и корректные результаты, равно как и минимизировать погрешность вычислений.

Изобразим блок-схематическое представление поискового цикла:

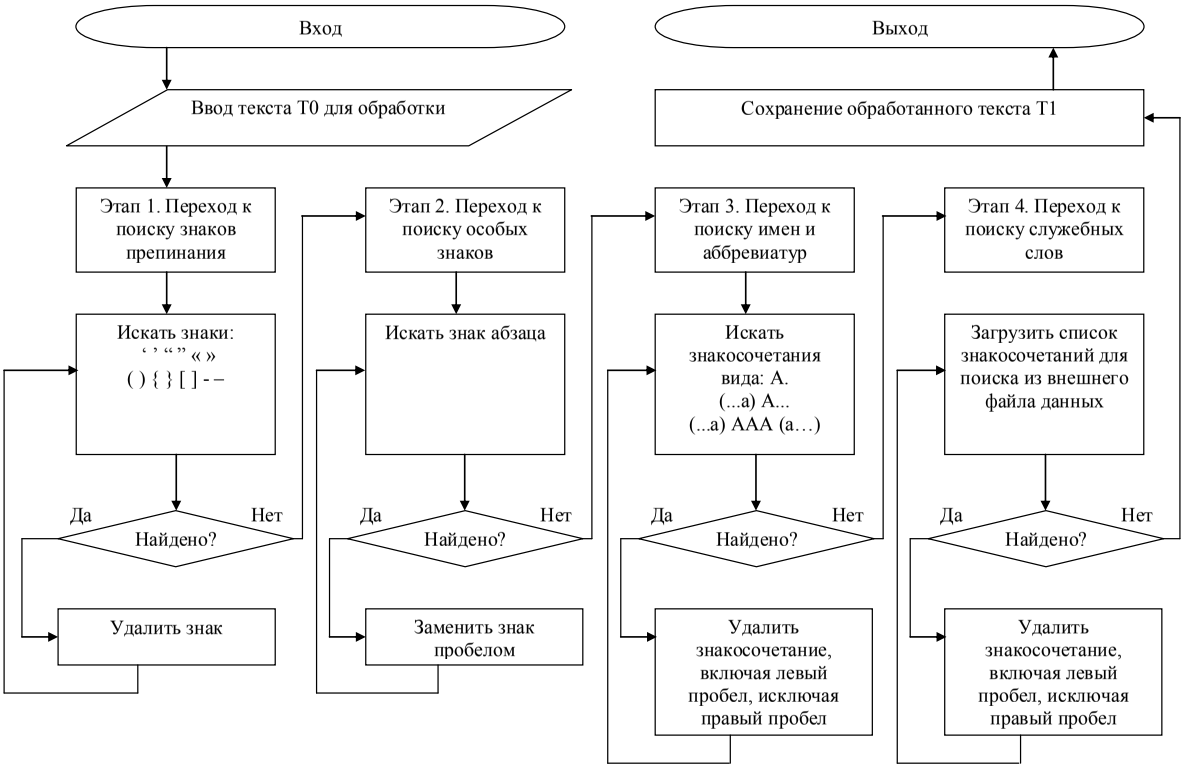

Рис. 1. Блок-схема алгоритма предварительной обработки

Заметим, что наиболее существенный результат мы рассчитываем получить за счет очистки исследуемого текста от служебных слов – в первую очередь союзов «а» и «и», активное использование которых способно внести определенную погрешность в оценку массовой доли слабо полисемантизованных словарных групп.

Важно, однако, указать, что механизм предварительной обработки не может быть реализован в качестве отдельного модуля перспективного программного анализатора: он должен быть органично встроен в процесс исследования текста. Для простоты изложения мы изобразили его в виде отдельного цикла, однако при разработке программного обеспечения, реализующего МФСП-алгоритм, цикл очистки не будет замкнутым и обособленным; напротив, он будет являться конструктивным элементом общей последовательности машинно-логических действий.

В первую очередь необходимость интеграции связана с тем, что, в то время как первые два шага предварительной обработки текста представляют собой внесение в исследуемый материал исключительно формальных изменений, упрощающих процедуру обнаружения слов, вторые два шага подразумевают уничтожение определенных лексических единиц, не имеющих собственной семантики. Иными словами, шаги 3 и 4 вносят в текст количественные изменения с целью более эффективного определения его качественных параметров; это полезно с позиции второго детектирующего критерия (измерения потенциальной полисемантичности текста), но с точки зрения первого критерия (расчета средней длины слова) принципиально неприемлемо. В силу этого, строго говоря, перед запуском исследования по МФСП-алгоритму будут работать только шаги 1-2; шаги 3-4 будут подключены уже в процессе.

Теперь нам необходимо перейти к формированию основной последовательности операций для последующего построения на ее основе аналитического программного обеспечения. Данный алгоритм со всей очевидностью окажется существенно более сложным, поскольку в нем необходимо учесть ряд необходимых условий и описать некоторые нестандартные процедуры. Полагаем целесообразным представить алгоритм в виде следующего набора шагов:

1. Расчет данных по первому детектирующему критерию (средняя длина слова).

1.1. Рассчитать общее количество любых символов в тексте.

1.2. Рассчитать количество знаков препинания в тексте в соответствии с заданным списком знаков, включая пробелы.

1.3. На основании разности данных, полученных на этапах 1.1 и 1.2, определить суммарное количество неслужебных символов, входящих в состав слов.

1.4. Рассчитать количество неразрывных знакосочетаний, начинающихся с пробела и заканчивающихся другим пробелом либо служебным символом (точкой, запятой, двоеточием, точкой с запятой, вопросительным или восклицательным знаком).

1.5. Найти среднюю длину слова на основании частного данных, полученных на этапах 1.3 и 1.4.

2. Расчет данных по второму детектирующему критерию (потенциальная полисемантичность).

2.1. Рассчитать количество знакосочетаний вида «пробел + буква» для каждой словарной группы по алфавиту.

2.2. Рассчитать массовую долю (индивидуальную и суммарную) для характеристических словарных групп на основании данных, полученных на этапах 1.4 и 2.1.

3. Анализ данных по обоим детектирующим критериям.

3.1. Предварительно определить ФС исследуемого текста на основании сравнения данных, полученных на этапе 1.5, с заданным диапазоном средней длины слова.

3.2. Предварительно определить ФС исследуемого текста на основании сравнения данных, полученных на этапе 2.2, с заданными диапазонами массовых долей характеристических словарных групп.

3.2.1. Установить особые отметки для последующего анализа в случае соответствия предварительно заданных данных о тематике текста и аномального отличия фактических массовых долей от заданных диапазонов.

3.3. Принять решение о запуске дополнительных верификационных процедур на основании сравнения и анализа вердиктов и данных по обоим детектирующим критериям.

3.4. Вынести окончательный вердикт по итогам всех расчетов и аналитических процедур третьего этапа.

Расчеты в основном алгоритме, как и в случае предварительной обработки, будут производиться на основании циклических процедур поиска определенных знакосочетаний, а также инкрементирования счетчиков. Кроме того, считаем необходимым предусмотреть протоколирование процедуры анализа на каждом из шагов.

Таким образом, мы располагаем возможностью подготовить блок-схематическое представление логики работы программного обеспечения, реализующего МФСП-алгоритм. Поскольку интегрированная блок-схема, увязывающая предварительный обработчик исследуемого материала и все три этапа основного программного модуля, неизбежно окажется излишне громоздкой, при ее построении мы ограничимся указанием мест вхождения промежуточных данных, получаемых от циклов очистки исходного текста.

Отметим, что в наши непосредственные задачи на данном этапе исследования не входит изображение полного объема возможных вариантов работы аналитического программного обеспечения. Мы считаем необходимым представить лишь основную последовательность действий, которая позволяет составить базовое представление о принципах и логике функционирования проектируемого анализатора; технические детали устройства и организации программного продукта, равно как и теоретически допустимые отклонения от изначально намеченного «курса» работы программы, могут быть исключены из представления без существенного ущерба для понимания сути МФСП-алгоритма и принципов, на которых он построен.

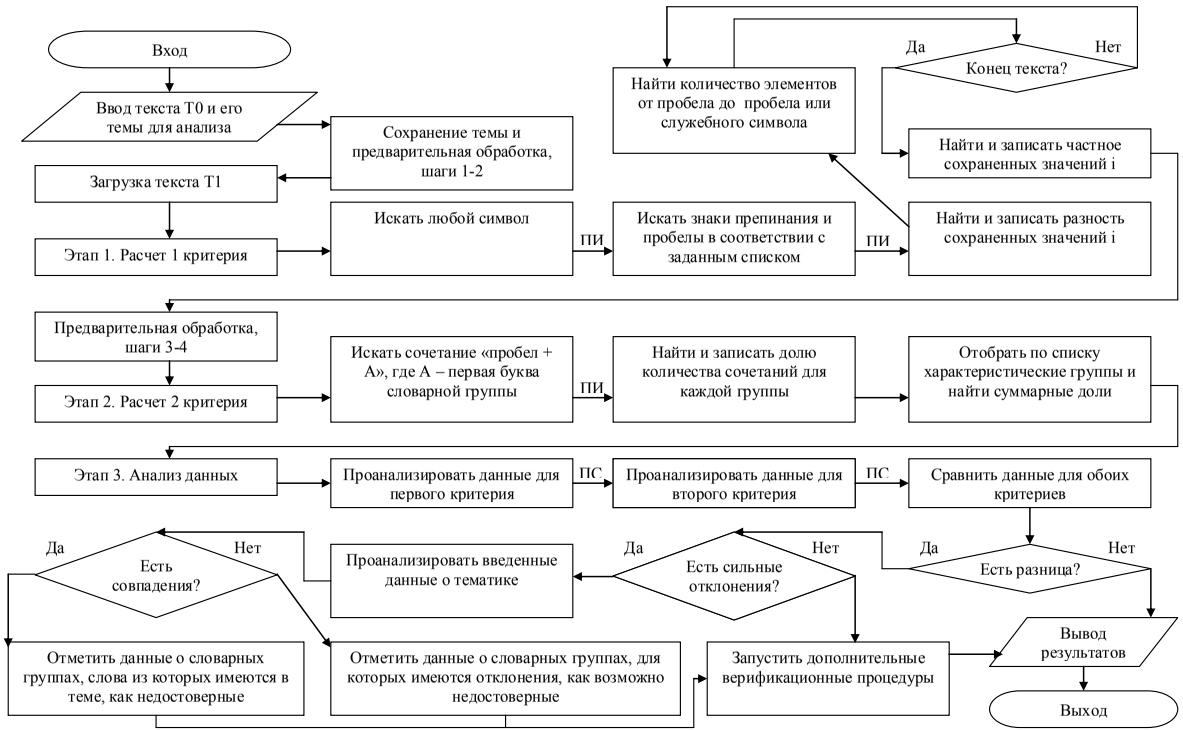

Рис. 2. Блок-схема МФСП-алгоритма

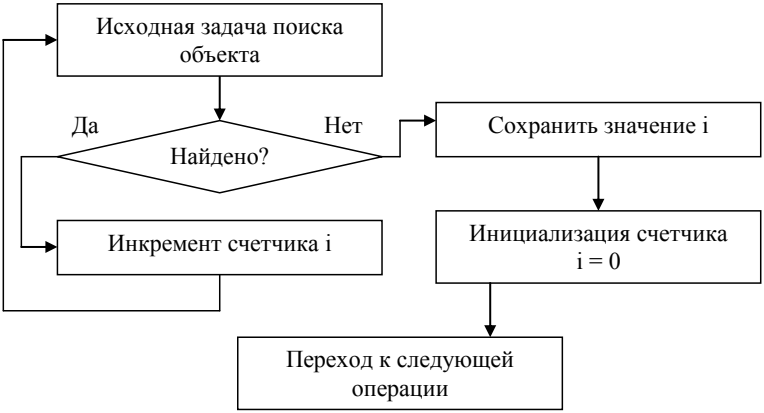

Рис. 3. Встроенная процедура инкремента (ПИ)

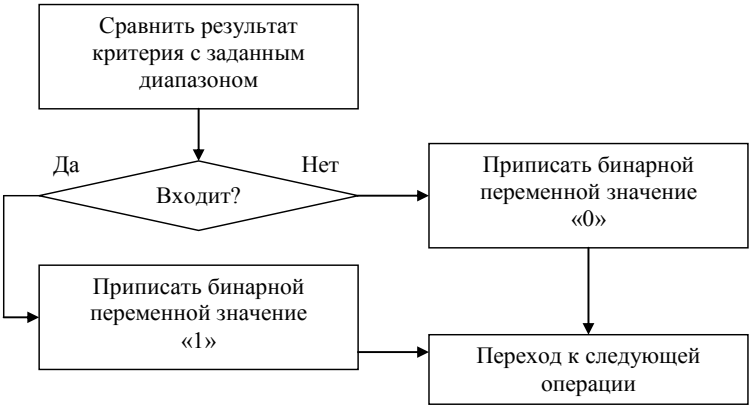

Рис. 4. Встроенная процедура сравнения (ПС)

На двух последних рисунках изображены т.н. встроенные процедуры, которые в интересах целесообразности, эргономики и компактности представления были вынесены за пределы общей блок-схемы. Места вхождения указанных процедур в основную последовательность операций на ри-

сунке 2 обозначены соответствующими аббревиатурами.

Границы базовых диапазонов мы будем определять на основании описанных выше предварительных исследований, с возможным последующим их уточнением. Таким образом, по первому детектирующему критерию (средняя длина слова) диапазон будет иметь вид (0 ... 6), так что значение «1» будет соответствовать свободным книжным стилям, а значение «0» – формальным. Что же касается второго критерия, то в этом случае мы имеем три базовых диапазона – «% СП - % СиП» (0 ... 13) , «% СиП» [9 ... +∞) , «СиП / СП» [0,4 ... +∞); результаты для каждого из диапазонов будут записываться в отдельные бинарные переменные, и окончательный вердикт будет формироваться на основании сопоставления итогов (если не менее двух переменных имеют значение «1», то и результат обработки критерия будет равен единице). Изобразим процедуру анализа и сопоставления в виде таблицы:

Таблица 4. Анализ данных для второго детектирующего критерия

Значение 1	Значение 2	Значение 3	Сумма	Принцип сравнения	Итог
0	0	0	0	Итоговое значение равно 1, если сумма больше либо равна 10. Если сумма меньше 10, то итоговое значение равно 0.	0
0	0	1	1		0
0	1	0	1		0
0	1	1	10		1
1	0	0	1		0
1	0	1	10		1
1	1	0	10		1
1	1	1	11		1

Итоговое значение «1» со всей очевидностью также будет соответствовать свободным книжным стилям, а «0» - формальным. Как видно из таблицы, единица появляется на выходе тогда и только тогда, когда не менее чем два исходных параметра имеют значение «1», что нам и требуется для успешного вынесения вердикта по второму детектирующему критерию.

Значения, полученные в результате обработки каждого из детектирующих критериев, будут сопоставлены посредством

операции суммирования. Если значение итоговой переменной составит «0» (0 + 0 = 0), то исследуемый текст будет признан относящимся к формальному книжному стилю речи – научному либо официально-деловому; при значении «10» (1 + 1 = 10) – к свободному (художественному либо публицистическому). Оставшиеся два сочетания дадут в сумме единицу (0 + 1 = 1, 1 + 0 = 1); такое значение будет свидетельствовать о наличии разницы между критериями и о необходимости запуска дополнительных верификационных процедур.

Одной из подобных процедур может являться отсечение недостоверных словарных групп на основании данных о тематике текста. Для этих целей при написании программного обеспечения мы намереваемся предусмотреть особое текстовое поле, предназначенное для указания темы либо ключевых слов; словарные группы, к которым будут принадлежать введенные в поле слова, потребуется проверить на соответствие выявленным отклонениям и при необходимости пометить как характеризующиеся сниженной либо нулевой достоверностью.

Поясним сущность функционирования данной секции общего алгоритма на примере. Предположим, что данные по двум детектирующим критериям разошлись, и на выходе получен код «1», который соответствует нечеткости вердикта – положим, средняя длина слова в тексте соответствует формальным стилям, но параметры «% СП - % СиП» и «% СиП» вошли в рамки обозначенных диапазонов, тем самым недвусмысленно свидетельствуя о наличии у текста признаков свободных стилей. В таком случае необходимо удостовериться в том, что высокие показатели сильно полисемантизованных словарных групп не обусловлены тематикой текста: так, нет сомнений в том, что, хотя словарная группа Щ характеризуется высокими показателями потенциальной полисемантичности, по массовой доле слов из этой группы научный текст из области ботаники, посвященный, к примеру, растениям семейства щавелевых, превзойдет любой художественный или публицистический текст. Располагая данными о тематике и / или ключевыми словами, алгоритм сможет отсечь аномальные показатели и выполнить повторное сопоставление результатов, полученных для обоих детектирующих критериев.

Если изначально несовпадение будет вызвано именно тематической отнесенностью, то описанный механизм устранит несоответствие.

Если же критерии будут отличаться даже после анализа тематики, может быть запущена проверка по особо достоверным словарным группам. Предварительные итоги расчетов массовых долей различных словарных групп, приведенные выше, указывают на устойчивость показателей «ЕЁ», «Ж», «Ц», «Щ» и «Я»; следовательно, вспомогательная процедура верификации будет состоять в суммировании массовых долей этих словарных групп (за исключением тех, которые будут признаны недостоверными при обработке тематики) и последующей проверке на вхождение суммы в заданный диапазон. В дальнейшем мы рассчитываем получить более точные данные, которые, в свою очередь, позволят более точно настроить разрабатываемое программное обеспечение и задать достаточно достоверные границы диапазонов для каждого конкретного случая, включая отсечение некоторых групп по тематике.

Подобная аналитическая модель представляется нам достаточно разработанной и не требующей дальнейшей детализации. Мы раскрыли сущность всех исходных ветвей и опорных точек алгоритма, которые не представлялось возможным подробно описать в пределах самого блок-схематического изображения, и на этом полагаем работу по формированию как основного алгоритма, так и вспомогательных под-алгоритмов оконченной.

Таким образом, в целом предложенное выше базовое описание МФСП-алгоритма можно считать завершенным. В нашем распоряжении на данный момент имеются предварительные результаты исследований, позволяющие первоначально задать критерии и диапазоны для проверки исследуемых текстов, нами предложен механизм предварительной очистки текстового материала от потенциальных помех и построена основная логическая последовательность действий, которые машина будет исполнять в процессе анализа. Соответственно, мы располагаем полноценным набором данных и логических построений, который позволяет перейти непосредственно к разработке программного обеспечения, реализующего МФСП-алгоритм, с его последующим детализиро-

ванным описанием и практическим применением на более масштабной выборке текстов, позволяющим уточнить границы заданных аналитических диапазонов и снизить количество вероятных ошибок, недочетов и ложных срабатываний.

3.2. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ, РЕАЛИЗУЮЩЕЕ МФСП-АЛГОРИТМ

Аналитическое приложение, реализующее МФСП-алгоритм, построено нами в визуальной среде разработки программного обеспечения для операционных систем Microsoft Windows – CodeGear RAD Studio 2007. Исходный код продукта написан на языке программирования Delphi, возможности которого оптимально подходят для создания приложений подобного рода.

В продукте реализованы все шаги, изложенные выше в блок-схематическом представлении. Приложение собрано, протестировано и готово к работе. Изложим вкратце основные сведения о разработанном программном обеспечении, а также о его ключевых особенностях и принципах функционирования.

МФСП-анализатор реализован в виде стандартного исполняемого файла операционной системы Microsoft Windows (формат EXE), сопровождаемого вспомогательными файлами данных, размещаемых в субдиректориях. В подкаталоге \HELP\ находится справочный аппарат, папка \LISTS\ является местом расположения данных, необходимых для осуществления аналитических процедур, а в директории \LOGS\ сохраняются результаты протоколирования выполняемых приложением операций.

Анализатору было присвоено нами наименование «Style Recognition System» (букв. «система распознавания стилей»), сокращенно «SRS». В дальнейшем мы будем использовать данное наименование в качестве альтернативного обозначения разработанного программного продукта. Использование англоязычного имени диктуется требованиями совместимости.

Для начала работы с анализатором необходимо запустить основной исполняемый файл – SRS.EXE. В результате этого действия на дисплее персональной ЭВМ будет отображено

главное окно приложения, содержащее все необходимые элементы для полноценного использования возможностей продукта. Внешний вид окна представлен на изображении (рис. 5).

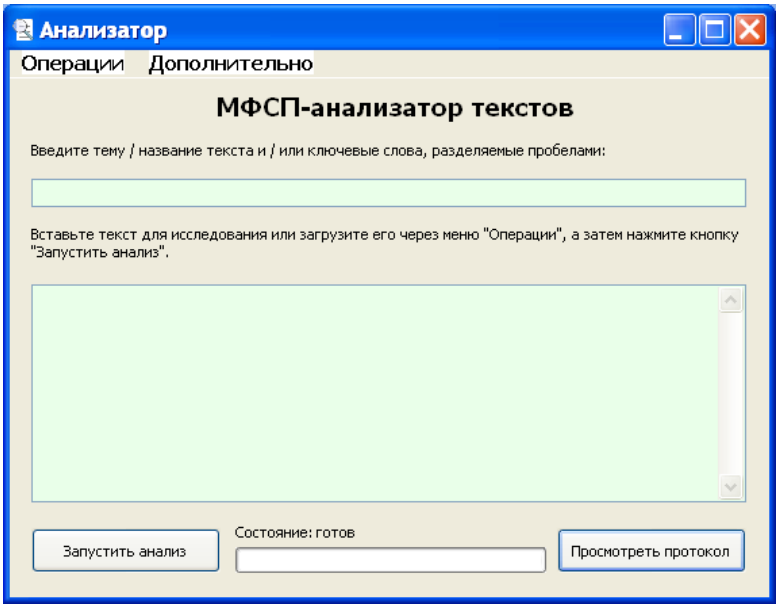

Рис. 5. Внешний вид главного окна МФСП-анализатора

Основными элементами окна являются:

1. Строка меню, содержащая средства быстрого доступа к различным функциям SRS. Элементы разделены на две группы: «Операции» и «Дополнительно». В группу «Операции» входят следующие команды:

1.1. «Загрузить текст». Данная команда позволяет открыть стандартный диалог выбора объектов файловой системы и автоматически поместить анализатор текст для исследования. К загрузке принимаются файлы в формате TXT (текстовые файлы).

1.2. «Вставить текст». Данная команда позволяет поместить в анализатор текст, находящийся в буфере обмена операционной системы Windows, если таковой там присутствует.

1.3. «Пакетная обработка». Инструмент последовательного анализа текстов позволяет в полуавтоматическом режиме исследовать сразу несколько образцов, находящихся в одном и том же каталоге. После назначения необходимой директории продукт извлекает список расположенных в ней текстовых файлов (как и в случае с командой 1.1, к загрузке принимаются только текстовые файлы, сохраненные в формате TXT) и начинает циклическую поочередную обработку каждого из них. Содержимое объекта при этом загружается в основное текстовое поле анализатора, а имя файла автоматически подставляется в строку ключевых слов. После завершения анализа объекта SRS переходит к следующему файлу. При этом следует отметить, что уведомления о завершении исследования не отображаются; итоги изучения совокупности текстов доступны в окне протокола (см. далее).

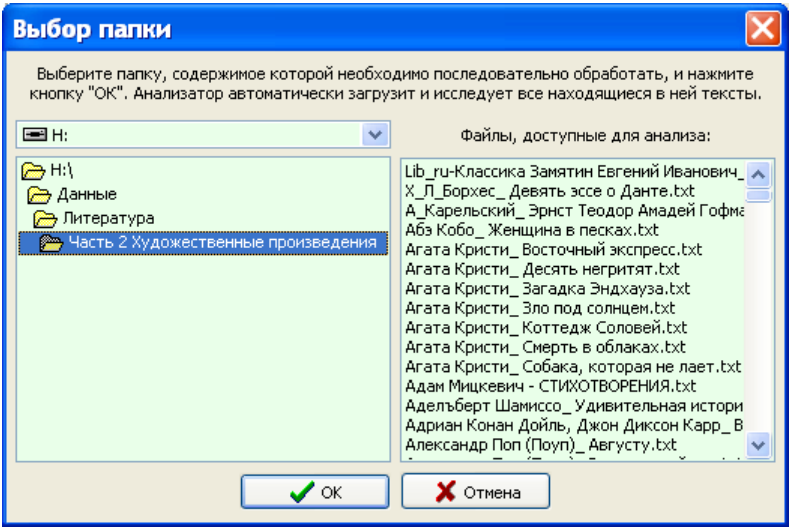

Рис. 6. Окно выбора папки для пакетного анализа файлов

Полагаем целесообразным отметить, что описанный выше инструмент пакетной обработки является демонстрацией принципиальной возможности выхода на потенциальное применение SRS в качестве компонента комплексных систем автоматизированного исследования целевых текстов, где мо-

жет существовать потребность в последовательном изучении и последующей классификации множества образцов.

1.3. «Запустить анализ». Данная команда инициирует процесс автоматического исследования текста, находящегося в анализаторе, при условии, что таковой был введен, вставлен либо загружен пользователем.

1.4. «Просмотреть протокол». Данная команда открывает окно просмотра и сохранения результатов протоколирования действий, выполняемых программным продуктом.

1.5. «Инструменты». Данная команда предоставляет доступ к специализированному средству поиска, позволяющему рассчитывать отдельные параметры текста без запуска процедур автоматического анализа (см. далее).

1.6. «Выход». Данная команда позволяет прекратить работу с приложением.

В группу «Дополнительно» входят следующие команды:

1.7. «Справка». Данная команда позволяет активировать справочный аппарат SRS.

1.8. «О программе». Данная команда вызывает краткие сведения о продукте.

2. Поле для ввода названия текста, ключевых слов либо иных сведений о тематике. Настоятельная необходимость использования ключевых слов в процессе анализа была подробно освещена нами выше.

3. Поле для ввода исследуемого текста. Текст для анализа можно поместить в поле при помощи пункта меню «Загрузить файл», либо скопировать его из произвольного текстового редактора и вставить стандартными средствами операционной системы Windows или соответствующей команды в строке меню.

4. Управляющие кнопки «Запустить анализ» и «Просмотреть протокол». По своему назначению они аналогичны одноименным командам в строке меню.

5. Индикатор выполнения («прогресс-бар»). Данный компонент в режиме реального времени отображает, на какой стадии находится автоматический анализ текста.

При активации команды «Просмотреть протокол» открывается окно просмотра и сохранения отчетов. Внешний вид данного окна представлен на изображении (рис. 7).

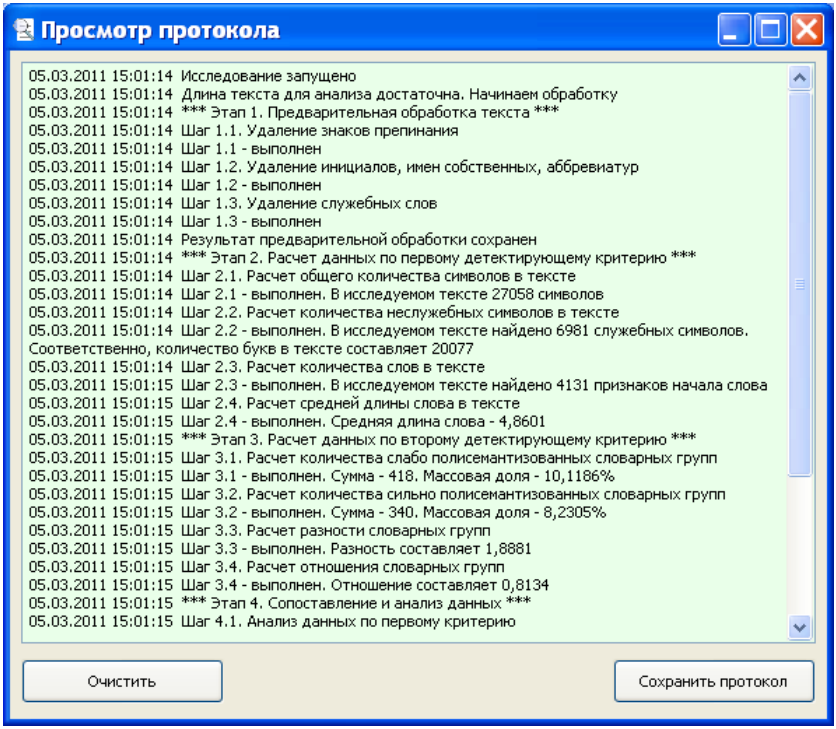

Рис. 7. Внешний вид окна просмотра и сохранения отчетов

Рассматриваемое окно содержит три базовых элемента:

1. Поле для ввода уведомлений. В процессе работы МФСП-анализатор автоматически помещает в это поле информационные сообщения о ходе исследования, формируя тем самым протокол. Каждое сообщение сопровождается отметкой о дате и времени его наступления.

2. Кнопка «Очистить протокол». Данная команда позволяет стереть все содержимое поля ввода.

3. Кнопка «Сохранить протокол». Данная команда вызывает стандартное диалоговое окно, позволяющее сохранить содержимое поля в текстовый файл (формат TXT).

При вызове команды «Инструменты» в строке меню активируется окно специализированных средств поиска. Внешний вид данного окна представлен на изображении (рис. 8).

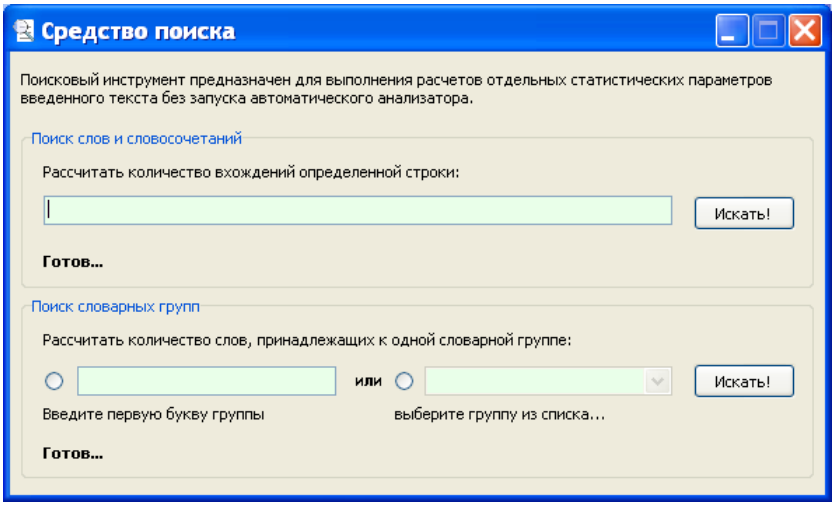

Рис. 8. Внешний вид окна средств поиска

В рассматриваемом окне представлены два поисковых инструмента, каждый из которых размещается на своей группирующей панели.

1. Поиск слов и словосочетаний. Данный инструмент позволяет подсчитывать количество вхождений заданной строки в текст, который находится в анализаторе. При этом не учитываются и не задействуются никакие корректирующие механизмы; к примеру, если в строку поиска будет введено значение «а», то будут учтены вообще все буквы «а» в тексте. Поиск является регистрозависимым, т.е. прописные буквы отличаются от строчных.

2. Поиск словарных групп. Данный инструмент позволяет рассчитывать суммарное количество и массовую долю слов, относящихся к определенным словарным группам. Первая буква группы для поиска может быть введена вручную либо выбрана из предлагаемого списка. Необходимо заметить, что данный инструмент запускает некоторые процедуры предварительной обработки текста, поэтому выдаваемые им результаты могут существенно отличаться от итогов работы первого инструмента.

В обоих случаях поиск запускается посредством кнопки «Искать»; финальные значения счетчиков выводятся в том же окне под текстовыми строками.

Отдельно рассматривать ход автоматического анализа нам не представляется целесообразным, поскольку он не характеризуется сколь-либо значимыми отличиями от разработанных и визуализированных нами ранее алгоритмов. Тем не менее, полагаем необходимым упомянуть о процедурах, сопровождающих начальный и завершающий этапы исследования.

Перед запуском анализа SRS осуществляет проверку введенного в анализатор текста на соответствие трем формальным параметрам: собственно наличие текста (т.е. объект исследования загружен, вставлен либо введен в поле), минимальная длина (в интересах репрезентативности рассчитываемых показателей изучаемый текст должен состоять из более чем 5 000 знаков), присутствие ключевых слов или сведений о тематике. Если какое-либо из указанных требований не выполнено, автоматический анализ не будет запущен, а на дисплей будет выведено соответствующее предупреждение. Пользователь может отказаться от ввода ключевых слов, поместив в соответствующее поле простой одинарный пробел, хотя делать это не рекомендуется.

После завершения всех аналитических процедур продукт предпримет попытку вынести окончательный вердикт. Если результаты по обоим диагностическим критериям будут совпадать, то на дисплее будет отображено уведомление о завершении работы анализатора, содержащее в том числе и итог исследования; если же критерии будут противоречить друг другу, и уточняющий анализ не позволит снять указанное противоречие, то SRS сообщит о невозможности осуществить надлежащую классификацию текста и пояснит вероятные причины (рис. 9, 10).

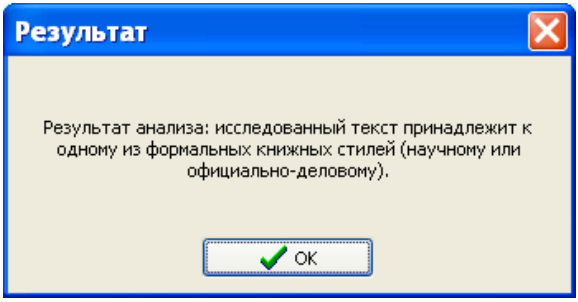

Рис. 9. Итог успешного исследования текста

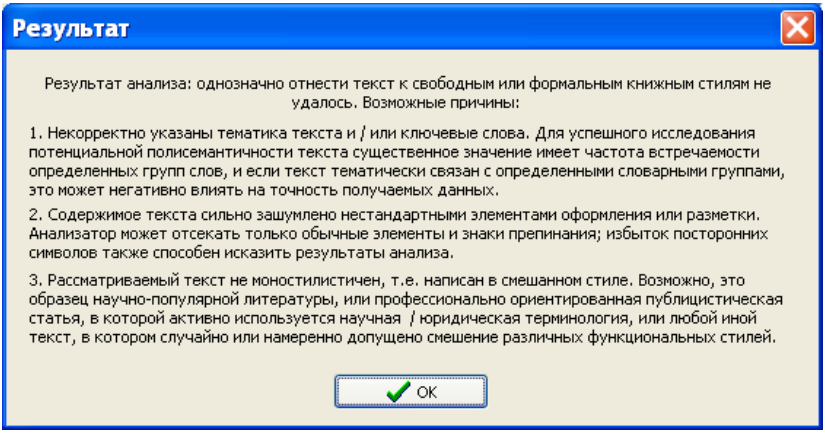

Рис. 10. Уведомление о невозможности однозначно классифицировать текст

Заметим также, что приложение обладает базовой защитой от попыток подвергнуть анализу иноязычные тексты. Вывод о некорректности языка делается на основании результатов расчетов по характеристическим словарным группам (рис. 11).

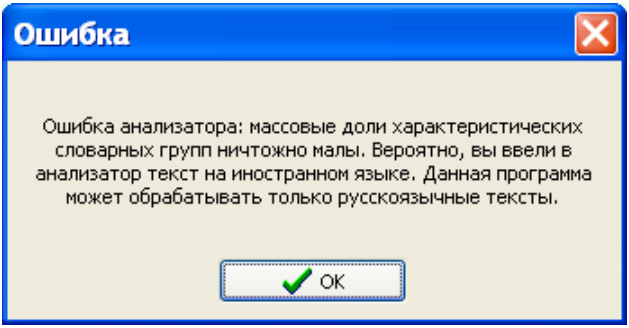

Рис. 11. Уведомление о вероятной попытке проанализировать иноязычный текст

При построении первого действующего образца анализатора его база данных и внутренние инструкции основывались на предварительных результатах, полученных нами ранее при

проверочном изучении четырех эталонных текстов. Выборочное тестирование показало, что в целом программное обеспечение функционировало успешно даже при использовании ориентировочных диагностических значений, найденных с определенной погрешностью ввиду отсутствия надлежащих средств оперативной обработки и очистки исследуемых текстов от помех. Тем не менее, в силу того, что таковое средство было реализовано в SRS, нами было осуществлено уточняющее исследование тех же опорных текстов, которое проводилось уже не в текстовом процессоре Microsoft Word 2003, как при предварительном их изучении, а собственно в построенном МФСП-анализаторе, чьи алгоритмы обеспечивают получение более надежных и достоверных сведений.

Поскольку предварительная обработка влияет главным образом на расчеты по второму диагностическому критерию, пороговый показатель для первого критерия (средней длины слова) не подвергался модификациям и остался на уровне 6 знаков. В то же время необходимо заметить, что вычисляемая SRS средняя длина слова несколько ниже, чем аналогичное значение, рассчитываемое на основании данных от редактора Word 2003; это связано преимущественно с тем, что МФСП-анализатор обрабатывает только буквенные символы, в то время как текстовый процессор Microsoft осуществляет более грубую оценку, включая в расчеты знаки препинания. Кроме того, SRS показывает несколько иное количество слов, нежели Word 2003, что обусловлено работой внутреннего алгоритма фильтрации. Тем не менее, радикального влияния на вычисляемый показатель это не произвело (табл. 5).

Таблица 5. Различия между количественными показателями средней длины слова, полученными при помощи SRS и Word 2003

Текст	Word 2003		SRS	
	КС	СДС	КС	СДС
ХТ	4123	5,1	4131	4,9
ПТ	3929	5,6	3987	5,2
НТ	3215	6,6	3205	6,3
ОТ	3046	7,1	3034	6,9

Что же касается второго диагностического критерия, то уточнение и предварительная очистка также оказали определенное воздействие на результаты. Тенденции, выявленные при первичной оценке параметров текстов, сохранились и в большинстве случаев стали более выраженными (табл. 6).

Таблица 6. Различия между количественными показателями потенциальной полисемантичности, полученными при помощи SRS и Word 2003

	Текст	КССП	КССиП	%СП	%СиП	%СП-%СиП	СиП/СП
Word 2003	ХТ	774	341	20,89%	9,20%	11,69%	0,44
	ПТ	742	397	19,73%	10,56%	9,17%	0,54
	НТ	662	242	23,04%	8,42%	14,62%	0,37
	ОТ	553	157	19,71%	5,60%	14,11%	0,28
SRS	ХТ	390	340	9,44%	8,23%	1,21%	0,87
	ПТ	372	346	9,33%	8,68%	0,65%	0,93
	НТ	386	238	12,04%	7,43%	4,62%	0,62
	ОТ	312	155	10,28%	5,11%	5,18%	0,50

Исходя из уточненных сведений, в границы диапазонов были внесены требуемые модификации. Новые контрольные значения и числовые отрезки приобрели следующий вид:

1) разность массовых долей слабо и сильно полисемантизованных словарных групп – (0 ... 3];

2) массовая доля слов, принадлежащих к сильно полисемантизованным словарным группам – [6,75 ... +∞);

3) отношение суммарного количества слов из сильно полисемантизованных словарных групп к суммарному количеству слов из слабо полисемантизованных словарных групп – [0,75 ... +∞).

Заметим, что, несмотря на существенную модификацию пороговых показателей, общий принцип диагностирования по второму критерию остался неизменен. Если найденный показатель входит в обозначенный диапазон, то промежуточной переменной присваивается значение «1», соответствующее свободным стилям речи, и наоборот; окончательный вер-

дикт принимается на основании преобладающих значений.

Изменения произошли также и в индивидуальных показателях характеристических словарных групп. После обработки эталонных текстов в МФСП-анализаторе и определения их массовых долей при помощи поискового инструмента были получены новые данные (табл. 7).

Таблица 7. Различия между количественными показателями диагностически значимых словарных групп, полученными при помощи SRS и Word 2003

Word 2003	**СП-группа**	**А**	**Б**	**Д**	**И**	**Й**	**Э**	**Ю**
	ХТ	1,08%	3,27%	3,13%	12,17%	0,00%	0,84%	0,40%
	ПТ	2,47%	3,30%	4,23%	7,26%	0,00%	1,73%	0,11%
	НТ	1,95%	3,86%	5,57%	9,71%	0,00%	1,18%	0,00%
	ОТ	1,18%	0,53%	4,24%	11,44%	0,00%	0,36%	0,25%
	СиП-группа	**ЕЁ**	**Ж**	**У**	**Ц**	**Ч**	**Щ**	**Я**
	ХТ	2,43%	0,94%	2,00%	0,27%	2,48%	0,08%	1,00%
	ПТ	2,23%	1,68%	2,31%	0,32%	2,95%	0,03%	1,04%
	НТ	1,50%	0,31%	3,27%	0,03%	2,68%	0,00%	0,63%
	ОТ	1,07%	0,07%	2,17%	0,25%	1,57%	0,00%	0,46%
SRS	**СП-группа**	**А**	**Б**	**Д**	**И**	**Й**	**Э**	**Ю**
	ХТ	0,48%	3,03%	2,71%	2,78%	0,00%	0,07%	0,36%
	ПТ	0,83%	2,53%	3,26%	2,33%	0,00%	0,30%	0,08%
	НТ	1,28%	2,25%	3,84%	4,52%	0,00%	0,15%	0,00%
	ОТ	0,59%	0,59%	3,59%	5,04%	0,00%	0,23%	0,23%
	СиП-группа	**ЕЁ**	**Ж**	**У**	**Ц**	**Ч**	**Щ**	**Я**
	ХТ	2,23%	0,85%	1,72%	0,29%	2,28%	0,07%	0,80%
	ПТ	1,81%	1,13%	1,86%	0,25%	2,68%	0,03%	0,93%
	НТ	1,28%	0,06%	3,03%	0,03%	2,46%	0,00%	0,56%
	ОТ	0,99%	0,07%	1,81%	0,23%	1,58%	0,00%	0,43%

Считаем необходимым отметить, что в результате уточняющего анализа массовые доли многих словарных групп стали более четко различимы. Наиболее заметным следует признать разделение для слабо полисемантизованной группы «И», ко-

торая, таким образом, приобрела потенциал, в перспективе позволяющий использовать ее в качестве надежной словарной группы при вспомогательном исследовании текстов с противоречием базовых диагностических критериев; определенные изменения произошли и в случае групп «Д» и «Э». Отличия в параметрах иных надежных словарных групп также оказались более выраженными, что, в свою очередь, привело к уточнению их пороговых значений.

Отметим, что работа с надежными словарными группами в итоговом алгоритме построена не на суммировании показателей, а на поочередной сверке индивидуальных показателей каждой из групп с установленным эталоном. Подобная реализация позволяет учитывать введенные перед запуском анализа ключевые слова и сведения о тематике, за счет чего достигается отсечение недоверенных словарных групп. Вердикт при этом выносится в пользу того типа функционального стиля, на который указывает более половины значений промежуточных переменных.

В результате были установлены следующие диапазоны для надежных словарных групп:

1) ЕЁ – [1,54 ... +∞); 2) Ж – [0,5 ... +∞); 3) Ц – [0,235 ... +∞); 4) Щ – [0,0001 ... +∞); 5) Я – [0,68 ... +∞).

Таким образом, по итогам верификационного анализа на материале текстов, взятых за эталон, нами были перепроверены и уточнены основные численные показатели, на базе которых строится автоматическое детектирование типа функционального стиля текста. Введение системы предварительной очистки материала от элементов, создающих помехи для успешной работы исследовательского приложения, позволило снизить погрешности вычислений и получить более достоверные данные по обоим диагностическим критериям. Вследствие применения преданалитических процедур фильтрации выявленные нами ранее тенденции оказались более четко выраженными, а некоторые данные претерпели весьма существенные изменения; в частности, отдельные словарные группы приобрели характер надежных.

После обновления статистических таблиц были внесены соответствующие модификации в собственную базу данных МФСП-анализатора, равно как и в его внутренние алгорит-

мы, обеспечивающие проведение автоматического исследования. Вследствие этого детектирующие механизмы продукта стали функционировать более четко и точно, позволяя достичь уверенного разделения вводимых в SRS текстов на типы в соответствии с признаками их функционально-стилевой принадлежности.

Итак, разработанный продукт построен, протестирован, отлажен и готов к проверке на широкой выборке текстов.

3.3. РЕЗУЛЬТАТЫ ОПЫТНОГО ИСПОЛЬЗОВАНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ, РЕАЛИЗУЮЩЕГО МФСП-АЛГОРИТМ

Непосредственной задачей опытной проверки МФСП-анализатора на широкой выборке текстов является не только изучение эффективности разработанного программного обеспечения, но и сбор более объемных и детализированных статистических данных, которые позволили бы убедиться в точности предложенных критериев и в случае необходимости произвести надлежащую корректировку. Поскольку опорные значения, на основании которых выносятся вердикты SRS, были получены на базе малой выборки из четырех текстов, то сколь бы ни была существенна репрезентативность указанных текстов и собранных статистических сведений, размер выборки не позволяет считать ее единственно верным образцом и требует проведения расширенных исследований на материале более объемной совокупности образцов.

Тем не менее, одновременно со сбором данных возможно решить и задачу оценивания точности функционирования аналитического алгоритма на основании тех данных, каковые имеются в его распоряжении на данный момент. Вполне очевидно, что нам необходимо убедиться в принципиальной способности анализатора классифицировать художественные и публицистические тексты как свободные, а научные и официально-деловые – как формальные, сохраняя при этом достаточно высокую степень достоверности и низкую долю ложных либо нечетких вердиктов.

При этом мы не исключаем, что алгоритм может подвергнуться дальнейшей модификации с учетом как собственно

статистических сведений, так и итогов проверки на точность. Внесение коррективов по результатам опытного применения программы представляется нам естественным конструктивным элементом цикла разработки программного продукта с учетом тех особенностей определения его действующих диагностических параметров, которые были обозначены выше.

Нами было принято решение установить суммарный объем выборки текстов на уровне 100 штук. В совокупность образцов, соответственно, будут в равных количествах отобраны все четыре рассматриваемых нами функциональных стиля – размер каждой группы составит по 25 текстов. Материал будет отбираться нами в произвольном порядке без явных требований к объему, жанру и преобладающей тематике, однако в то же время мы предпримем необходимые усилия по обеспечению надлежащего многообразия выборки по упомянутым выше аспектам; это позволит повысить репрезентативность корпуса текстов и получить более достоверные данные, крайне слабо зависящие от конкретных особенностей тех или иных подвергаемых изучению образцов.

Поскольку выборка должна не только носить диагностический характер, но и обеспечивать сбор и анализ статистических сведений для возможного уточнения критериев исследования МФСП-алгоритма, мы считаем необходимым обеспечить определенную стерильность обрабатываемого материала. Под этим мы подразумеваем размещение в выборке преимущественно таких текстов, которые можно считать достаточно чистыми образцами вышеобозначенных стилей, в меру возможности, разумности и достаточности воздерживаясь от внесения в корпус полистилистичных текстов, т.е. промежуточных вариантов, которые располагаются на стыке двух или более функциональных стилей.

В связи с этим требует рассмотрения вопрос о принципиальной неоднородности публицистического стиля. Согласно исследованиям М.Н. Кожиной, которым мы ранее следовали при выборе доминирующей классификации стилей, публицистика характеризуется порой существенным разбросом показателей и значительной открытостью для разнообразных элементов иных ФС – таких, к примеру, как речевые обороты и фразеологизмы, происходящие из формализованных стилей.

Теоретически описанная специфика может негативно влиять на результаты анализа и, соответственно, на репрезентативность и чистоту выборки.

В сложившейся ситуации мы считаем целесообразным отобрать в корпус главным образом такие газетные публикации, которые тяготеют к свободному стилю изложения, преимущественно в интересах целостности публицистического стиля (который в противном случае потребовалось бы разделять на два противоположных типа и тем самым создавать неравновесность свободной и формальной групп). Поскольку в противоположной группе уже наличествуют два стиля, являющиеся источниками проникающих в публицистический стиль специфических языковых средств, то необходимый объем данных, позволяющий выявлять элементы формальных стилей и оценивать степень их представленности в том или ином тексте, будет нами собран в любом случае.

Заметим, что поскольку ключевыми понятиями для МФСП-анализатора являются чисто математические показатели средней длины слова и потенциальной полисемантичности, которые, как было разъяснено выше, позволяют с достаточной достоверностью оценивать степень свободы или формализации функционального стиля, то применение аналитического алгоритма предоставляет выход на определение сущности предъявленного анализатору текста. Существует презумпция, согласно которой факт размещения того или иного текста в публицистическом издании создает основания для отнесения данного материала к публицистическому же стилю; SRS не обладает принципиальной возможностью учета описанной презумпции и в определенной степени может противодействовать ей, свидетельствуя, что научный или официально-деловой текст, будучи помещенным в газету, журнал или иное подобное издание, не утрачивает своих отличительных особенностей и не преобразуется в публицистический текст, оставаясь высоко формализованным и характеризуясь низкими показателями потенциальной полисемантичности. То же справедливо и в отношении авторского стиля пишущего: нет сомнений, что опубликованный в периодической печати, к примеру, экспертный комментарий ученого или специалиста практически неизбежно будет нести на себе своеобразный отпечаток формализации.

Считаем необходимым также отметить, что в той части выборки, которая представляет художественный стиль, в целях разрешения потенциальных проблемных вопросов будет соблюдено приближенно равное соотношение текстов, созданных непосредственно носителями русского языка, и переводов на указанный язык различных произведений иностранных авторов. Потребность в установлении подобного соотношения обусловлена фактом использования в качестве одного из эталонных образцов именно текста, который является продуктом перевода (рассказа О. Уайлда «Молодой король»): необходимо подтвердить, что данный факт не оказывает существенного влияния на диагностические критерии.

Заметим, что мы не усматриваем принципиальной разницы между изначально русскими текстами и результатами переводческой деятельности, поскольку в обоих случаях мы имеем дело с фактами русского языка; несмотря на то, что содержание переводимого текста диктуется содержанием оригинала, сам по себе порождаемый переводчиком текст не приобретает специфических черт языка оригинала. Тем не менее, могут существовать также и обратные воззрения, которые мы полагаем уместным учитывать.

В том, что касается элементов остальных стилей, целенаправленное четкое уравновешивание количества оригинальных и переведенных текстов производиться не будет. Тем не менее, в рамках соблюдения принципа многообразия исследуемых материалов тексты, являющиеся продуктами перевода, могут быть введены в состав корпуса изучаемых образцов.

На основании вышеописанных принципов нами была сформирована совокупность текстов для статистического исследования. При составлении выборки использовались электронные копии текстов, доступные для свободной загрузки из сети Интернет. Предварительная внешняя обработка материала была минимизирована и сводилась преимущественно к преобразованию данных в формат, воспринимаемый инструментом пакетной обработки (т.е. к сохранению информации в виде простых текстовых файлов в кодировке ANSI). Анализ дал результаты, представленные в таблицах 8-11.

Таблица 8. Итоги исследования текстов художественного стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Длина	359964	400357	236025	447871	269075	25415	81711
Слов	53913	63380	35110	65386	34534	2931	9809
СДС	4,7831	4,8953	4,9669	5,0885	3,9340	3,9710	3,8842
%СП	8,5972%	8,9744%	9,7095%	9,3124%	7,9226%	6,5507%	6,2901%
%СиП	11,0790%	9,4320%	10,8003%	8,9346%	6,0694%	5,4930%	5,9027%
Р-ть	-2,4818	-0,4576	-1,0908	0,3778	1,8532	1,0577	0,3874
О-е	1,2887	1,0510	1,1123	0,9594	0,7661	0,8385	0,9384
УА	-	-	-	-	-	-	-
АНСГ	-	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	8	9	10	11	12	13	14
Длина	22095	376723	90674	183978	165793	23287	114118
Слов	3554	39600	14394	27688	31699	3587	16839
СДС	4,6629	3,4766	4,7587	4,5648	3,8478	4,4957	4,8982
%СП	9,0321%	6,4747%	9,8166%	8,7691%	8,8646%	7,8617%	9,3711%
%СиП	7,5689%	5,6818%	9,5456%	10,4919%	10,5934%	8,2520%	7,1263%
Р-ть	1,4632	0,7929	0,2710	-1,7228	-1,7288	-0,3903	2,2448
О-е	0,8380	0,8775	0,9724	1,1965	1,1950	1,0496	0,7605
УА	-	-	-	-	-	-	-
АНСГ	-	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	15	16	17	18	19	20	21
Длина	313322	26592	75628	15183	17401	29917	20800
Слов	46329	3808	10983	1784	2859	4693	3171
СДС	4,8489	5,0089	4,9390	5,0639	4,4008	4,0324	4,8802
%СП	10,5701%	10,2416%	9,0595%	10,8744%	6,4708%	7,7136%	8,4516%
%СиП	7,5935%	10,4254%	7,2202%	7,6794%	4,4071%	6,9252%	7,3478%
Р-ть	2,9766	-0,1838	1,8393	3,1950	2,0637	0,7884	1,1038
О-е	0,7184	1,0179	0,7970	0,7062	0,6811	0,8978	0,8694
УА	-	-	-	+	+	-	-
АНСГ	-	-	-	+	+	-	-
Вердикт	+	+	+	+	+	+	+
Текст	22	23	24	25	Средний показатель		
Длина	118760	107931	72194	101108	147836,9		

Продолжение табл. 8

Слов	17920	16403	11278	16748	21536
СДС	4,2513	4,5122	4,4370	4,1192	4,5089
%СП	11,5848%	9,2178%	7,9003%	9,8460%	8,7791%
%СиП	9,5703%	9,1203%	8,5476%	10,6162%	8,2570%
Р-ть	2,0145	0,0975	-0,6473	-0,7702	0,5221
О-е	0,8261	0,9894	1,0819	1,0782	0,9403
УА	-	-	-	-	
АНСГ	-	-	-	-	
Вердикт	+	+	+	+	

Итого вердиктов: 25 верных, 0 нечетких, 0 неверных
Уточняющий анализ запускался: 2 раза
Анализ по надежным словарным группам запускался: 2 раза

Таблица 9. Итоги исследования текстов публицистического стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Длина	7230	8655	15477	29767	11681	5459	6486
Слов	964	1268	2251	4856	1632	752	848
СДС	4,9481	5,4495	5,4860	4,7780	5,8658	5,9535	5,8255
%СП	10,6846%	9,0694%	8,9738%	9,3699%	10,4779%	12,5000%	10,8491%
%СиП	8,2988%	6,7823%	6,5304%	9,2463%	7,1078%	7,4468%	10,2594%
Р-ть	2,3858	2,2871	2,4434	0,1236	3,3701	5,0532	0,5897
О-е	0,7767	0,7478	0,7277	0,9868	0,6784	0,5957	0,9457
УА	-	-	+	-	+	+	-
АНСГ	-	-	-	-	+	-	-
Вердикт	+	+	+	+	=	+	+
Текст	8	9	10	11	12	13	14
Длина	6623	15360	13956	19398	6621	10501	12455
Слов	1055	2149	1959	2908	819	1513	1822
СДС	4,8275	5,8381	5,7626	5,3855	5,4481	5,5056	5,4748
%СП	9,6682%	11,1680%	9,2905%	9,9037%	9,6459%	8,6583%	11,5258%
%СиП	8,9100%	7,3523%	7,4017%	6,4993%	4,7619%	6,8738%	7,4643%
Р-ть	0,7582	3,8157	1,8888	3,4044	4,8840	1,7845	4,0615
О-е	0,9216	0,6583	0,7967	0,6562	0,4937	0,7939	0,6476

Продолжение табл. 9

УА	-	+	-	+	+	-	+
АНСГ	-	+	-	-	+	-	+
Вердикт	**+**	**+**	**+**	**+**	**=**	**+**	**+**
Текст	15	16	17	18	19	20	21
Длина	11695	15516	6054	5257	5468	6675	6685
Слов	1826	2256	881	753	843	952	936
СДС	4,9578	5,4756	5,5392	5,6521	5,1851	5,1134	5,2799
%СП	8,8171%	10,0621%	9,9886%	9,4290%	8,8968%	9,8739%	9,6154%
%СиП	7,0646%	7,5355%	9,1941%	7,5697%	9,1340%	5,1471%	7,0513%
Р-ть	1,7525	2,5266	0,7945	1,8593	-0,2372	4,7268	2,5641
О-е	0,8012	0,7489	0,9205	0,8028	1,0267	0,5213	0,7333
УА	-	-	-	-	-	+	-
АНСГ	-	-	-	-	-	+	-
Вердикт	**+**	**+**	**+**	**+**	**+**	**=**	**+**
Текст	22	23	24	25	Средний показатель		
Длина	8503	7547	11982	8662	10548,52		
Слов	1205	1120	1877	1225	1546,8		
СДС	5,7734	5,3848	4,9989	5,7241	5,4253		
%СП	8,8797%	8,1250%	8,7906%	7,4286%	9,6677%		
%СиП	8,2158%	4,6429%	8,8439%	5,7959%	7,4052%		
Р-ть	0,6639	3,4821	-0,0533	1,6327	2,2625		
О-е	0,9252	0,5714	1,0061	0,7802	0,7706		
УА	-	+	-	-			
АНСГ	-	-	-	-			
Вердикт	**+**	**+**	**+**	**+**			

Итого вердиктов: 22 верных, 3 нечетких, 0 неверных
Уточняющий анализ запускался: 9 раз
Анализ по надежным словарным группам запускался: 5 раз

Таблица 10. Итоги исследования текстов научного стиля

Текст	1	2	3	4	5	6	7
Длина	8082	11794	7151	6307	5871	153660	522418
Слов	1016	1320	773	746	649	18607	62956
СДС	6,6181	7,4258	7,9405	7,1032	7,5932	6,5052	6,9933
%СП	11,0236%	10,8333%	12,0310%	12,7346%	12,6348%	9,3030%	13,5603%
%СиП	5,5118%	8,1818%	5,1746%	6,3003%	8,3205%	4,6703%	5,3878%
Р-ть	5,5118	2,6515	6,8564	6,4343	4,3143	4,6327	8,1724
О-е	0,5000	0,7552	0,4301	0,4947	0,6585	0,5020	0,3973
УА	-	+	-	-	-	-	-
АНСГ	-	+	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	8	9	10	11	12	13	14
Длина	640763	611274	294539	251950	8339	158064	342976
Слов	86264	76340	32636	29555	988	18346	38439
СДС	5,8213	6,4292	7,0700	6,9543	6,7247	7,2798	7,4348
%СП	9,9427%	10,6916%	10,3076%	9,5855%	10,3239%	13,7905%	12,9348%
%СиП	5,2571%	7,0789%	3,7566%	6,5133%	7,4899%	5,0583%	4,9351%
Р-ть	4,6856	3,6127	6,5510	3,0722	2,8340	8,7322	7,9997
О-е	0,5287	0,6621	0,3644	0,6795	0,7255	0,3668	0,3815
УА	+	-	-	-	+	-	-
АНСГ	+	-	-	-	+	-	-
Вердикт	=	+	+	+	+	+	+
Текст	15	16	17	18	19	20	21
Длина	576735	538073	638815	434023	10314	10577	9600
Слов	76182	67485	79477	49650	1183	1247	1130
СДС	6,1412	6,5304	6,5416	7,2522	7,4142	7,1460	6,9956
%СП	9,8068%	10,6024%	10,1828%	12,8540%	11,2426%	8,6608%	10,8850%
%СиП	7,3206%	5,5894%	5,9351%	5,3575%	6,5934%	8,8212%	6,4602%
Р-ть	2,4862	5,0130	4,2477	7,4965	4,6492	-0,1604	4,4248
О-е	0,7465	0,5272	0,5828	0,4168	0,5865	1,0185	0,5935
УА	+	-	-	-	-	+	-
АНСГ	+	-	-	-	-	+	-

Продолжение табл. 10

Вердикт	=	+	+	+	+	+	+
Текст	22	23	24	25	Средний показатель		
Длина	9142	12669	6324	5623	211003,3		
Слов	1143	1444	667	654	25955,88		
СДС	6,4182	7,4017	8,0435	7,2446	7,0009		
%СП	7,3491%	9,8338%	8,3958%	8,8685%	10,7352%		
%СиП	8,2240%	6,7175%	9,2954%	8,5627%	6,5005%		
Р-ть	-0,8749	3,1163	-0,8996	0,3058	4,2346		
О-е	1,1190	0,6831	1,1071	0,9655	0,6317		
УА	+	-	+	+			
АНСГ	+	-	+	+			
Вердикт	+	+	+	+			

Итого вердиктов: 23 верных, 2 нечетких, 0 неверных
Уточняющий анализ запускался: 8 раз
Анализ по надежным словарным группам запускался: 8 раз

Таблица 11. Итоги исследования текстов официально-делового стиля

Текст	1	2	3	4	5	6	7
Длина	174699	52122	9220	27346	131609	41427	36367
Слов	20664	6716	1066	3410	15172	4872	4659
СДС	6,9340	6,4242	6,8405	6,2651	6,9853	6,4002	6,3634
%СП	12,3355%	13,3859%	10,5066%	9,5601%	10,9478%	12,1716%	9,6802%
%СиП	4,4764%	6,7004%	4,0338%	6,4223%	6,3011%	6,7323%	7,0831%
Р-ть	7,8591	6,6855	6,4728	3,1378	4,6467	5,4393	2,5971
О-е	0,3629	0,5006	0,3839	0,6718	0,5756	0,5531	0,7317
УА	-	-	-	-	-	-	+
АНСГ	-	-	-	-	-	-	+
Вердикт	+	+	+	+	+	+	+
Текст	8	9	10	11	12	13	14
Длина	67000	11092	22194	13540	54614	26047	308198
Слов	5866	1242	2543	1505	6304	2241	26883
СДС	5,8529	7,2262	7,0822	7,5056	6,3165	6,9942	5,0633
%СП	6,5973%	11,5137%	11,2466%	15,5482%	9,4385%	8,2106%	8,3808%

%СиП	3,0685%	4,5089%	6,0558%	6,7110%	4,9810%	6,3811%	2,6113%
Р-ть	3,5288	7,0048	5,1908	8,8372	4,4575	1,8295	5,7695
О-е	0,4651	0,3916	0,5385	0,4316	0,5277	0,7772	0,3116
УА	+	-	-	-	-	+	+
АНСГ	+	-	-	-	-	+	+
Вердикт	**=**	**+**	**+**	**+**	**+**	**+**	**=**
Текст	15	16	17	18	19	20	21
Длина	35743	463287	105466	102350	26770	68811	47690
Слов	3890	39250	11425	11922	3013	7466	5739
СДС	7,6270	6,2498	7,6229	7,0478	7,6777	6,8924	7,0585
%СП	9,6144%	10,7414%	17,4354%	9,8893%	15,3004%	8,2641%	16,2223%
%СиП	5,0386%	5,1261%	4,1138%	3,7913%	11,5831%	3,9379%	6,3774%
Р-ть	4,5758	5,6153	13,3216	6,0980	3,7173	4,3262	9,8449
О-е	0,5241	0,4772	0,2359	0,3834	0,7570	0,4765	0,3931
УА	-	-	-	-	+	-	-
АНСГ	-	-	-	-	-	-	-
Вердикт	**+**	**+**	**+**	**+**	**+**	**+**	**+**
Текст	22	23	24	25	Средний показатель		
Длина	69531	140400	80977	47342	86553,68		
Слов	8277	17074	9739	5411	9053,96		
СДС	6,9447	6,7259	6,8607	7,2689	6,8092		
%СП	9,8586%	13,5996%	11,8493%	6,3574%	11,1462%		
%СиП	4,0715%	6,4894%	4,0867%	9,2220%	5,5962%		
Р-ть	5,7871	7,1102	7,7626	-2,8646	5,5500		
О-е	0,4130	0,4772	0,3449	1,4506	0,5262		
УА	-	-	-	+			
АНСГ	-	-	-	+			
Вердикт	**+**	**+**	**+**	**+**			

Итого вердиктов: 23 верных, 2 нечетких, 0 неверных
Уточняющий анализ запускался: 6 раз
Анализ по надежным словарным группам запускался: 5 раз

В выше приведенных таблицах приняты следующие сокращения: СДС – показатель средней длины слова, %СП – массовая доля слабо полисемантизованных словарных групп, %СиП – массовая доля сильно полисемантизованных словарных групп, Р-ть – разность массовых долей слабо и сильно по-

лисемантизованных словарных групп, О-е – отношение массовых долей сильно и слабо полисемантизованных словарных групп, УА – факт запуска уточняющего анализа, АНСГ – факт запуска анализа по надежным словарным группам. В строках УА и АНСГ знак «-» означает, что соответствующий анализ не запускался, знак «+» имеет противоположное значение. В строке «Вердикт» знак «+» означает верное определение типа функционального стиля, знак «=» – нечеткое решение (означающее, что в тексте имеется противоречие базовых критериев, которое не удалось устранить при помощи дополнительных аналитических процедур); знак «-», означающий неверный вердикт, в таблицах не встречается.

На основании данных, систематизированных в статистических таблицах, может быть сделан ряд базовых выводов о степени эффективности текущей реализации разработанного программного обеспечения и о тенденциях, обнаруживаемых в результате анализа выборки. Представим указанные выводы отдельно для каждого из функциональных стилей с их последующей интеграцией.

Итак, согласно итогам анализа художественных текстов все они (100%) были классифицированы корректно и отнесены к группе свободных стилей. Соответственно, нечеткие и ошибочные вердикты отсутствовали. В 2 случаях из 25 (8%) возникло первоначальное противоречие основных критериев, которое было успешно устранено посредством дополнительных аналитических процедур; заметим, что в обоих упомянутых случаях исследовался довольно короткий текст (15183 и 17401 знак соответственно). Сколь-либо значимых различий в статистических показателях между изначально русскими текстами и продуктами перевода не обнаружено; родовые и жанровые различия также не оказали существенного влияния на точность детектирования. В большинстве случаев не было отмечено и достоверной связи между статистическими показателями и объемом исследуемого образца.

Общая длина текстов варьировалась в диапазоне от 15183 до 447871 знака; среднее значение составило в итоге 147836,9 знака. Количество слов в изучаемых текстах располагалась в диапазоне от 1784 до 65386 штук; усредненный показатель установился на уровне 21536 слов. Минимум средней длины

слова насчитывает 3,4766 знака, максимум – 5,0885; среднее значение данного параметра равно 4,5089 знака.

Массовая доля слабо полисемантизованных словарных групп изменялась в пределах от 6,2901% до 11,5848%, сильно полисемантизованных – от 4,4071% до 11,0790%. Диапазон значений разности составил от -2,4818 до 3,195 процентов; отношение варьируется в границах от 0,6811 до 1,2887. В среднем вышеупомянутые показатели укладываются в изначально заданные диагностические критерии: %СП - 8,7791%, %СиП - 8,2570%, разность - 0,5221%, отношение - 0,9403. Напомним, что согласно имеющимся у нас на текущем этапе данным, по мере увеличения степени свободы стиля разность стремится ко множеству отрицательных чисел, а отношение стремится превысить единицу; это означает, что массовая доля сильно полисемантизованных словарных групп стремится превзойти массовую долю слабо полисемантизованных словарных групп. Данные по текстам художественного стиля в целом подтверждают эту тенденцию.

Результаты изучения публицистических текстов позволили верно классифицировать 22 из 25 текстов (88%). В трех оставшихся случаях (12%) возникло противоречие базовых критериев, которое не удалось снять посредством процедур вспомогательного исследования. В 9 случаях (36%) запускался уточняющий анализ; в 5 из них (20%) возникла также и потребность в использовании анализа по надежным словарным группам. Необходимо отметить, что в сравнении с прочими функциональными стилями данный фрагмент выборки характеризуется абсолютным доминированием коротких текстов: минимальная длина изучаемого образца в этом случае составила 5257 знаков (при нижней границе в 5000), а максимальная – 29767. Средняя длина текста насчитывает, таким образом, 10548,52 знака. По очевидным причинам, связанным с необходимостью накопления репрезентативного объема данных, точность определения коротких текстов может быть ниже (что, собственно, и демонстрируется фактом наличия трех нечетких вердиктов).

В изученных текстах насчитывалось от 752 до 4856 слов (средний показатель - 1546,8); параметр средней длины слова варьировался в пределах от 4,778 до 5,9535 со среднеариф-

метическим значением в 5,4253 знака. Массовая доля слабо полисемантизованных словарных групп установилась в диапазоне от 7,4286% до 12,5%, а сильно полисемантизованных – от 4,6429% до 10,2594%; средние показатели составили 9,6677% и 7,4052% соответственно, дав тем самым разность в 2,2625 процента (от -0,2372 до 5,0532) и отношение в 0,7706 (минимальное значение - 0,4937, максимальное - 1,0267). Это дает основания заключить, что в целом среднестатистические показатели текстов публицистического стиля также уложились в рамки диагностических критериев для свободных функциональных стилей.

Таким образом, из 50 образцов свободного типа МФСП-анализатор без каких-либо доработок и уточнений своего алгоритма успешно классифицировал 47 текстов (94%). В трех оставшихся случаях вердикт оказался нечетким; ошибочных заключений обнаружено не было. Усредненные диагностически значимые показатели обоих свободных стилей соответствуют заданным критериям: средняя длина слова равна 4,9671, массовая доля слабо полисемантизованных словарных групп составляет 9,2234%, сильно полисемантизованных - 7,8311%, что дает разность 1,3923% и отношение 0,8554. Полученные результаты имеют основания считаться успешными и позволяют нам непосредственно перейти к рассмотрению итогов исследования образцов формализованных функциональных стилей.

После обработки корпуса научных текстов были корректно детектированы 23 текста из 25 (92%); в двух других случаях вердикт оказался нечетким. В 8 случаях (32%) потребовались запуск уточняющего исследования и анализа по надежным словарным группам. В выборке были в приблизительно равной пропорции представлены короткие тексты (статьи из сборника материалов широкопрофильной научной конференции) и длинные образцы, являющие собой объемные научные исследования и монографии из различных областей знания. Соответственно, минимум длины текста в данном случае составил 5623 знака, а максимум – 640763; усредненное значение равно 211003,3 знака.

В изученных текстах содержалось от 649 до 86264 слов (в среднем - 25955,88), параметр СДС для которых изменялся в диапазоне от 5,8213 до 8,0435 знака – среднее арифметическое

составило 7,0009. Массовая доля слабо полисемантизованных словарных групп характеризовалась минимальным значением в 7,3491% и максимальным – в 13,7905%; аналогичные границы для сильно полисемантизованных словарных групп составили 3,7566% и 9,2954% соответственно. Диапазон разности обладает границами в -0,8996 и 8,7322 процента, а отношение варьируется от 0,3644 до 1,119. Что касается средних значений, то они соответствуют изначально заданным диагностическим критериям: %СП - 10,7352%, %СиП - 6,5005%, разность - 4,2346%, отношение - 0,6317. Напомним, что для формальных функциональных стилей разность стремится к возрастанию в область положительных чисел, а отношение бесконечно стремится к нулю; это отражает тенденцию к преобладанию слабо полисемантизованных словарных групп над сильно полисемантизованными.

Из 25 текстов официально-делового стиля были верно классифицированы также 23 (92%), а в остальных двух случаях вердикт оказался нечетким по вине сниженного показателя средней длины слова, вышедшего за пределы диагностического диапазона. Уточняющий анализ запускался 6 раз (24% случаев), в пяти из которых (20%) возникла также и потребность в исследовании по надежным словарным группам.

Длина обработанных текстов находилась в диапазоне от 9220 до 463287 знаков, так что средний показатель составил 86553,68 знака. Количество слов варьировалось от 1066 до 39250 со среднеарифметическим значением в 9053,96 штуки. Что касается параметра средней длины слова, то его нижняя и верхняя границы установились на уровне 5,0633 и 7,6777 знаков соответственно; усредненный показатель равен 6,8092. Массовая доля слабо полисемантизованных словарных групп располагалась в диапазоне от 6,3574% до 17,4354%, а сильно полисемантизованных – от 2,6113% до 11,5831%; соответственно, минимальная разность составила -2,8646, а максимальная – 13,3216 процентов; отношение же уложилось в пределы от 0,2359 до 1,4506. Таким образом, средние показатели для официально-делового стиля также оказались соответствующими заданным диагностическим критериям: %СП – 11,1462%, %СиП – 5,5962%, разность – 5,55%, отношение – 0,5262.

Таким образом, из 50 текстов, принадлежащих к формаль-

ным функциональным стилям, МФСП-анализатор успешно классифицировал 46 (92%). В четырех оставшихся случаях вердикт оказался нечетким; ошибочных решений зафиксировано не было. Усредненные значения основных диагностических критериев соответствуют изначально заданным диапазонам: средняя длина слова составляет 6,9051 знака, массовая доля слабо полисемантизованных словарных групп равна 10,9407%, сильно полисемантизованные словарные группы насчитывают 6,0484%, разность между ними установилась на уровне 4,8923%, а отношение – на отметке в 0,579.

Сведем полученные среднеарифметические показатели в таблицу 12.

Таблица 12. Средние значения основных диагностических параметров

Параметр	ХС	ПС	НС	ОС	СС	ФС
СДС	4,5089	5,4253	7,0009	6,8092	4,967088	6,9051
%СП	8,7791%	9,6677%	10,7352%	11,1462%	9,2234%	10,9407%
%СиП	8,2570%	7,4052%	6,5005%	5,5962%	7,8311%	6,0484%
Р-ть	0,5221	2,2625	4,2346	5,5500	1,392308	4,892324
О-е	0,9403	0,7706	0,6317	0,5262	0,855434	0,578972

Примечание. ХС – художественный стиль, ПС – публицистический стиль, НС – научный стиль, ОС – официально-деловой стиль, СС – свободные стили (суммарно), ФС – формальные стили (суммарно)

Данные в таблице наглядно свидетельствуют о том, что для свободных стилей действительно характерны относительно низкие показатели средней длины слова, массовой доли слабо полисемантизованных словарных групп, разности %СП-%СиП и отношения %СиП/%СП, а также относительно высокие показатели массовой доли сильно полисемантизованных словарных групп, в то время как в случае формальных функциональных стилей ситуация является обратной.

Таким образом, проведя исследования на широкой выборке текстов, мы смогли установить, что в своей текущей реализации без каких-либо модификаций, но исключительно на основании диагностических критериев, полученных по итогам

предварительных исследований на четырех эталонных текстах, МФСП-анализатор успешно классифицировал в общей сложности 93 из 100 предъявленных ему текстов (что со всей очевидностью равно 93%), а в оставшихся 7 случаях завершил анализ с нечетким вердиктом, не допустив ни одного ошибочного детектирования. При этом в 18 случаях принципиальное несоответствие базовых диагностических критериев удалось успешно ликвидировать посредством вспомогательных аналитических процедур, что говорит об их эффективности. Кроме того, итоги анализа выборки позволили подтвердить факт присутствия принципиальных различий в средней длине слова и в массовых долях слабо / сильно полисемантизованных словарных групп в текстах свободных и формальных функциональных стилей.

Тем не менее, нам необходимо также изучить среднестатистические количественные показатели словарных групп, на нахождении массовых долей которых строится процедура определения потенциальной полисемантичности текста, чтобы определить, подтверждают или опровергают они выбор конкретных словарных групп в качестве показательных. Данная часть исследования представляется нам целесообразной, поскольку индивидуальные характеристики эталонных текстов могли оказать влияние не только на показатели средней длины слова, но и на преобладание в них лексических единиц, ассоциированных с теми или иными группами.

Представим среднеарифметические показатели долей основных показательных словарных групп в текстах различных функциональных стилей в статистической таблице.

Таблица 13. Средние массовые доли словарных групп в текстах различных функциональных стилей

СГ	ХС	ПС	НС	ОС	СС	ФС
А	0,3818	0,7792	1,2733	1,5392	0,5805	1,4062
Б	3,2023	2,8651	1,9355	1,3503	3,0337	1,6429
Д	3,4709	3,2472	3,1434	3,9605	3,3591	3,5519
И	1,4698	2,2922	3,3330	3,6504	1,8810	3,4917
Й	0,0147	0,0000	0,0220	0,0000	0,0073	0,0110
Э	0,1964	0,4390	0,9470	0,4518	0,3177	0,6994
Ю	0,0433	0,0474	0,0811	0,1941	0,0453	0,1376

Продолжение табл. 13

СГ	ХС	ПС	НС	ОС	СС	ФС
Е	1,5353	1,4675	0,8500	0,6964	1,5014	0,7732
Ж	0,8424	0,6390	0,2657	0,1723	0,7407	0,2190
У	1,7155	1,8030	2,6991	3,0701	1,7593	2,8846
Ц	0,2339	0,3430	0,4894	0,5139	0,2885	0,5016
Ч	2,4071	2,4605	1,6164	0,8647	2,4338	1,2406
Щ	0,0607	0,0237	0,0209	0,0018	0,0422	0,0113
Я	1,4625	0,6685	0,5591	0,2802	1,0655	0,4197

Согласно данным таблицы, состав словарных групп, используемых для расчетов потенциальной полисемантичности, нуждается в корректировке. Очевидно, что четыре словарные группы из исходных 14 показывают тенденцию, обратную предполагаемой, либо не демонстрируют устойчивости и четкой различимости массовых долей; в случае слабо полисемантизованных словарных групп это Б и Д, а в случае сильно полисемантизованных – У и Ц. В силу этого мы считаем необходимым исключить данные словарные группы из основного алгоритма расчетов, провести анализ повторно и установить, претерпят ли результаты исследования выборки существенные изменения. В данном случае нас будут интересовать исключительно количественные показатели, поэтому вторичная проверка четкости вердиктов (включая проставление отметок о фактах запуска уточняющего анализа и анализа по надежным словарным группам) производиться на этом этапе не будет; также мы не станем упоминать параметры общей длины текста, количества слов в нем и средней длины слова, поскольку вносимые в алгоритм модификации никак не касаются соответствующих шагов анализа.

По итогам расчетов с удаленными из основной базы данных недостаточно достоверными словарными группами нами получены новые результаты, которые мы также систематизируем и изложим в таблицах.

Таблица 14. Результаты повторного анализа для художественного стиля

Текст	1	2	3	4	5	6	7
%СП	2,3186%	2,3825%	2,3697%	2,5021%	2,0212%	1,8765%	1,5700%
%СиП	8,6565%	7,3872%	8,5246%	6,4387%	4,3609%	4,1624%	4,1493%

Разность	-6,3379	-5,0047	-6,1549	-3,9366	-2,3397	-2,2859	-2,5793
Отношение	3,7336	3,1007	3,5974	2,5733	2,1576	2,2182	2,6429
Текст	**8**	**9**	**10**	**11**	**12**	**13**	**14**
%СП	1,8008%	1,3586%	2,1120%	1,9864%	2,8708%	2,0072%	1,7459%
%СиП	5,9932%	4,4141%	7,6699%	8,6752%	8,3441%	6,2448%	5,0716%
Разность	-4,1924	-3,0555	-5,5579	-6,6888	-5,4733	-4,2376	-3,3257
Отношение	3,3281	3,2491	3,6316	4,3673	2,9066	3,1111	2,9048
Текст	**15**	**16**	**17**	**18**	**19**	**20**	**21**
%СП	1,9642%	2,4422%	1,8028%	3,0830%	1,2592%	1,9604%	2,4913%
%СиП	5,5386%	8,1933%	5,1261%	5,7175%	2,9731%	4,9435%	5,3611%
Разность	-3,5744	-5,7511	-3,3233	-2,6345	-1,7139	-2,9831	-2,8698
Отношение	2,8198	3,3548	2,8434	1,8545	2,3611	2,5217	2,1519
Текст	**22**	**23**	**24**	**25**	**Средний показатель**		
%СП	3,4208%	2,0484%	1,6049%	1,6480%	2,1059%		
%СиП	7,6228%	7,0841%	6,7122%	8,3234%	6,3075%		
Разность	-4,2020	-5,0357	-5,1073	-6,6754	-4,2016		
Отношение	2,2284	3,4583	4,1823	5,0507	3,0540		

Таблица 15. Результаты повторного анализа для публицистического стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**
%СП	2,0747%	4,6530%	3,7317%	2,4918%	3,9828%	5,3191%	4,9528%
%СиП	6,1203%	4,6530%	4,4425%	6,8987%	4,8407%	4,6543%	7,9009%
Разность	-4,0456	0,0000	-0,7108	-4,4069	-0,8579	0,6648	-2,9481
Отношение	2,9500	1,0000	1,1905	2,7686	1,2154	0,8750	1,5952
Текст	**8**	**9**	**10**	**11**	**12**	**13**	**14**
%СП	3,2227%	4,4672%	3,8285%	3,2325%	4,2735%	2,6438%	4,1164%
%СиП	6,8246%	5,4444%	5,3088%	4,5392%	2,6862%	4,3622%	5,8178%
Разность	-3,6019	-0,9772	-1,4803	-1,3067	1,5873	-1,7184	-1,7014
Отношение	2,1176	1,2188	1,3867	1,4043	0,6286	1,6500	1,4133
Текст	**15**	**16**	**17**	**18**	**19**	**20**	**21**
%СП	3,0120%	3,1472%	2,8377%	3,4529%	3,2028%	2,4160%	4,0598%
%СиП	5,3669%	5,4965%	7,0375%	4,6481%	7,2361%	4,0966%	4,5940%
Разность	-2,3549	-2,3493	-4,1998	-1,1952	-4,0333	-1,6806	-0,5342

Продолжение табл. 15

Отношение	1,7818	1,7465	2,4800	1,3462	2,2593	1,6957	1,1316

Текст	**22**	**23**	**24**	**25**	**Средний показатель**
%СП	3,6515%	3,0357%	3,7294%	3,3469%	3,5553%
%СиП	4,3154%	3,4821%	6,7128%	4,0000%	5,2592%
Разность	-0,6639	-0,4464	-2,9834	-0,6531	-1,7039
Отношение	1,1818	1,1471	1,8000	1,1951	1,5672

Таблица 16. Результаты повторного анализа для научного стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**
%СП	6,3976%	5,0000%	5,1746%	7,2386%	6,0092%	5,7505%	7,8277%
%СиП	4,0354%	2,7273%	1,5524%	3,0831%	2,4653%	3,2300%	2,4477%
Разность	2,3622	2,2727	3,6222	4,1555	3,5439	2,5205	5,3800
Отношение	0,6308	0,5455	0,3000	0,4259	0,4103	0,5617	0,3127
Текст	**8**	**9**	**10**	**11**	**12**	**13**	**14**
%СП	5,7892%	6,6138%	6,1864%	4,4561%	5,1619%	6,2411%	8,1766%
%СиП	3,5461%	4,8913%	1,9672%	3,3869%	1,5182%	3,2868%	2,7004%
Разность	2,2431	1,7225	4,2192	1,0692	3,6437	2,9543	5,4762
Отношение	0,6125	0,7396	0,3180	0,7601	0,2941	0,5266	0,3303
Текст	**15**	**16**	**17**	**18**	**19**	**20**	**21**
%СП	5,0287%	5,3434%	5,3047%	8,0685%	6,5089%	5,1323%	6,0177%
%СиП	5,6352%	2,8110%	3,9986%	2,5277%	2,1978%	3,4483%	4,7788%
Разность	-0,6065	2,5324	1,3061	5,5408	4,3111	1,6840	1,2389
Отношение	1,1206	0,5261	0,7538	0,3133	0,3377	0,6719	0,7941
Текст	**22**	**23**	**24**	**25**	**Средний показатель**		
%СП	3,3246%	5,0554%	2,9985%	2,5994%	5,6562%		
%СиП	5,3368%	3,1856%	2,9985%	5,0459%	3,3121%		
Разность	-2,0122	1,8698	0,0000	-2,4465	2,3441		
Отношение	1,6053	0,6301	1,0000	1,9412	0,6585		

Таблица 17. Результаты повторного анализа для официально-делового стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**
%СП	7,7478%	7,8023%	4,5028%	4,2522%	6,7888%	3,2430%	4,9152%
%СиП	1,5970%	3,2162%	1,2195%	3,6070%	1,9641%	2,9762%	3,5415%
Разность	6,1508	4,5861	3,2833	0,6452	4,8247	0,2668	1,3737
Отношение	**0,2061**	**0,4122**	**0,2708**	**0,8483**	**0,2893**	**0,9177**	**0,7205**

Продолжение табл. 18

Текст	8	9	10	11	12	13	14
%СП	2,7787%	3,6232%	6,3704%	7,5083%	4,6003%	4,9041%	5,0329%
%СиП	0,5967%	1,0467%	2,1628%	1,3289%	1,6022%	0,4904%	0,8109%
Разность	2,1820	2,5765	4,2076	6,1794	2,9981	4,4137	4,2220
Отношение	**0,2147**	**0,2889**	**0,3395**	**0,1770**	**0,3483**	**0,1000**	**0,1611**
Текст	**15**	**16**	**17**	**18**	**19**	**20**	**21**
%СП	5,5270%	4,5172%	14,0919%	5,7876%	9,8241%	3,8843%	4,6001%
%СиП	1,6710%	1,8497%	1,7155%	1,8202%	0,9957%	1,3662%	2,9273%
Разность	3,8560	2,6675	12,3764	3,9674	8,8284	2,5181	1,6728
Отношение	**0,3023**	**0,4095**	**0,1217**	**0,3145**	**0,1014**	**0,3517**	**0,6364**
Текст	**22**	**23**	**24**	**25**	**Средний показатель**		
%СП	5,5817%	8,3870%	7,2081%	2,3840%	5,8345%		
%СиП	1,5344%	3,0749%	1,8072%	5,3964%	2,0127%		
Разность	4,0473	5,3121	5,4009	-3,0124	3,8218		
Отношение	**0,2749**	**0,3666**	**0,2507**	**2,2636**	**0,4275**		

Простейший визуальный анализ приведенных данных позволяет установить, что исключение слабо надежных словарных групп дало заметный положительный результат: между свободными и формализованными функциональными стилями окончательно сформировалось четкое различие в основных диагностических показателях. В случае художественного и публицистического стилей подавляющее большинство образцов (47 из 50) характеризуются преобладанием массовой доли сильно полисемантизованных словарных групп над массовой долей слабо полисемантизованных, что отражается в отрицательной разности и отношении, превышающем единицу. В случае научного и официально-делового стилей результат является прямо противоположным: 45 из 50 текстов характеризуются доминированием слабо полисемантизованных словарных групп, что приводит к образованию положительной разности и отношения, меньшего единицы. Аналогичные тенденции демонстрируют и среднестатистические показатели.

С учетом новых сведений мы полагаем необходимым пересмотреть процесс вынесения вердикта по второму аналитическому критерию. Поскольку обнаруженная тенденция уравнивает диагностическую ценность разности и отноше-

ния, а абсолютные значения массовых долей претерпевают существенные изменения в зависимости от конкретного текста, нам представляется уместным сократить количество подкритериев до одного (разности массовых долей слабо и сильно полисемантизованных словарных групп), установив в качестве порогового значения нуль. Соответственно, при анализе по второму критерию свободными будут считаться тексты, у которых разность меньше нуля, а формальными – тексты, у которых аналогичный показатель больше нуля.

Приняв описанные выше изменения за основу, сравним итоги обновленного анализа по второму критерию с результатами вычислений средней длины слова, приведенными ранее в данном параграфе, и представим сопоставляемые параметры в таблицах. При этом считаем целесообразным не указывать собственно значения повторно, ограничившись отметками о субвердиктах по каждому из критериев и об итогах их сравнения для каждого из вошедших в выборку текстов.

Таблица 18. Сравнение субвердиктов для художественного стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
К-1	+	+	+	+	+	+	+	+	+	+	+	+
К-2	+	+	+	+	+	+	+	+	+	+	+	+
Итог	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**
13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+
+	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**

Таблица 19. Сравнение субвердиктов для публицистического стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
К-1	+	+	+	+	+	+	+	+	+	+	+	+
К-2	+	?	+	+	+	-	+	+	+	+	+	-
Итог	**+**	**?**	**+**	**+**	**+**	**=**	**+**	**+**	**+**	**+**	**+**	**=**
13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+
+	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**

Таблица 20. Сравнение субвердиктов для научного стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
К-1	+	+	+	+	+	+	+	-	+	+	+	+
К-2	+	+	+	+	+	+	+	+	+	+	+	+
Итог	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**=**	**+**	**+**	**+**	**+**
13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	-	+	+	+	+	+	+	-	+	?	-
+	**+**	**=**	**+**	**+**	**+**	**+**	**+**	**+**	**=**	**+**	**?**	**=**

Таблица 21. Сравнение субвердиктов для официально-делового стиля

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
К-1	+	+	+	+	+	+	+	-	+	+	+	+
К-2	+	+	+	+	+	+	+	+	+	+	+	+
Итог	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**=**	**+**	**+**	**+**	**+**
13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
+	-	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	-
+	**=**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**+**	**=**

В таблицах 18-21 знаком «+» обозначается вердикт, соответствующий типу текста, знаком «-» - вердикт, не соответствующий типу текста, знаком «=» - противоречие субвердиктов, знаком «?» - неопределенный вердикт.

Суммируем данные из приведенных таблиц. В двух случаях точно определить вердикт невозможно по причине того, что разность равна нулю (т.е. массовая доля слабо полисемантизованных словарных групп равна массовой доле сильно полисемантизованных словарных групп). Нуль не может быть классифицирован в пользу какого-либо вердикта, поэтому итоговое решение считается сомнительным и обозначается в таблицах условным символом «?». Еще в девяти случаях наблюдается противоречие субвердиктов. Для остальных 89 текстов оба критерия свидетельствуют в пользу верного решения; ошибочных итоговых вердиктов нет. Полученные данные, помимо прочего, говорят об увеличении точности работы

анализатора, поскольку при исследовании по 14 словарным группам противоречие базовых критериев наблюдалось в 25 случаях.

Таким образом, на данный момент мы имеем 11 проблемных текстов: П-2, П-6, П-12, Н-8, Н-15, Н-22, Н-24, Н-25, ОД-8, ОД-14, ОД-25. Предпримем попытку изыскать сведения, позволяющие ликвидировать несоответствия и неопределенности. В первую очередь нам представляется уместным привлечь ранее полученные данные (т.к. анализ по 14 словарным группам, хотя и не характеризовался достаточной четкостью, все же продемонстрировал убедительный результат в 93% корректно классифицированных текстов) и произвести сопоставление итоговых вердиктов. Изобразим итоги сравнения в таблице.

Таблица 21. Сравнение вердиктов для анализа по 10 и 14 словарным группам

Текст	**П-2**	**П-6**	**П-12**	**Н-8**	**Н-15**
ИВ-14	+	+	=	=	=
ИВ-10	?	=	=	=	=
Н-22	**Н-24**	**Н-25**	**ОД-8**	**ОД-14**	**ОД-25**
=	+	+	=	=	+
=	?	=	=	=	=

Примечание. ИВ-14 – итоговый вердикт при анализе по 14 словарным группам, ИВ-10 – итоговый вердикт при анализе по 10 словарным группам

Из таблицы видно, что в 5 случаях из 11 исследование по 14 словарным группам позволяет ликвидировать неопределенность либо несоответствие. Заметим также, что активную роль в этом играют вспомогательные процедуры уточняющего анализа и исследования по надежным словарным группам. Из этого следует, что, во-первых, целесообразно сохранить анализ по 14 словарным группам в качестве дополнительного инструмента, который теоретически способен оказывать уточняющее воздействие на вердикт, а во-вторых, дополнительное

исследование по надежным словарным группам также не следует исключать из аналитического алгоритма, подвергнув его, однако, некоторым модификациям.

В первую очередь надлежит вывести из состава надежных словарную группу «Ц», поскольку ранее она была удалена из базового алгоритма как недостаточно достоверная. Это сокращает количество групп до четырех: в списке остаются «ЕЁ», «Ж», «Щ» и «Я». Указанные группы, согласно имеющимся в нашем распоряжении статистическим сведениям, действительно характеризуются относительным постоянством показателей, в то время как прочие группы могут довольно существенно варьировать свои статистические результаты. Мы считаем допустимым сохранить для них обозначение «надежные», однако в дальнейшем будем подразумевать именно относительную устойчивость демонстрируемых ими показателей.

По итогам сравнения имеющихся пороговых значений со среднестатистическими нами были определены новые границы числовых отрезков, факт принадлежности к которым оказывает непосредственное влияние на вердикт. Указанные значения приобрели следующий вид: ЕЁ – 1,45%; Ж – 0,38%; Щ – 0,03%; Я – 0,68%. Превышение данных порогов свидетельствует о наличии у текста признаков свободного стиля, в то время как более низкие показатели дают основания относить его к формальным ФС.

Соответственно, базовую последовательность действий, выполняемых в случае противоречия диагностических критериев, необходимо адаптировать к новым сведениям, после чего внести модификации в МФСП-анализатор.

Изобразим фрагмент алгоритма, составляющий упомянутую выше последовательность, в блок-схематическом представлении (рис. 12).

Аббревиатурой «УС» на изображении обозначен уточняющий анализ (исключение словарных групп по тематике), аббревиатурой «РС» - расширенный анализ (по 14 словарным группам); сокращение «ПС», как и в приведенных ранее блок-схемах, соответствует процедуре сравнения.

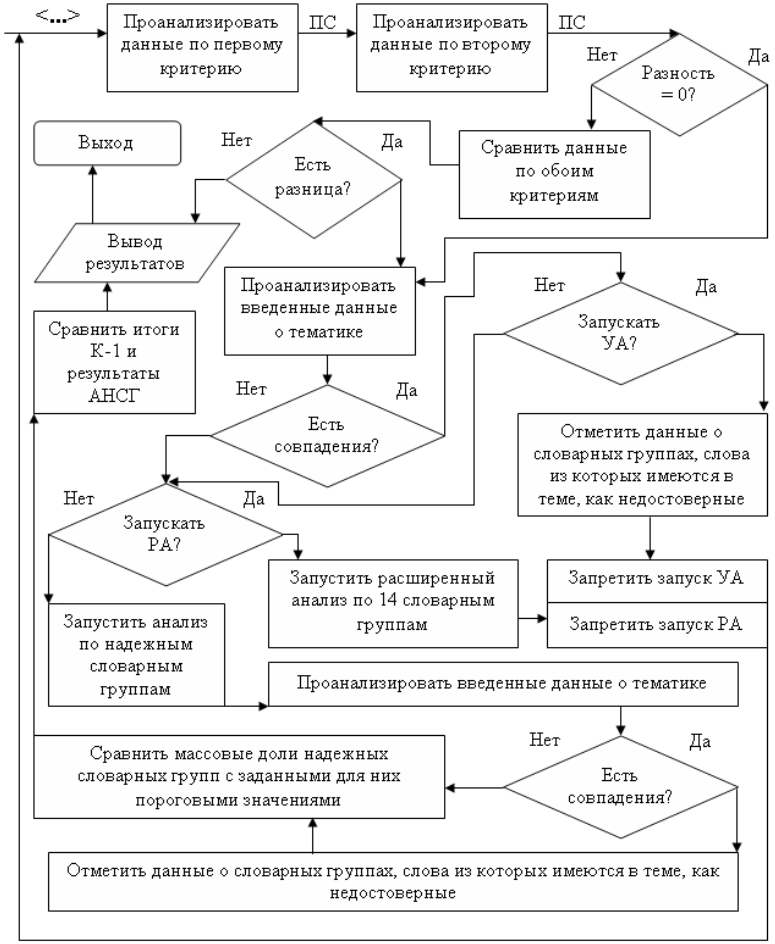

Рис. 12. Фрагмент аналитического алгоритма со внесенными изменениями

Необходимо заметить, что процедура анализа по второму критерию в данной реализации не является однотипной и варьируется в зависимости от конкретного шага, на котором находится исполнение алгоритма. Если при стандартной обработке текста, как уже было указано нами выше, единствен-

ным диагностическим критерием выступает разность массовых долей слабо и сильно полисемантизованных словарных групп, то при расширенном анализе процедура сохраняется в ее исходном виде – субвердикт выносится на основании трех подкритериев (%СиП, разность и отношение).

Процедура уточняющего анализа, таким образом, фактически может запускаться трижды. Учет ключевых слов будет осуществляться как при обычном, так и при расширенном исследовании, а также при анализе по надежным словарным группам. Мы рассчитываем, что подобная последовательная четырехуровневая защита от нечетких вердиктов в конечном счете позволит практически полностью их ликвидировать.

Следует, однако, напомнить, что дополнительные процедуры предназначены для снятия противоречий, которые обусловлены возникновением ошибочных субвердиктов по второму диагностическому критерию. В ряде случаев (на рассматриваемой выборке – в 3 из 100) могут возникать аномалии средней длины слова, вызванные специфическими особенностями конкретных текстов; для них вердикт неизбежно будет нечетким, поскольку ошибки по первому критерию ликвидировать невозможно.

Итак, мы внесли требуемые изменения в аналитический алгоритм и в разработанное программное обеспечение, а также в файлы его базы данных, после чего проанализировали имеющуюся выборку повторно. Это позволяет нам представить результаты исследования с указанием фактов запуска тех или иных проверок и итоговых вердиктов в аналитических таблицах 22-25. Заметим, что собственно значения критериев приводиться не будут, поскольку они уже содержатся в таблицах, приведенных выше.

Таблица 22. Результаты повторного анализа для художественного стиля

Текст	1	2	3	4	5	6	7
ЗДП	-	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	8	9	10	11	12	13	14
ЗДП	-	-	-	-	-	-	-

Продолжение табл. 22

Вердикт	+	+	+	+	+	+	+
Текст	15	16	17	18	19	20	21
ЗДП	-	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	22	23	24	25			
ЗДП	-	-	-	-			
Вердикт	+	+	+	+			

Таблица 23. Результаты повторного анализа для публицистического стиля

Текст	1	2	3	4	5	6	7
ЗДП	-	УА-1, РА	-	-	-	УА-1 - УА-2	-
Вердикт	+	+	+	+	+	+	+
Текст	8	9	10	11	12	13	14
ЗДП	-	-	-	-	УА-1	-	-
Вердикт	+	+	+	+	+	+	+
Текст	15	16	17	18	19	20	21
ЗДП	-	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	22	23	24	25			
ЗДП	-	-	-	-			
Вердикт	+	+	+	+			

Таблица 24. Результаты повторного анализа для научного стиля

Текст	1	2	3	4	5	6	7
ЗДП	-	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	8	9	10	11	12	13	14
ЗДП	УА-1 - АНСГ	-	-	-	-	-	-
Вердикт	=	+	+	+	+	+	+
Текст	15	16	17	18	19	20	21
ЗДП	УА-1 - АНСГ	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+

Продолжение табл. 24

Текст	22	23	24	25	
ЗДП	УА-1 - АНСГ	-	УА-1 - АНСГ	УА-1 - АНСГ	
Вердикт	+	+	+	+	

Таблица 25. Результаты повторного анализа для официально-делового стиля

Текст	1	2	3	4	5	6	7
ЗДП	-	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	8	9	10	11	12	13	14
ЗДП	УА-1 - АНСГ	-	-	-	-	-	УА-1 - АНСГ
Вердикт	=	+	+	+	+	+	=
Текст	15	16	17	18	19	20	21
ЗДП	-	-	-	-	-	-	-
Вердикт	+	+	+	+	+	+	+
Текст	22	23	24	25			
ЗДП	-	-	-	УА-1			
Вердикт	+	+	+	+			

В таблицах 22-25 используются следующие сокращения и условные обозначения: ЗДП – запускавшиеся дополнительные проверки (в данной строке указывается, какие именно процедуры вспомогательного анализа из четырех имеющихся были применены программой), «+» - верный вердикт, «=» - нечеткий вердикт. В строке ЗДП знак «-» означает, что никакие уточняющие и расширенные исследования не запускались.

В целях экономии места в случаях, когда количество запускавшихся дополнительных проверок превышает 3, они описываются не путем их перечисления, а путем указания диапазона. Во всех случаях вспомогательные процедуры следуют в установленном порядке: 1) УА-1 (уточняющий анализ первого уровня – для исследования по 10 словарным группам), 2) РА (расширенный анализ по 14 словарным группам), 3) УА-2 (уточняющий анализ второго уровня – для исследования по 14 словарным группам), 4) АНСГ – анализ по надежным словарным группам. Соответственно, диапазон вида, к примеру,

«УА-1 - АНСГ» означает, что были последовательно запущены все четыре дополнительные проверки.

Таким образом, мы можем заключить, что переход на новую версию аналитического алгоритма улучшил качество детектирования, демонстрируемое МФСП-анализатором. Общая точность работы программного обеспечения возросла до 97%; в трех оставшихся случаях вердикт был нечетким по вине аномалий средней длины слова. В целом противоречие базовых диагностических критериев было отмечено на 11 текстах из 100; из них в двух случаях оно было снято посредством уточняющего анализа первого уровня, в одном случае – при помощи расширенного анализа, еще в одном случае потребовался уточняющий анализ второго уровня, а в четырех случаях несоответствие было ликвидировано за счет анализа по надежным словарным группам. Тексты художественного и публицистического стилей были верно классифицированы в 100% случаев, так что точность определения свободных стилей также составила 100%. Для научного стиля надежность классификации оказалась равна 96%, для официально-делового – 92%; суммарная точность определения формальных стилей составила, таким образом, 94%.

По завершении отладки программное обеспечение «МФСП-анализатор текстов» было сопоставлено с веб-сервисом определения функционально-стилевой принадлежности текстов «Худломер» (http://teneta.rinet.ru/hudlomer/), в описании которого заявлены возможности различения разговорного, художественного, публицистического и научного стилей. В работе данного сервиса задействованы спектр длин слов, инвариант Фоменко, законы Ципфа и Гиро. При проверке указанного анализатора на описанной выше тестовой выборке было зарегистрировано 36 абсолютно верных вердиктов (ФС образца совпал с данными сервиса), 28 частично верных (ФС образца принадлежал к тому же типу, что и выведенный анализатором) и 11 ошибочных вердиктов (ФС образца не только был определен неверно, но и принадлежал к другому типу). Тексты официально-делового стиля, не обрабатываемые ПО «Худломер», были проверены дополнительно на возможность их типизации; из 25 образцов 7 были классифицированы ошибочно (как тексты одного из свободных стилей),

18 – условно корректно (как тексты формализованного ФС).

Итак, полученные в результате испытаний данные подтверждают выдвинутое нами исходное предположение о том, что для довольно уверенного разграничения свободных и формальных функциональных стилей речи русского языка может быть достаточной оценка на основании двух диагностических критериев – средней длины слова и потенциальной полисемантичности, определяемой по преобладанию в тексте слов, относящихся к определенным словарным группам. Итоги проверки разработанного нами программного обеспечения на широкой выборке текстов свидетельствуют о том, что выявленные закономерности в целом имеют универсальный характер. Алгоритм, реализованный в виде демонстрационного образца – продукта SRS, – после необходимой адаптации может использоваться для скоростного разграничения потока текстов на свободные и формальные в условиях реальной работы систем автоматизированной обработки, позволяя тем самым оптимизировать и улучшать их качество и производительность посредством внедрения формально-семантических аналитических механизмов. Соответственно, глобальную задачу, поставленную нами перед началом проведения исследовательских работ по данному направлению, можно считать выполненной успешно.

Тем не менее, мы считаем необходимым вкратце рассмотреть дополнительный вопрос о вероятном разграничении конкретных функциональных стилей на основании тех данных, которые в настоящее время имеются в нашем распоряжении (т.е. статистических сведений о массовых долях тех или иных словарных групп в текстах различной функционально-стилевой принадлежности, полученных на тестовой выборке). Как уже было отмечено выше, критерии средней длины слова и потенциальной полисемантичности имеют достаточную силу для различения типов стилей (свободных и формальных), однако внутри указанных групп не обладают необходимым разграничительным потенциалом, что обусловлено относительной близостью соответствующих показателей однотипных стилей. Однако не исключено, что их специфика все же будет находить заметное выражение в массовых долях тех или иных словарных групп; к примеру, М.Н. Кожина указывает,

что публицистический стиль характеризуется активным использованием слов с префиксом «не-», а, следовательно, доля словарной группы «Н» в газетных текстах может быть выше. Очевидно, что при таком исследовании мы будем учитывать уже не только слабо и сильно полисемантизованные, а вообще все имеющиеся словарные группы, пытаясь выявить частотные закономерности, не связанные непосредственным образом с потенциальной полисемантичностью текстов.

Для этих целей мы построили сопоставительные графики, отражающие изменение среднеарифметических значений массовых долей различных словарных групп в свободных стилях (рис. 13) и в формальных стилях (рис. 14). Подобное визуальное представление позволяет успешно выявлять несоответствия и несовпадения показателей (в случае наличия таковых), которые могут быть использованы в качестве детектирующих критериев. Очевидные различия среднеарифметических показателей могут указывать на принципиальную разницу индивидуальных параметров и, следовательно, на существование определенной тенденции, которая потенциально способна выступить в роли диагностического инструмента.

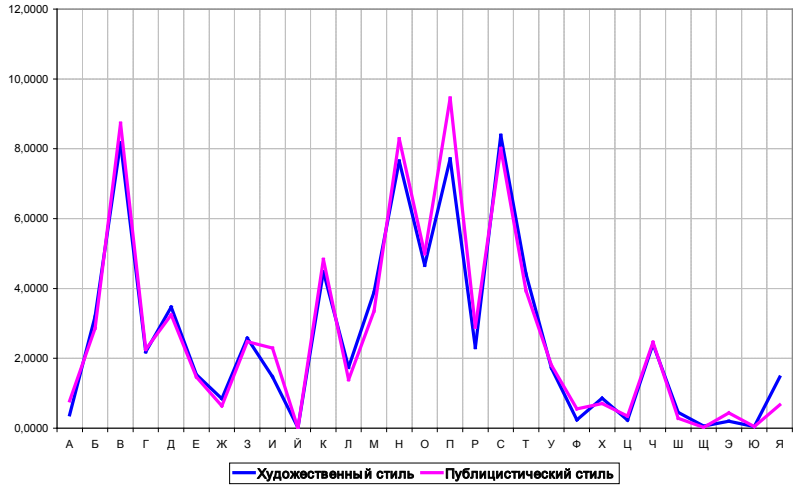

Рис. 13. Сопоставительный график для свободных стилей

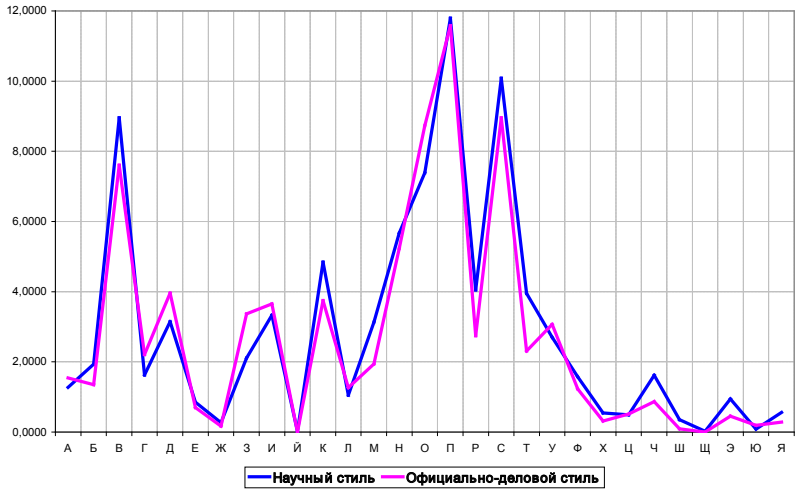

Рис. 14. Сопоставительный график для формальных стилей

Графики приведены в единой размерности: количество делений оси ординат не подобрано под конкретные экстремумы в автоматическом режиме, а установлено вручную на идентичном уровне – от 0 до 12% с ценой деления 2%.

На основании анализа графиков можно отметить, что словарная группа «Н» действительно является довольно четким разграничителем свободных и формальных стилей наряду со слабо и сильно полисемантизованными словарными группами, но, как и в случае с показателями потенциальной полисемантичности, не позволяет проводить четких различий между художественным и публицистическим стилями.

В то же время на сопоставительных графиках заметны некоторые выраженные несоответствия: к примеру, переход от «З» к «И» для художественного стиля является более резким, а для публицистического – более сглаженным. Похожее явление наблюдается и на графике для формальных стилей. Всего наш анализ позволил выявить 10 достаточно четких несоответствий, по 5 для каждого типа функциональных стилей; все они обладают тем же характером, что и приведенный в пример, т.е. разница состоит в резкости либо сглаженности пере-

ходов. Это приводит нас к заключению, что, изучая разности массовых долей некоторых словарных групп и сопоставляя их с эталоном, можно с некоторой точностью классифицировать текст по его конкретной функционально-стилевой принадлежности внутри свободного и формального типов.

Укажем отобранные нами потенциальные диагностические показатели.

Таблица 26. Разности, выбранные для испытаний

Свободные стили:		Формальные стили:	
1) З-И	4) Э-Ю	1) В-Г	4) Э-Ю
2) Ф-Х	5) П-Я	2) З-И	5) Т-У
3) Щ-Э		3) Ч-Ш	

Для каждой пары мы будем находить разности и определять эталон в виде среднего арифметического, после чего сравнивать полученные разности с эталоном. Отличие от указанного эталона будет трактоваться соответствующим образом. Так как всего критериев пять для каждого случая, финальный вердикт будет приниматься по трем и более преобладающим показателям.

Представим полученные данные в статистических таблицах 27-30.

Таблица 27. Исследование разностей, художественный стиль

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**
З-И	0,9868	0,8094	0,5753	0,6393	0,6139	0,4094	1,0399
Ф-Х	-0,4285	-0,6532	-0,6181	-0,3166	-0,5096	-0,3412	-0,0918
Щ-Э	-0,1836	-0,0379	-0,1082	-0,1178	-0,0174	0,0341	-0,0306
Э-Ю	0,1892	0,0584	0,1225	0,0734	-0,0405	0,0000	-0,0102
П-Я	6,2508	8,2487	6,2803	8,4575	5,1891	4,4012	5,4440
Критерии	1-1-1-1-1	1-1-1-1-0	0-1-1-1-1	0-0-1-1-0	0-1-1-1-1	0-0-1-1-1	1-0-1-1-1
Сумма	**5 (ХС)**	**4 (ХС)**	**4 (ХС)**	**2 (ПС)**	**4 (ХС)**	**3 (ХС)**	**4 (ХС)**
Текст	**8**	**9**	**10**	**11**	**12**	**13**	**14**
З-И	1,4913	0,9621	1,0282	0,5634	1,3502	1,3382	1,0689
Ф-Х	-0,4221	-0,5278	-0,6114	-0,7440	-1,5269	-0,3345	-0,6651

Щ-Э	0,0844	-0,0606	-0,5141	-0,1084	-0,6846	-0,1673	**-0,0178**
Э-Ю	-0,0281	0,0556	0,5280	0,1011	0,6940	0,1115	**0,0119**
П-Я	5,4305	6,7323	5,0785	4,4713	4,7636	6,2448	**8,5278**
Критерии	1-1-1-1-1	1-1-1-1-1	1-1-0-0-1	0-1-1-1-1	1-1-0-0-1	1-0-1-1-1	**1-1-1-1-0**
Сумма	**5 (ХС)**	**5 (ХС)**	**3 (ХС)**	**4 (ХС)**	**3 (ХС)**	**4 (ХС)**	**4 (ХС)**
Текст	**15**	**16**	**17**	**18**	**19**	**20**	**21**
З-И	1,5066	0,5777	2,4857	0,1121	3,0080	0,5540	1,0091
Ф-Х	-0,9303	-0,6565	-1,1928	-0,3924	-0,6296	-0,1278	-0,1261
Щ-Э	-0,0971	-0,3939	-0,0546	0,0561	0,1049	0,0000	-0,0631
Э-Ю	0,1533	0,4202	0,1366	0,1121	0,0000	0,0852	0,0315
П-Я	7,7014	6,9065	9,5875	6,8946	7,7300	5,8172	6,0864
Критерии	1-1-1-1-0	0-1-0-0-1	1-1-1-1-0	0-1-1-1-1	1-1-1-1-0	0-0-1-1-1	1-0-1-1-1
Сумма	**4 (ХС)**	**2 (ПС)**	**4 (ХС)**	**4 (ХС)**	**4 (ХС)**	**3 (ХС)**	**4 (ХС)**
Текст	**22**	**23**	**24**	**25**	**Среднее значение**		**Эталон**
З-И	1,1719	1,0852	1,5428	1,7256	1,1062		0,6452
Ф-Х	-1,2109	-1,0730	-0,4611	-0,9792	-0,6228		-0,3896
Щ-Э	-0,4743	-0,1768	-0,1507	-0,2150	-0,1358		-0,2755
Э-Ю	0,4967	0,1524	0,1419	0,2329	0,1532		0,2724
П-Я	5,1116	5,6880	4,9832	4,1915	6,2487		7,5159
Критерии	1-1-0-0-1	1-1-1-1-1	1-1-1-1-1	1-1-1-1-1			
Сумма	**3 (ХС)**	**5 (ХС)**	**5 (ХС)**	**5 (ХС)**			

Итого: 23 вердикта в пользу художественного стиля, 2 вердикта в пользу публицистического стиля (точность определения – 92%). Совпадение всех 5 субвердиктов в пользу художественного стиля отмечено в 6 случаях, 4 из 5 – в 12 случаях, 3 из 5 – в 5 случаях. В остальных 2 случаях – 3 из 5 вердиктов совпали в пользу публицистического стиля.

Соотношение верных / неверных субвердиктов:

«З-И» – 17 / 8 ; «Ф-Х» – 19 / 6; «Щ-Э» – 21 / 4;

«Э-Ю» – 21 / 4; «П-Я» – 19 / 6.

Таблица 28. Исследование разностей, публицистический стиль

Текст	1	2	3	4	5	6	7
З-И	0,2075	1,0252	0,3554	1,0502	-0,4902	-0,7979	0,2358
Ф-Х	0,1037	0,0000	0,6664	-0,3707	0,0000	-0,2660	-1,2972
Щ-Э	-0,2075	-0,5521	-0,6219	-0,3089	-0,4289	-1,1968	-1,1792
Э-Ю	0,2075	0,4732	0,6219	0,3089	0,3676	1,0638	1,1792
П-Я	10,3734	8,7539	8,7517	6,6516	9,5588	8,1117	6,9575
Критерии	0-0-1-1-0	1-0-0-0-0	0-0-0-0-0	1-0-0-0-1	0-0-0-0-0	0-0-0-0-0	0-1-0-0-1
Сумма	**2 (ПС)**	**1 (ПС)**	**0 (ПС)**	**2 (ПС)**	**0 (ПС)**	**0 (ПС)**	**2 (ПС)**
Текст	**8**	**9**	**10**	**11**	**12**	**13**	**14**
З-И	-0,2844	0,6049	0,2042	0,7221	-0,1221	2,1811	0,8782
Ф-Х	-0,4739	0,3257	0,3063	-0,0688	-0,1221	-1,2558	-0,4391
Щ-Э	-0,1896	-0,0465	-0,7147	-0,0688	-0,2442	-0,3966	-0,9879
Э-Ю	0,0948	0,0000	0,6126	0,0688	0,2442	0,3966	1,3172
П-Я	4,7393	10,0512	10,1072	11,0385	10,0122	8,0635	6,4215
Критерии	0-1-1-1-1	0-0-1-1-0	0-0-0-0-0	1-0-1-1-0	0-0-1-1-0	1-1-0-0-0	1-1-0-0-1
Сумма	**4 (ХС)**	**2 (ПС)**	**0 (ПС)**	**3 (ХС)**	**2 (ПС)**	**2 (ПС)**	**3 (ХС)**
Текст	**15**	**16**	**17**	**18**	**19**	**20**	**21**
З-И	-0,7119	-0,1330	0,9081	-0,7968	-0,7117	1,3655	-1,2821
Ф-Х	0,6572	0,1330	-1,0216	0,0000	-0,2372	0,3151	-0,1068
Щ-Э	0,0000	-0,3103	0,0000	-0,3984	-0,1186	-0,4202	-0,4274
Э-Ю	0,0000	0,3103	0,0000	0,5312	0,1186	0,4202	0,3205
П-Я	8,9266	9,7518	10,7832	7,8353	6,8802	9,7689	8,3333
Критерии	0-0-1-1-0	0-0-0-0-0	1-1-1-1-0	0-0-0-0-0	0-0-1-1-1	1-0-0-0-0	0-0-0-0-0
Сумма	**2 (ПС)**	**0 (ПС)**	**4 (ХС)**	**0 (ПС)**	**3 (ХС)**	**1 (ПС)**	**0 (ПС)**
Текст	**22**	**23**	**24**	**25**	**Среднее значение**		**Эталон**
З-И	0,2490	0,2679	-0,3197	0,0000	0,1842		0,6452
Ф-Х	-0,2490	-0,0893	-0,7459	0,3265	-0,1564		-0,3896
Щ-Э	-0,4149	-0,4464	-0,2131	-0,4898	-0,4153		-0,2755
Э-Ю	0,3320	0,4464	-0,0533	0,4082	0,3916		0,2724
П-Я	8,7967	10,1786	9,7496	8,9796	8,7830		7,5159
Критерии	0-0-0-0-0	0-0-0-0-0	0-1-1-1-0	0-0-0-0-0			
Сумма	**0 (ПС)**	**0 (ПС)**	**3 (ХС)**	**0 (ПС)**			

Итого: 19 вердиктов в пользу публицистического стиля, 6 вердиктов в пользу художественного стиля (точность определения – 76%). Совпадение всех 5 субвердиктов в пользу

публицистического стиля отмечено в 10 случаях, 4 из 5 – в 2 случаях, 3 из 5 – в 7 случаях. В 2 случаях – 4 из 5 вердиктов совпали в пользу художественного стиля, в 4 случаях – 3 из 5.

Соотношение верных / неверных субвердиктов:

«З-И» – 18 / 7; «Ф-Х» – 19 / 6; «Щ-Э» – 16 / 9;

«Э-Ю» – 16 / 9; «П-Я» – 20 / 5.

Таблица 29. Исследование разностей, научный стиль

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**
В-Г	9,5472	8,3333	6,5977	6,9705	9,7072	5,2883	5,8787
З-И	2,1654	-0,3788	-0,6468	1,4745	2,7735	1,2361	1,7679
Ч-Ш	1,6732	0,4545	0,5175	0,9383	1,2327	0,8384	0,8371
Э-Ю	0,4921	0,3788	0,2587	0,2681	0,6163	1,1232	1,0754
Т-У	2,7559	-1,0606	-1,4230	5,7641	-3,0817	1,6983	2,1920
Критерии	1-1-1-0-1	1-0-0-0-0	1-0-0-0-0	1-1-0-0-1	1-1-1-1-0	0-1-0-1-1	0-1-0-1-1
Сумма	**4 (НС)**	**1 (ОС)**	**1 (ОС)**	**3 (НС)**	**4 (НС)**	**3 (НС)**	**3 (НС)**
Текст	**8**	**9**	**10**	**11**	**12**	**13**	**14**
В-Г	4,9777	7,7653	6,9218	7,1934	8,4008	7,2768	8,2364
З-И	1,7180	1,1960	1,9212	1,4448	0,4049	1,8914	1,9303
Ч-Ш	1,6334	3,1569	0,9805	1,5192	0,4049	1,4935	1,2461
Э-Ю	0,7767	0,8803	1,3972	0,5887	0,6073	0,7904	1,8913
Т-У	2,3625	1,6309	0,9928	0,5786	-1,9231	3,0361	0,6920
Критерии	0-1-1-1-1	1-1-1-1-1	1-1-0-1-1	1-1-1-1-1	1-0-0-1-0	1-1-1-1-1	1-1-1-1-1
Сумма	**4 (НС)**	**5 (НС)**	**4 (НС)**	**5 (НС)**	**2 (ОС)**	**5 (НС)**	**5 (НС)**
Текст	**15**	**16**	**17**	**18**	**19**	**20**	**21**
В-Г	7,6987	9,8289	6,0974	6,3867	6,0862	6,9767	8,4956
З-И	1,5765	0,9306	1,8018	1,3092	0,8453	1,4435	2,4779
Ч-Ш	3,0453	0,9987	1,8056	1,1621	0,4227	1,0425	1,4159
Э-Ю	0,3715	0,8046	0,4970	3,7784	2,1978	1,0425	0,6195
Т-У	2,8918	1,7248	1,6181	1,6939	1,1834	0,8821	4,4248
Критерии	1-1-1-0-1	1-1-0-1-1	0-1-1-0-1	1-1-1-1-1	0-1-0-1-1	1-1-1-1-1	1-1-1-1-1
Сумма	**4 (НС)**	**4 (НС)**	**3 (НС)**	**5 (НС)**	**3 (НС)**	**5 (НС)**	**5 (НС)**
Текст	**22**	**23**	**24**	**25**	**Среднее значение**		**Эталон**
В-Г	3,4996	8,0332	6,1469	10,8563	7,3281		6,3595

Продолжение табл. 29

З-И	1,0499	0,7618	0,0000	-0,4587	1,2254	0,7545
Ч-Ш	1,3123	-0,2770	0,7496	2,9052	1,2604	1,0180
Э-Ю	0,2625	0,6233	0,0000	0,3058	0,8659	0,5618
Т-У	3,1496	1,3158	-2,9985	1,2232	1,2529	0,0000
Критерии	0-1-1-0-1	1-1-0-1-1	0-0-0-0-0	1-0-1-0-1		
Сумма	**3 (НС)**	**4 (НС)**	**0 (ОС)**	**3 (НС)**		

Итого: 21 вердикт в пользу научного стиля, 4 вердикта в пользу официально-делового стиля (точность определения – 84%). Совпадение всех 5 субвердиктов в пользу научного стиля отмечено в 7 случаях, 4 из 5 – в 7 случаях, 3 из 5 – также в 7 случаях. В 1 случае – 5 из 5 вердиктов совпали в пользу официально-делового стиля, в 2 случаях – 4 из 5, в 1 случае – 3 из 5.

Соотношение верных / неверных субвердиктов:
«В-Г» – 18 / 7; «З-И» – 20 / 5; «Ч-Ш» – 14 / 11;
«Э-Ю» – 16 / 9; «Т-У» – 20 / 5.

Таблица 30. Исследование разностей, официально-деловой стиль

Текст	**1**	**2**	**3**	**4**	**5**	**6**	**7**
В-Г	7,3074	4,8839	1,6886	6,3636	4,7851	8,1897	7,7270
З-И	1,2776	2,9929	-1,5009	0,4399	-0,1977	-2,7094	1,5669
Ч-Ш	0,5033	1,4592	0,0938	1,4370	0,9755	1,4163	1,4810
Э-Ю	0,7065	0,0447	0,0938	0,5279	0,1582	0,5337	0,4507
Т-У	-0,3484	-0,1340	-0,5629	0,0587	-2,2542	-0,5542	-1,6313
Критерии	1-1-0-1-0	0-1-1-0-0	0-0-0-0-0	1-0-1-0-1	0-0-0-0-0	1-0-1-0-0	1-1-1-0-0
Сумма	**3 (НС)**	**2 (ОС)**	**0 (ОС)**	**3 (НС)**	**0 (ОС)**	**2 (ОС)**	**3 (НС)**
Текст	**8**	**9**	**10**	**11**	**12**	**13**	**14**
В-Г	5,2676	7,1659	4,9941	5,5150	6,0438	2,7666	2,1575
З-И	0,6819	-0,5636	-1,3763	3,6545	-3,9975	0,8478	0,1711
Ч-Ш	0,1534	0,5636	0,5505	0,1329	0,2379	0,2231	0,4315
Э-Ю	0,1023	-0,3221	0,1573	-0,1993	0,4124	-0,4462	-0,1786
Т-У	-0,8524	2,2544	-1,6123	-3,9203	-0,9518	-3,9268	-0,2120
Критерии	0-0-0-0-0	1-0-0-0-1	0-0-0-0-0	0-1-0-0-0	0-0-0-0-0	0-1-0-0-0	0-0-0-0-0
Сумма	**0 (ОС)**	**2 (ОС)**	**0 (ОС)**	**1 (ОС)**	**0 (ОС)**	**1 (ОС)**	**0 (ОС)**

Продолжение табл. 30

Текст	15	16	17	18	19	20	21
В-Г	8,7404	5,2051	3,9475	5,0579	7,8327	0,6295	4,1993
З-И	1,4396	0,9860	0,3151	0,4697	2,6220	-1,0715	-0,2614
Ч-Ш	0,5398	0,5478	0,7177	0,7046	0,1328	0,4152	0,9409
Э-Ю	0,4370	0,0484	0,2188	0,0084	3,3190	0,2545	-0,1045
Т-У	-1,0797	-0,3465	-0,4114	-0,0839	-4,9120	1,8350	-0,0174
Критерии	1-1-0-0-0	0-1-0-0-0	0-0-0-0-0	0-0-0-0-0	1-1-0-1-0	0-0-0-0-1	0-0-0-0-0
Сумма	2 (ОС)	1 (ОС)	0 (ОС)	0 (ОС)	3 (НС)	1 (ОС)	0 (ОС)

Текст	22	23	24	25	Среднее значение	Эталон
В-Г	2,1989	5,7866	8,9845	7,3369	5,3910	6,3595
З-И	1,3290	-1,0132	2,0023	-1,0164	0,2835	0,7545
Ч-Ш	0,4833	0,5974	0,5134	4,1397	0,7757	1,0180
Э-Ю	-0,1571	-0,1054	0,3696	0,1109	0,2576	0,5618
Т-У	0,3987	-0,7028	1,5505	-0,6653	-0,7633	0,0000
Критерии	0-1-0-0-1	0-0-0-0-0	1-1-0-0-1	1-0-1-0-0		
Сумма	2 (ОС)	0 (ОС)	3 (НС)	2 (ОС)		

Итого: 20 вердиктов в пользу официально-делового стиля, 5 вердиктов в пользу научного стиля (точность определения – 80%). Совпадение всех 5 субвердиктов в пользу научного стиля отмечено в 10 случаях, 4 из 5 – в 4 случаях, 3 из 5 – в 6 случаях. В 5 случаях – 3 из 5 вердиктов совпали в пользу научного стиля.

Соотношение верных / неверных субвердиктов:
«В-Г» – 16 / 9; «З-И» – 15 / 10; «Ч-Ш» – 20 / 5;
«Э-Ю» – 23 / 2; «Т-У» – 20 / 5.

Итак, можно заключить, что на основании дополнительной процедуры анализа некоторых ключевых разностей можно разграничивать функциональные стили с относительно высокой точностью – от 76 до 92 процентов, в среднем - 83%. Соответственно, МФСП-анализатор может быть дополнен такой процедурой для последующего четкого определения функциональных стилей.

Отметим, что подобное исследование применимо лишь после того, как основной аналитический алгоритм классифицирует изучаемый текст по его типу («свободный» - «формаль-

ный»), поскольку разностная процедура построена именно на отличиях однотипных стилей, а не на признаках, которые дифференцируют тот или иной функциональный стиль вообще ото всех прочих. Таким образом может быть реализовано последовательное сужение вердикта и снятие неопределенности: первым шагом является разделение по типам, вторым шагом – разграничение по конкретным функциональным стилям.

Кроме того, мы считаем очевидно необходимым указать, что полученный результат существенно менее точен, нежели итог, получаемый посредством основного аналитического алгоритма. Вердикты, выносимые в процессе работы разностной процедуры, обладают недостаточной четкостью; для улучшения показателей может требоваться последующее накопление и изучение статистики на материале более объемной выборки, а также построение вспомогательных подалгоритмов, позволяющих устранять ложные срабатывания. Подобные расширенные исследования неизбежно остаются вне пределов нашей текущей работы, в силу чего мы полагаем уместным ограничиться лишь указанием на потенциальную применимость разностного анализа в интересах детектирования конкретных функциональных стилей с намерением вернуться к решению этой задачи в будущем.

Таким образом, поставленные ранее вопросы, связанные с применением формальных маркеров функционально-стилевой принадлежности текстов в интересах их автоматической классификации, а также с реализацией МФСП-алгоритма в виде программного обеспечения, можно считать рассмотренными.

3.4. ВЫВОДЫ

Описан алгоритм классификации текстов на основании минимального набора МФСП и реализующее его программное обеспечение, а также изложены результаты проверки надежности построенного программного продукта на тестовой выборке. Проведенное исследование характеристик типичных образцов текстов различных функциональных стилей позволило определить эталонные значения средней длины слова и потенциальной полисемантичности текста, а также иссле-

довать возможность применения тех или иных полученных показателей в целях идентификации стилевой принадлежности текста. Установлено, что соотношения между массовыми долями сильно полисемантизованных и слабо полисемантизованных групп, а также между количеством слов, принадлежащих к сильно и слабо полисемантизованным словарным группам, различаются для текстов всех стилей и обладают необходимым потенциалом для успешной типизации текстовых образцов. Также в ходе практического исследования было статистически подтверждено, что художественный и публицистический тексты характеризуются относительно низким показателем средней длины слова, в то время как научный и официально-деловой тексты имеют относительно высокие показатели СДС. На основании последующих статистических исследований было установлено, что возможности аналитического алгоритма, основанного на оценивании средней длины слова и потенциальной полисемантичности, достаточны для уверенной классификации текстов по признаку функционально-стилевой отнесенности на свободные и формальные.

Данный алгоритм был реализован в виде аналитического приложения, проверка которого на выборке из 100 текстов позволила не только изучить эффективность разработанного программного обеспечения, но и собрать более объемные и детализированные статистические данные, которые позволили верифицировать точность предложенных критериев и произвести их надлежащую корректировку, а также убедиться в принципиальной способности анализатора классифицировать художественные и публицистические тексты как свободные, а научные и официально-деловые – как формальные, сохраняя при этом достаточно высокую степень достоверности и низкую долю ложных либо нечетких вердиктов. По результатам тестовых прогонов и отладочных работ были определены окончательные параметры аналитического алгоритма. Общая точность его работы в итоге составила 97%; тексты художественного и публицистического стилей были верно классифицированы в 100% случаев, так что точность определения свободных стилей также составила 100%. Для научного стиля надежность классификации оказалась равна 96%, для официально-делового – 92%; суммарная точность опре-

деления формальных стилей составила, таким образом, 94%.

Полученные данные подтвердили выдвинутое исходное предположение о том, что для довольно уверенного разграничения свободных и формальных функциональных стилей речи русского языка может быть достаточной оценка на основании двух диагностических критериев – средней длины слова и потенциальной полисемантичности, определяемой по преобладанию в тексте слов, относящихся к определенным словарным группам.

Кроме того, в качестве дополнительного изучен вопрос о вероятном использовании частотных характеристик тех или иных словарных групп для нужд детализированной классификации текстов в пределах свободного / формального типов. Анализ данных, полученных на тестовой выборке, позволил разработать дополнительный этап автоматического исследования текстов, базирующийся на сопоставлении разностей показателей некоторых словарных групп; в результате было установлено, что разностный анализ позволяет разграничивать функциональные стили с относительно высокой точностью – от 76 до 92 процентов, в среднем - 83%. Установлено, что последующее накопление и изучение статистики на материале более объемной выборки, а также построение вспомогательных подалгоритмов, позволяющих устранять ложные срабатывания, может способствовать дальнейшему повышению точности выносимых вердиктов.

ЗАКЛЮЧЕНИЕ

Изучение формально-семантического анализа как средства оптимизации систем автоматизированной обработки текстов позволило нам сделать ряд выводов и обобщений.

Исследователями в области лексической семантики отмечается, что значение слова формируется не только на основании лишь его собственной семантики как неотъемлемой характеристики языкового знака, но и на базе некоторых внешних элементов – в частности, контекста, как физического (непосредственного окружения слова в тексте), так и идеального (экстралингвистических особенностей речевой ситуации). В то время как физический контекст неплохо изучен, и его специфика учитывается во многих алгоритмах АСОТ, идеальный контекст редко отражается в упомянутых алгоритмах, поскольку с трудом поддается формализации.

В ходе работы с теоретическими источниками установлено, что существующие и перспективные системы автоматизированной обработки текстов на естественных языках хотя и учитывают новейшие разработки в области искусственного интеллекта, лишь изредка приближаются к адекватному выполнению своих непосредственных задач, что формирует потребность в оптимизации их производительности и качества посредством особых вспомогательных аналитических алгоритмов.

Определено, что средством разрешения упомянутых выше проблем выступает формально-семантический анализ многозначной лексики, который предоставляет возможность дальнейшего совершенствования АСОТ и позволяет учитывать в процессе выполняемых ими аналитических исследований не только физический, но и идеальный контекст. При этом существует настоятельная необходимость реализации системного подхода к построению и выполнению аналитических процедур при формировании надежного и достоверного лингвистического обеспечения подобных вспомогательных оптимизационных алгоритмов.

В ходе исследования был сделан вывод, что из всех элементов идеального контекста, которые были рассмотрены и

проанализированы, наиболее эффективным с точки зрения оптимизации автоматизированных систем обработки текстов является функциональный стиль, или подъязык. Автоматическая классификация текстовых образцов по признаку их функционально-стилевой принадлежности характеризуется полезностью с точки зрения всех трех основных разновидностей АСОТ – поисковых роботов и каталогизаторов, машинных переводчиков и средств человеко-машинного взаимодействия. Кроме того, специфические черты основных функциональных стилей являются потенциально исчисляемыми, а следовательно, существует возможность выявления и описания машиночитаемых формальных маркеров функционально-стилевой принадлежности естественноязыковых текстов.

Рассмотрев характерные классифицирующие черты различных функциональных стилей, мы пришли к заключению, что образцы книжных стилей, с которыми автоматизированным системам обработки текстов чаще всего приходится иметь дело (художественного, публицистического, научного и официально-делового) могут быть подвергнуты типизации и разделены на свободный и формальный типы. К свободному типу мы отнесли художественный и публицистический стили, а к формальному, соответственно, научный и официально-деловой. В рамках указанных типов стили объединяются на основании отсутствия / наличия стремления к предельной однозначности речи, что в случае формальных стилей выражается в активном использовании терминов и слабо полисемантизованной лексики вообще, а в случае свободных стилей – в обширном применении сильно полисемантизованных слов, создании образности и общем стремлении к многозначности речи. Представленные основания для группировки позволяют актуализировать средства формально-семантического анализа, которые предоставляют возможность формирования представлений о семантике и значении путем исследования внешней стороны языковых знаков (т.е. допускают выход на план содержания через анализ плана выражения).

Обоснована мысль о том, что может существовать непосредственная взаимосвязь между внешними характеристиками слов и степенью их полисемантизации (т.е. количеством

значений), и установлено, что русскоязычные тексты можно характеризовать энтропийной мерой потенциальной полисемантичности, т.е. суммарной мерой неопределенности, которую вносит в текст каждое многозначное слово. Ввиду вышеупомянутой связи выразительных и содержательных аспектов слова определено, что потенциальная полисемантичность текста может оцениваться путем подсчета массовых долей в нем слов, принадлежащих к сильно и слабо полисемантизованным словарным группам.

После изучения характерных особенностей функциональных стилей речи нами был сделан вывод о целесообразности классификации русскоязычных текстов на основании двух диагностических критериев (машиночитаемых формальных маркеров функционально-стилевой принадлежности) – средней длины слова и потенциальной полисемантичности. Предварительные исследования на эталонных текстах показали, что минимизированный аналитический алгоритм, базирующийся на указанных двух критериях, обладает достаточным потенциалом для дальнейшего развития исследований в этом направлении.

В ходе исследования были определены и охарактеризованы формальные маркеры функционально-стилевой принадлежности русскоязычных текстов, позволяющие реализовать формально-семантический подход к их анализу и, соответственно, достичь оптимизации автоматизированных систем обработки текстовых образцов, на основании чего был составлен алгоритм классификации текстов на свободные и формальные, в котором задействуются расчет и сопоставление вышеуказанных диагностических критериев, а также вспомогательные под-алгоритмы предварительной очистки текста от помех и уточнения нечетких вердиктов. На базе данного алгоритма было построено демонстрационное программное обеспечение – МФСП-анализатор текстов, - которое в полуавтоматическом режиме осуществляет классификацию предъявляемых ему образцов.

По завершении работ над алгоритмом и программным обеспечением были проведены испытания программного продукта на выборке из 100 текстов различной жанровой и функционально-стилевой отнесенности. В результате перво-

го тестового прогона МФСП-анализатор продемонстрировал эффективность в 93%; после уточнения диагностических критериев, не влияющих на их сущность, точность работы исследовательского механизма была повышена до 97%.

Полученные результаты позволили доказать выдвинутые ранее предположения. Установлено, что минимизированный аналитический алгоритм, основанный на анализе и сопоставлении двух диагностических критериев – средней длины слова и потенциальной полисемантичности – позволяет с высокой степенью точности производить типизацию русскоязычных текстов по признаку их функционально-стилевой отнесенности на свободные и формальные; мы обосновали наличие связи между мерой потенциальной полисемантичности и типом функционального стиля текста на русском языке; мы доказали, что существует непосредственная взаимосвязь между массовыми долями слабо и сильно полисемантизованных словарных групп с одной стороны и мерой потенциальной полисемантичности русскоязычного текста с другой стороны. Это позволило сделать вывод о том, что задача оптимизации автоматизированных систем обработки текстов может быть решена при использовании средств формально-семантического анализа, реализуемых в виде исследования машиночитаемых формальных маркеров функционально-стилевой принадлежности текстовых образцов.

Перспективы исследования видятся в изыскании таких машиночитаемых формальных маркеров, которые позволяли бы эффективно проводить различение функциональных стилей в пределах описанных групп, т.е. разграничивать художественный стиль с публицистическим, а научный – с официально-деловым, а также успешно определять менее актуальные для основных разновидностей современных АСОТ разговорный и церковно-религиозный функциональные стили речи.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Анисимова Н.П. Современные французские семантические теории (историко-эпистемологический анализ). – Тверь: ТвГУ, 2002. – 196 с.
2. Апресян Ю.Д. Избранные труды. Т.1: Лексическая семантика: Синонимические средства языка. – М.: Яз. рус. культуры, 1995. – 472 с.
3. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. Лингвистическое обеспечение системы ЭТАП-2. – М.: Наука, 1988. – 295 с.
4. Арутюнова Н. Д. Типы языковых значений. Оценка. Событие. Факт. – М.: Наука, 1988. – 339 с.
5. Арутюнова Н.Д. К проблеме функциональных типов лексического значения // Аспекты семантических исследований. – М., 1980. – С. 156-249.
6. Арутюнова Н.Д. Языковая метафора (синтаксис и лексика) // Лингвистика и поэтика. – М., 1979. – С. 147-173.
7. Бакулов А.Д., Леонтьева Н.Н. Теоретические аспекты машинного перевода // Искусственный интеллект. Системы общения и экспертные системы. – М., 1990. – С. 201-216.
8. Баранов А.Н. Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема // Русистика сегодня. – 1998. - № 1-2. – С. 179-191.
9. Баранов А.Н. Введение в прикладную лингвистику. – М.: Эдиториал УРСС, 2001. – 360 с.
10. Баранов А.Н. Категории искусственного интеллекта в лингвистической семантике: фреймы и сценарии. – М.: Наука, 1987. – 53 с.
11. Баранов А.Н. Речевое воздействие и аргументация // Рекламный текст. Семиотика и лингвистика. – М., 2000. – С. 109-163.
12. Баранов А.Н., Караулов Ю.Н. Русская политическая метафора (материалы к словарю). – М.: Институт русского языка РАН, 1991. – 193 с.
13. Беляева Л.Н., Откупщикова Н.И. Автоматический (машинный) перевод // Прикладное языкознание. – СПб., 1996. – С. 360-388.
14. Богуславский И.М. Сфера действия лексических единиц. – М.: Яз. рус. культуры, 1996. – 460 с.

15. Богуславский И.М., Цинман Л.Л. Семантический компонент лингвистического процессора // Семиотика и информатика. – Вып. 30. – М., 1990. – С. 5-30.
16. Бондалетов В.Д. Стилистика русского языка. – Л.: Просвещение, 1989. – 223 с.
17. Вежбицка А. Из книги «Семантические примитивы» //Семиотика. – М., 1983. – С.225-252.
18. Вейценбаум Дж. Возможности вычислительных машин и человеческий разум. – М.: Радио и связь, 1982. – 306 с.
19. Виноград Т. Программа, понимающая естественный язык. – М.: Мир, 1976. – 294 с.
20. Виноград Т. Работа с естественными языками // Современный компьютер. – М., 1985. – С. 145-150.
21. Виноградов В.В. Лексикология и лексикография. – М.: Наука, 1977. – 312 с.
22. Виноградов В.В. Стилистика. Теория поэтической речи. Поэтика. – М.: Изд-во АН СССР, 1963. – 256 с.
23. Гальперин И.Р. Текст как объект лингвистического исследования. – М.: Наука, 1981. – 138 с.
24. Герд А.С. Научно-техническая лексикография // Прикладное языкознание. – СПб., 1996. – С. 287-308.
25. Герд А.С. Проблемы становления и унификации научной терминологии // История отечественного терминоведения. Направления и методы терминологических исследований. – М., 1995. – С.180-196.
26. Головко Н.В. К вопросу о формальной идентификации функционального стиля в русском языке // Вестник Пятигорского государственного лингвистического университета. – 2009. – №2. – С. 16-19.
27. Головко Н.В. Логико-квантитативный аспект теории фиксирования типов языковой информации // Вестник Ставропольского государственного университета. – 2008. – № 3. – Вып. 56. – С. 72-79.
28. Городецкий Б.Ю. Компьютерная лингвистика: моделирование языкового общения. – Вступит. статья в кн.: Новое в зарубежной лингвистике: Вып. 24. – М.: Прогресс, 1989. – С. 5-29.
29. Гринбаум О.Н. Компьютерные аспекты стилеметрии // Прикладное языкознание. – СПб., 1996. – С. 451–465.

30. Демьянков В.З. Использование естественного языка для человеко-машинного общения (на примере языка экономических показателей предприятия) // Теория и практика научно-технической информации. – М., 1976. – С. 87-94.
31. Добров Б.Д., Лукашевич Н.В. Построение структурной тематической аннотации текста // Труды Международного семинара Диалог-98. – Т. 2. – Казань, 1998.
32. Дударь З.В., Шуклин Д.Е. Семантическая нейронная сеть как формальный язык описания и обработки смысла текстов на естественном языке // Радиоэлектроника и информатика. – №3. – Харьков, 2000. – С.72-76.
33. Загоруйко Н.Г. Искусственный интеллект и эмпирическое предсказание. – Новосибирск: НГУ, 1975. – 82 с.
34. Закревская Н.С. Исследование однородности текста с помощью модели скользящего среднего // Квантитативная лингвистика: исследования и модели (КЛИМ-2005). Материалы Всероссийской научной конференции (6-10 июня 2005 г.). – Новосибирск, 2005. – С. 26-34.
35. Звегинцев В.А. Мысли о лингвистике. – М.: МГУ, 1996. – 335 с.
36. Звегинцев В.А. Семасиология. – М.: МГУ, 1957. – 321 с.
37. Звегинцев В.А. Язык и лингвистическая теория. – М.: Эдиториал УРСС, 2001. – 248 с.
38. Зейналова С.С., Самарина Н.В., Шевенко С.М. Разработка системы автоматического анализа японских текстов // Международный семинар по машинному переводу. Тезисы докладов и сообщений. – М., 1975. – С. 37-39.
39. Иорданская Л. Н., Мельчук И. А. К семантике русских причинных предлогов // Московский лингвистический журнал. – 1996. – Т. 2. – С.162-211.
40. Исаев И.А. Опыт автоматизации лексикографических исследований. Система DIALEX // Слово Достоевского. – М., 1996. – С. 254–265.
41. Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. – М.: Наука, 1981. – 368 с.
42. Кибрик А.Е. Очерки по общим и прикладным вопросам языкознания. – М.: МГУ, 1992. – 214 с.
43. Кожина М.Н. Стилистика русского языка. – М.: Флинта, 2008. – 464 с.

44. Крамер Г. Математические методы статистики. – М.: Мир, 1976. – 648 с.
45. Крылов С.А. Композиционно-интегральный металингвистический подход к условно-семантической интерпретации текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2005». – М., 2005. – С. 298-302.
46. Крылов С.А. Семантическая роль как элемент метаязыков общей и специальной типологии // 40 лет Санкт-Петербургской типологической школе. – М., 2004. - С. 233-252.
47. Крылов С.А., Падучева Е.В. Дейксис: общетеоретические и прагматические аспекты // Языковая реальность в аспекте лингвистической прагматики. – М., 1983. – С. 25-26.
48. Крылова О.А. Лингвистическая стилистика. – М.: Высш. школа, 2006. – 320 с.
49. Кубрякова Е.С. Типы языковых значений: Семантика производного слова. – М.: ЛКИ, 2008. – 208 с.
50. Курилович Е. Очерки по лингвистике. – М.: Изд-во иностр. лит, 1962. – 293 с.
51. Лайонз Дж. Введение в теоретическую лингвистику. – М.: Прогресс, 1978. – 543 с.
52. Левин Ю.И. О семантике местоимений // Проблемы грамматического моделирования. – М., 1973. – С. 108-121.
53. Левицкий Ю.А. Семантика русских сочинительных союзов // Проблемы структурной лингвистики. – М., 1981. – С. 83-91.
54. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 1. Моделирование системы «мягкого понимания» текста: информационно-лингвистическая модель. – М.: МГУ, 2000. – 42 с.
55. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 2. Семантические словари: состав, структура, методика создания. – М.: МГУ, 2001. – 40 с.
56. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Часть 3. Семантический компонент. Локальный семантический анализ. – М.: МГУ, 2002. – 49 с.
57. Лидсей Р. Память, способная к выводам, как основа машин, понимающих человеческий язык // Вычислительные машины и мышление. – М., 1967. – С. 233-255.
58. Лосев А.Ф. Знак. Символ. Миф. – М.: МГУ, 1982. – 480 с.

59. Лотте Д.С. Вопросы заимствования и упорядочения иноязычных терминов и терминоэлементов. – М.: Наука, 1982. – 149 с.
60. Максимов Л.Ю. Литературный язык и язык художественной литературы // Анализ художественного текста. – М., 1975. С.11-20.
61. Мартеньянов Ю.С. Заметки о строении ситуации и форме ее описания // Машинный перевод и прикладная лингвистика. – М., 1964. – С. 125-149.
62. Мастерман М. Изучение семантической структуры текста для машинного перевода с помощью языка-посредника // Математическая лингвистика. – М., 1964. – С. 161-178.
63. Мдивани Р.Р. Лингвистические аспекты искусственного интеллекта. – Вступит. статья в кн.: Лингвистика и проблемы машинного «понимания» естественного языка. – М.: Изд-во АН СССР, 1985. – С. 5-13.
64. Мельников Г.П. О принципиальных возможностях передачи содержания текста при машинном переводе на современных ЭВМ // Теория и практика научно-технической информации. Материалы семинара. – М., 1974. – С. 80-85.
65. Мельников Г.П. Типы означаемых языкового знака и детерминанта языка // Проблемы семантики. – М., 1975. – С. 25-34.
66. Мельчук И.А. О некоторых типах языковых значений // О точных методах исследования языка. – М., 1961. – С. 74-102.
67. Мельчук И.А. Опыт теории лингвистических моделей «смысл-текст». – М.: Яз. рус. культуры, 1999. – 346 с.
68. Милль Дж. Ст. Система логики. – М.: Изд-во Г. А. Лемана, 1914. – 880 с.
69. Минский М. На пути к созданию искусственного разума // Вычислительные машины и мышление. – М., 1967. – С. 401-451.
70. Минский М. Структура для представления знаний // Психология машинного зрения. – М., 1978. – С. 250-338.
71. Нагао М. Государственный проект создания системы машинного перевода // Международный форум по информ. и докум. – Т. 9. – № 2. – М., 1984. – С. 11-16.
72. Напалков А.В. Эвристическое программирование. – Ростов-на-Дону: Изд-во Рост. ун-та, 1971. – 127 с.
73. Новиков А.И. Семантика текста и ее формализация. – М.: Наука, 1983. – 215 с.
74. Ожегов С.И. Лексикология. Лексикография. Одноязычные

филологические словари. Культура речи. – М.: Высш. школа, 1974. – 352 с.
75. Падучева Е.В. Высказывание и его соотнесенность с действительностью. – М.: Наука, 1985. – 271 с.
76. Панков И.П., Захаров В.П. Информационно-поисковые системы // Прикладное языкознание. – СПб., 1996. – С.347-349.
77. Панков М.П. Представление и использование знаний в системах понимания языка // Учен. зап. Тарт. ун-та. – Вып. 621. – Тарту, 1983. – С. 101-109.
78. Перцова Н.Н. Проблемы глубинной семантики. – Ч. 2. – М.: Изд-во ИРЯ АН . СССР, 1976. – 45 с.
79. Петров В.В., Переверзев В.Н. Ситуационная семантика. Научно-аналитический обзор. – М.: Наука, 1988. – 145 с.
80. Пеше М. Контент-анализ и теория дискурса // Квадратура смысла. – М., 1999. – С. 302-336.
81. Плотников Б.А. Исследование лексического значения путем квантитативного анализа дистрибутивных свойств слова: Автореф. ... док. филол. н. – Минск, 1981. – 32 с.
82. Поспелов Г.С. Системный анализ и искусственный интеллект. – М.: ВЦ АН СССР, 1980. – 48 с.
83. Преображенский А.Б. Принципы организации словарей лингвистического процессора // Диалоговые системы и представление знаний. – Тарту, 1981. – С. 117-124.
84. Рахилина Е. В. Когнитивный анализ предметных имен: семантика и сочетаемость. – М.: Русские словари, 2000. – 416 с.
85. Селезнев В.А., Исаева Е.В. Параметр Херста словарного ряда // Квантитативная лингвистика: исследования и модели (КЛИМ-2005). Материалы Всероссийской научной конференции (6-10 июня 2005 г.). – Новосибирск, 2005. – С. 146-152.
86. Селиверстова О.Н. Местоимения в языке и речи. – М.: Наука, 1988. – 151 с.
87. Селфридж М. Интегральная обработка обеспечивает надежное понимание // Новое в зарубежной лингвистике: Вып. 24. – М., 1989. – С. 161-208.
88. Семин Г.В. Анализ естественного языка под управлением семантической сети. – М.: МЭИ, 1980. – 60 с.
89. Сидоров Ю.В. Математическая и информационная поддержка методов обработки литературных текстов на основе фор-

мально-грамматических параметров. Автореф. ... канд. тех. наук. – Петрозаводск, 2002. – 19 с.
90. Скрэгг Г. Семантические сети как модели памяти // Новое в зарубежной лингвистике. – Вып. 12. – М.,1983. – С. 228-271.
91. Сокирко А.В. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ). Автореф ... канд. техн. наук. – М., 2001. – 24 с.
92. Солганик Г.Я. Стилистика текста. – М.: Флинта, 2005. – 253 с.
93. Солнцев В.М. Язык как системно-структурное образование. – М.: Наука, 1977. – 366 с.
94. Сэлтон Г. Автоматическая обработка, хранение и поиск информации. – М.: Сов. радио, 1973. – 328 с.
95. Татар Б. Русская лексикография. – Будапешт: Танкеньвкладо, 1992. – 370 с.
96. Телия В.Н. Коннотативный аспект семантики номинативных единиц. – М.: Наука, 1986. – 143 с.
97. Тодоров Ц. Теории символа. – М.: Дом интелл. книги, 1999. – 384 с.
98. Уфимцева А.А. Лексическое значение. Принцип семиологического описания лексики – М.: Наука, 1986. – 240 с.
99. Фитиалов С.Я. Лингвистические аспекты моделирования понимания естественного языка // Учен. зап. Тарт. ун-та. – Вып. 654. – Тарту, 1983. – С. 71-85.
100. Фомичев В.А. Математические основы представления смысла текстов для разработки лингвистических информационных технологий. Часть I. Модель системы первичных единиц концептуального уровня // Информационные технологии. – 2002. – №10. – С. 16-25.
101. Фомичев В.А. Математические основы представления смысла текстов для разработки лингвистических информационных технологий. Часть II. Система правил для построения семантических представлений фраз и сложных связных текстов // Информационные технологии. – 2002. – №11. – С. 34-45.
102. Хаймс Д. Этнография речи // Новое в лингвистике. Вып. 7: Социолингвистика. – М., 1975. – С. 42-95.
103. Хант Э. Искусственный интеллект. – М.: Мир, 1978. – 558 с.
104. Харламов А. Автоматический структурный анализ текстов // Открытые системы. – 2002. – №10. – С. 62-65.
105. Цейтин Г.С. Система обработки текста с процедурным пред-

ставлением словарной информации и перспективы ее применения в машинном переводе // Международный симпозиум по машинному переводу. Тезисы докладов. – М., 1979. С. 35-36.
106. Шаляпина З.М. К проблеме построения формальной модели процесса перевода // Теория перевода и научные основы подготовки переводчиков. – М., 1975. – С. 165-172.
107. Шахмайкин А.М. Некоторые принципы формализации семантики естественных языков (для целей семантической типологии): Автореф. ... канд. филол. наук. – М., 1986. – 23 с.
108. Шевелев О.Г. Представление набора текстов в реляционной базе данных для целей лингвистического анализа // Вестник Томского государственного университета. – 2004. – № 284. – С. 225-229.
109. Шевелев О.Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений. Автореф. ... канд. тех. наук. – Томск, 2006. – 19 с.
110. Шевенко С.М. Использование фреймов для представления знаний в лингвистическом процессоре // 9 Всесоюзный симпозиум по кибернетике. Тезисы докладов. – Т.1. – М., 1981. – С. 189-190.
111. Шевенко С.М. Лингвистика и проблемы машинного «понимания» естественного языка: аналитич. обзор. – М.: Изд-во АН СССР, 1985. – 78 с.
112. Шелов С.Д. Термин. Терминологичность. Терминологические определения. – СПб.: СПбГУ, 2003. – 280 с.
113. Шенк Р. Обработка концептуальной информации. – М.: Энергия, 1980. – 360 с.
114. Шенк Р., Бирнбаум Л., Мей Дж. К интеграции семантики и прагматики // Новое в зарубежной лингвистике. – Вып. 24. М., 1989. – С. 32-47.
115. Шмелев Д.Н. Проблемы семантического анализа лексики. – М.: Эдиториал УРСС, 2008. – 280 с.
116. Штерн И.Б. К интерпретации задачи лингвистического обеспечения человеко-машинных систем, реализующих «производственные» диалоги на естественном языке // Структурная и математическая лингвистика. – Вып. 5. – Киев, 1977. – С. 117-123.
117. Шуклин Д.Е. Структура семантической нейронной сети, из-

влекающей в реальном времени смысл из текста // Кибернетика и системный анализ. – №2. – Киев, 2001. – С.43-48.
118. Щерба Л.В. Языковая система и речевая деятельность. – Л.: Наука, 1974. – 428 с.
119. Ыйм Х., Хойт М., Литвак С. Система с полным циклом понимания естественного языка TARLUS: текущее состояние // 3 международный семинар по машинному переводу. Тезисы докладов. – М., 1983. – С. 249-257.
120. Якобсон Р. В поисках сущности языка // Семиотика. – М.,1983. – С. 102-117.
121. Якобсон Р. Лингвистика и поэтика // Структурализм «за» и «против». – М., 1975. – С. 193-230.
122. Crystal D. The Cambridge Encyclopedia of the English language. – Cambridge, 2003. – 500 p.
123. Fillmore Ch.J. Frames and the semantics of understanding // Quaderni di semantica, 1985. Vol. VI, no. 2, p. 222-254.
124. Gruber S.J. Frame information and lexically-based inference // Quaderni di semantica, 1985. Vol. VI, no. 2, p. 58-78.
125. Hofmann Th.R. Semantic frames and content representation // Quaderni di semantica, 1985. Vol. VI, no. 2, p. 267-284.
126. Holmes D.I. The Evolution of Stylometry in Humanities Scholarship // Literary and Linguistic Computing. – 1998. – №3. – Vol. 13. – Pp. 111-117.
127. Hudson R. Some basic assumptions about linguistic and non-linguistic knowledge // Quaderni di semantica, 1985. Vol. VI, no. 2, p. 284-287.
128. Kiefer F. How to account for situational meaning? // Quaderni di semantica, 1985. Vol. VI, no. 2, p. 288-295.
129. Ôim H., Saluveer M. Frames in linguistic descriptions // Quaderni di semantica, 1985. Vol. VI, no. 2, p. 295-305.
130. Peng R.D., Hengartner N.W. Quantitative analysis of literary styles // The American Statistician. – 2002. – №3. – Vol. 56. – Pp. 175-185.
131. Rocchio J.J. Jr. Relevance feedback in information retrieval // The SMART Retrieval System: Experiments in Automatic Document Processing. – 1971. – Pp. 313-323.
132. Schank R., Abelson R. Scripts, Plans, Goals and Understanding. – Hillsdale, 1977. – 307 p.
133. Tannen D. Frames and schemas in interaction // Quaderni di

semantica, 1985. Vol. VI, no. 2, p. 326-335.
134. Wierzbicka Anna. Negation - a Study in the Deep Grammar. – MIT, 1967. – 2 fld.
135. Wilks Y. Text structures and knowledge structures // Quaderni di semantica, 1985. Vol. VI, no. 2, p. 335-344.

ИНТЕРНЕТ-ИСТОЧНИКИ

1. Азарова И.В., Овчинникова Е.А. Семантическая структура пропозиции при извлечении фактов из текстов на русском языке. http://www.dialog-21.ru/
2. Ананян С., Харламов А. Автоматический анализ естественноязыковых текстов. http://www.megaputer.com/
3. Браславский П.И. Пути повышения эффективности поиска научной информации в Internet. http://www.nsc.ru/
4. Ленец А.В. Речевая ситуация лжи. http://lingvomaster.ru/
5. Матвеева Г.Г. Диагностирование личностных свойств автора по его речевому поведению. http://rspu.edu.ru/
6. Психолингвистическая экспертная система «ВААЛ». Руководство пользователя. http://www.vaal.ru/
7. Решения Cognitive Technologies. http://cognitive.ru/
8. Система TextAnalyst. Описание. http://www.megaputer.com/
9. Технологии анализа и поиска текстовой информации. http://rco.ru/
10. Хмелев Д.В. Классификация и разметка текстов с использованием методов сжатия данных. http://compression.graphicon.ru/
11. Apte C., Damerau F., Weiss S. Automated Learning of Decision Rules for Text Categorization. http://citeseer.ist.psu.edu/
12. Beil F., Ester M., Xu X. Frequent Term-Based Text Clustering. http://ifsc.ualr.edu
13. Cimiano P., Staab S., Tane J. Deriving Concept Hierarchies from Text by Smooth Formal Concept Analysis. http://www.aifb.uni-karlsruhe.de/
14. Ferrucci D. Text analysis as formal inference for the purposes of uniform tracing and explanation generation. http://domino.watson.ibm.com/
15. Mehler A., Waltinger U., Wegner A. A Formal Text Representation Model Based on Lexical Chaining. http://ikw.uni-osnabrueck.de/

16. NooLab – лаборатория информационных технологий. http://noolab.ru
17. Text Mining: добыча знаний. http://visti.net/

СПИСОК ИСТОЧНИКОВ ЛИНГВИСТИЧЕСКОГО МАТЕРИАЛА

1. Merriam-Webster Collegiate Dictionary // Encyclopaedia Britannica Ultimate Reference Suite 2007.
2. Азар И. Настоящих буйных мало. http://lenta.ru/
3. Аксенов В. Сюрпризы. http://lib.ru/
4. Аполлинер Г. Ранние стихотворения (1896-1910). http://lib.ru/
5. Архангельский А. Критические дни. http://vz.ru/
6. Архипов В.В. Заболевания легких при беременности. – М., 2002.
7. Байрон Дж.Г. Паломничество Чайльд-Гарольда. http://lib.ru/
8. Бакаев А.И. Вопросы подготовки и переподготовки служащих: правовой аспект // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
9. Баянова Г.И. Профориентационная работа в общеобразовательном учреждении // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
10. Белецкий Р. Роман романа. http://lib.ru/
11. Белковский С. Сценарий-2012: бизнес времен конца Владимира Путина. http://news.km.ru/
12. Белов В. Привычное дело. http://lib.ru/
13. Березовский В.М. Химия витаминов. – М., 1959.
14. Боровкова Н.В., Боровкова А.С., Зудилин С.Н. Эффективность минеральных удобрений, хелатных комплексов и гуминовых веществ при возделывании озимых культур в лесостепи среднего Поволжья // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.

15. Бродский И. Вертумн. http://lib.ru/
16. Будаев Э.В., Чудинов А.П. Зарубежная политическая метафорология. – Екатеринбург, 2008.
17. Веллер М. Ничего не происходит. http://lib.ru/
18. Воздушный кодекс Российской Федерации от 19 марта 1997 г. № 60-ФЗ. http://garant.ru/
19. Гайсина Л.Ф. Готовность студентов вуза к общению в мультикультурной среде и ее формирование. – Оренбург, 2004.
20. Гиймоль А. Украинский миллиардер выходит из тени. http://inosmi.ru/
21. Голубев А.Е. О стратегических приоритетах в производственной сфере // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
22. Горький М. На дне. http://lib.ru/
23. Дмитриев Т.А. Проблема методического сомнения в философии Р. Декарта. – М., 2007.
24. Дойл А.К. Знак четырех. http://lib.ru/
25. Дрожжаков А.И. Проблемы развития теории физической культуры // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
26. Есенин С. 28 стихотворений. http://lib.ru/
27. Задорнов М. Реформа образования. http://mzadornov.livejournal.com/
28. Зайнетдинов А.Ш., Махмутова А.Ф. Предпосылки организации речевой деятельности на основе субъектно-модальных значений // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
29. Зайцев И. Да не оскудеет рука берущего. http://news.km.ru/
30. Закон Российской Федерации «О государственной тайне» (с изменениями на 6 октября 1997 года). http://garant.ru/
31. Закон Российской Федерации «О рекламе». http://www.consultant.ru/
32. Ибсен Г. Пер Гюнт. http://lib.ru/
33. Иванов С.Н., Метелев А.В. Посттравматические стрессовые состояния сотрудников правоохранительных органов: проблемы психологической экспертизы. – Ижевск, 2008.

34. Ивлев Л.С., Довгалюк Ю.А. Физика атмосферных аэрозольных систем. – СПб., 1999.
35. Информационное письмо Президиума Высшего Арбитражного Суда РФ от 13 августа 2004 г. № 81. http://garant.ru/
36. Каныгин П. В Белоруссии не водятся хомячки, зато есть ручные бизнесмены. http://www.novayagazeta.ru/
37. Кафка Ф. Процесс. http://lib.ru/
38. Кольридж С.Т. Кристабель. http://lib.ru/
39. Конвенция о бесплатной выдаче и освобождении от легализации заверенных копий актов гражданского состояния (Люксембург, 26 сентября 1957 г.). http://garant.ru/
40. Конвенция Содружества Независимых Государств о правах и основных свободах человека (Минск, 26 мая 1995 г.). http://garant.ru/
41. Корт С.С. Теоретические основы защиты информации. М., 2004.
42. Кристи А. Зло под солнцем. http://lib.ru/
43. Ланина Т.Д. Процессы переработки пластовых вод месторождений углеводородов. – Ухта, 2006.
44. Леонтьев М. Помогать США «дожимать» Ливию с нашей стороны было бы бессмысленно и глупо. http://news.km.ru/
45. Лесной кодекс Российской Федерации от 4 декабря 2006 г. № 200-ФЗ. http://garant.ru/
46. Лукутин Б.В. Возобновляемая энергетика в децентрализованном электроснабжении. – М., 2008.
47. Лысенков А. Ступай с миром. http://lenta.ru/
48. Малый академический словарь русского языка в 4-х тт. / Под ред. А. П. Евгеньевой. – 4-е изд., стер. – М., 1999.
49. Мамонтов В. Игнат и асфальт. http://izvestia.ru/
50. Махнова И.Ф. Технология социально-психологического тренинга работников ювелирной отрасли // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
51. Методические рекомендации «Проведение физкультурных занятий на открытом воздухе с детьми 5-7 лет в дошкольных учреждениях» (утв. Минздравом СССР 29 октября 1984 г. № 11-14/26-6). http://garant.ru/
52. Мироненко С.Г., Преображенский А.П. Разработка подсисте-

мы обработки сигналов на основе wavelet-преобразования // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
53. Мольер Ж.-Б. Тартюф. http://lib.ru/
54. Москвина Т. Полководец Вульф. http://www.argumenti.ru/
55. Мурашкина Р. Тото Кутуньо: «В следующей жизни обещаю жениться на русской!». http://kp.ru/
56. Набоков В. Университетская поэма. http://lib.ru/
57. Обзор судебной практики Верховного Суда РФ за третий квартал 1999 года (По гражданским делам) (утв. постановлением Президиума Верховного Суда РФ от 29 декабря 1999 г.). http://garant.ru/
58. Ожегов С.И. Словарь русского языка / Под общ. ред. проф. Л.И. Скворцова. – 25-е изд., испр. и доп. – М., 2008.
59. Островский А. Гроза. http://lib.ru/
60. Петунин Ю.И. Приложение теории случайных процессов в биологии и медицине. – Киев, 1981.
61. Письмо Федеральной налоговой службы от 11 марта 2008 г. № ГИ-6-1/159 «Об основных положениях о порядке составления и передачи статистической налоговой отчетности в 2008 году». http://garant.ru/
62. Письмо ЦБР от 1 ноября 2008 г. № 137-Т «О повышении эффективности работы по предотвращению сомнительных операций». http://garant.ru/
63. Победоносцева Г.М. Перспективы применения теории устойчивого развития для районов Крайнего Севера и Арктики // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
64. Постановление Правительства РФ от 30 ноября 2005 г. № 706 «О мерах по обеспечению проведения обязательного аудита». http://garant.ru/
65. Постановление Правительства РФ от 31 января 2007 г. № 57 «О порядке присвоения имен благотворителей научным и образовательным учреждениям, учреждениям здравоохранения, культуры, социальной защиты, физической культуры и спорта, находящимся в ведении федеральных органов исполнительной власти, а также закрепленным за указанными

учреждениями объектам недвижимого имущества». http://garant.ru/
66. Постановление Совета Министров - Правительства РФ от 12 августа 1993 г. № 793 «Об утверждении Положения о пошлинах за патентование изобретений, полезных моделей, промышленных образцов, регистрацию товарных знаков, знаков обслуживания, наименований мест происхождения товаров, предоставление права пользования наименованиями мест происхождения товаров и Положения о регистрационных сборах за официальную регистрацию программ для электронных вычислительных машин, баз данных и топологий интегральных микросхем». http://garant.ru/
67. Приказ Минюста РФ от 6 октября 2008 г. № 224 «Об утверждении Инструкции по организации планирования деятельности Министерства юстиции Российской Федерации». http://garant.ru/
68. Пришвин М. Путешествие. http://lib.ru/
69. Проскуренко И.В. Фермерское рыбоводное хозяйство. – СПб., 2000.
70. Прядко Л.А. Педагогические условия формирования готовности будущих социальных работников к использованию средств физической культуры в повышении двигательной активности пожилых людей // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
71. Пугачев К. Выручит ли вновь «Спартак» Андрей Дикань? http://www.argumenti.ru/
72. Решение Государственной комиссии по радиочастотам при Минсвязи РФ от 26 февраля 2008 г. № 08-23-02-001 «Об утверждении проекта «Положения о порядке рассмотрения материалов, проведения экспертизы и принятия решения о присвоении (назначении) радиочастот или радиочастотных каналов для радиоэлектронных средств в пределах выделенных полос радиочастот». http://garant.ru/
73. Рулиене Л.Н. Перспективы использования технологий дистанционного обучения в университете // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с между-

народным участием. Т.XI. –Уфа, 2010.
74. Сабирова Л.А. Образ Петербурга в рассказе Л. Лунца «Ненормальное явление». Традиции и новаторство // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
75. Саркисова Е. Бываешь ты, Фемида, молодцом! http://opengaz.ru/
76. Саркисова Е. Культ порока, или Телеящик – на помойку! http://opengaz.ru/
77. Северянин И. Сонет серебряного века. http://lib.ru/
78. Современный толковый словарь русского языка / Под ред. С.А. Кузнецова. - М., 2004.
79. Соколов-Митрич Д. Время застенчивых. http://izvestia.ru/
80. Стуруа М. Цари и цареубийцы. http://www.mk.ru/
81. Субруйяр Р. Китай растет в атмосфере всеобщего безразличия. http://inosmi.ru/
82. Трапезников А. Я в наследстве отцов ощутил... http://www.litrossia.ru/
83. Уайлд О. Избранное. М., 1990.
84. Указ Президента РФ от 16 августа 2004 г. № 1083 «Вопросы Федеральной службы по военно-техническому сотрудничеству». http://garant.ru/
85. Указание ЦБР от 26 августа 2008 г. № 2055-У «О внесении изменений в Указание Банка России от 16 января 2004 года N 1376-У «О перечне, формах и порядке составления и представления форм отчетности кредитных организаций в Центральный банк Российской Федерации». http://garant.ru/
86. Устав (Основной закон) Ямало-Ненецкого автономного округа от 28 декабря 1998 г. № 56-ЗАО. http://garant.ru/
87. Устав Смоленской области от 15 мая 2001 г. № 37-з. http://garant.ru/
88. Утицин О. Россия бандитская. http://www.argumenti.ru/
89. Уэллс Г. Пища богов. http://lib.ru/
90. Федеральный закон «Об электронной цифровой подписи» от 10 января 2002 года №1-ФЗ. http://garant.ru/
91. Федеральный закон от 13 декабря 1996 г. № 150-ФЗ «Об оружии». http://garant.ru/
92. Федеральный закон от 16 июля 1998 г. № 102-ФЗ «Об ипотеке

(залоге недвижимости)». http://garant.ru/
93. Федеральный закон от 31 июля 1998 г. № 155-ФЗ «О внутренних морских водах, территориальном море и прилежащей зоне Российской Федерации». http://garant.ru/
94. Федеральный закон Российской Федерации от 27 июля 2006 г. № 152-ФЗ «О персональных данных». http://garant.ru/
95. Федеральный конституционный закон от 30 мая 2001 г. № 3-ФКЗ «О чрезвычайном положении». http://garant.ru/
96. Чебыкин И. Бывший СССР: вместо карет получили по тыкве. http://news.km.ru/
97. Черняховский С. Достоин презрения. http://news.km.ru/
98. Шагиахметова Л.Р. Воздействие шейпинг-тренировки психомоторной направленности на развитие духовных качеств студенток // Актуальные проблемы современной науки и образования. Материалы Всероссийской научно-практической конференции с международным участием. Т.XI. –Уфа, 2010.
99. Шеридан Р.Б. Школа злословия. http://lib.ru/
100. Шмелев И. Неупиваемая чаша. http://lib.ru/
101. Шолохов М. Батраки. http://lib.ru/
102. Шоу Б. Пигмалион. http://lib.ru/
103. Шустер С. Байден в Молдавии: вице-президент переполошил Москву – но церемонии соблюдены. http://inosmi.ru/
104. Элиот Т.С. Пепельная среда. http://lib.ru/
105. Эскобар П. Политическая романтика. http://inosmi.ru/
106. Юдин Ю. ЧМ по биатлону: Провал? Да, это – провал... http://sport.km.ru/
107. Юсин М. Иракский синдром. http://vz.ru/
108. Ярочкин В.И. Информационная безопасность. – М., 2004.

ПРИЛОЖЕНИЕ

Таблица 1. Полная таблица результатов исследования по словарю Ожегова

Буква	Слов	О	М	СП	СрП	СиП	% О	% М	М / О	% СП	СП / О	% СП (М)	% О+СП	% СрП	СрП / О	% СрП (М)	СрП / СП	% СиП	СиП / О	% СиП (М)	СиП / СП	% СрП+ СиП	СрП+ СиП / О	% рП+ СиП (М)	СрП+ СиП / СП	СрП+ СиП / О+СП
А	726	575	151	130	21	0	79,20%	20,80%	0,2626	17,91%	0,2261	86,09%	97,11%	2,89%	0,0365	13,91%	0,1615	0,00%	0,0000	0,00%	0,0000	2,89%	0,0365	13,91%	0,1615	0,0298
Б	1497	1176	321	250	67	4	78,56%	21,44%	0,2730	16,70%	0,2126	77,88%	95,26%	4,48%	0,0570	20,87%	0,2680	0,27%	0,0034	1,25%	0,0160	4,74%	0,0604	22,12%	0,2840	0,0498
В	2514	1915	599	414	169	16	76,17%	23,83%	0,3128	16,47%	0,2162	69,12%	92,64%	6,72%	0,0883	28,21%	0,4082	0,64%	0,0084	2,67%	0,0386	7,36%	0,0966	30,88%	0,4469	0,0794
Г	993	769	224	156	63	5	77,44%	22,56%	0,2913	15,71%	0,2029	69,64%	93,15%	6,34%	0,0819	28,13%	0,4038	0,50%	0,0065	2,23%	0,0321	6,85%	0,0884	30,36%	0,4359	0,0735
Д	1396	1030	366	275	76	15	73,78%	26,22%	0,3553	19,70%	0,2670	75,14%	93,48%	5,44%	0,0738	20,77%	0,2764	1,07%	0,0146	4,10%	0,0545	6,52%	0,0883	24,86%	0,3309	0,0697
ЕЁ	133	94	39	26	11	2	70,68%	29,32%	0,4149	19,55%	0,2766	66,67%	90,23%	8,27%	0,1170	28,21%	0,4231	1,50%	0,0213	5,13%	0,0769	9,77%	0,1383	33,33%	0,5000	0,1083
Ж	301	206	95	63	28	4	68,44%	31,56%	0,4612	20,93%	0,3058	66,32%	89,37%	9,30%	0,1359	29,47%	0,4444	1,33%	0,0194	4,21%	0,0635	10,63%	0,1553	33,68%	0,5079	0,1190
З	1788	1314	474	340	123	11	73,49%	26,51%	0,3607	19,02%	0,2588	71,73%	92,51%	6,88%	0,0936	25,95%	0,3618	0,62%	0,0084	2,32%	0,0324	7,49%	0,1020	28,27%	0,3941	0,0810
И	935	719	216	172	40	4	76,90%	23,10%	0,3004	18,40%	0,2392	79,63%	95,29%	4,28%	0,0556	18,52%	0,2326	0,43%	0,0056	1,85%	0,0233	4,71%	0,0612	20,37%	0,2558	0,0494
Й	5	4	1	1	0	0	80,00%	20,00%	0,2500	20,00%	0,2500	100,00%	100,00%	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,0000
К	2259	1706	553	396	149	8	75,52%	24,48%	0,3242	17,53%	0,2321	71,61%	93,05%	6,60%	0,0873	26,94%	0,3763	0,35%	0,0047	1,45%	0,0202	6,95%	0,0920	28,39%	0,3965	0,0747
Л	781	592	189	138	46	5	75,80%	24,20%	0,3193	17,67%	0,2331	73,02%	93,47%	5,89%	0,0777	24,34%	0,3333	0,64%	0,0084	2,65%	0,0362	6,53%	0,0861	26,98%	0,3696	0,0699
М	1404	1104	300	212	84	4	78,63%	21,37%	0,2717	15,10%	0,1920	70,67%	93,73%	5,98%	0,0761	28,00%	0,3962	0,28%	0,0036	1,33%	0,0189	6,27%	0,0797	29,33%	0,4151	0,0669
Н	2229	1670	559	411	142	6	74,92%	25,08%	0,3347	18,44%	0,2461	73,52%	93,36%	6,37%	0,0850	25,40%	0,3455	0,27%	0,0036	1,07%	0,0146	6,64%	0,0886	26,48%	0,3601	0,0711
О	2748	1974	774	560	193	21	71,83%	28,17%	0,3921	20,38%	0,2837	72,35%	92,21%	7,02%	0,0978	24,94%	0,3446	0,76%	0,0106	2,71%	0,0375	7,79%	0,1084	27,65%	0,3821	0,0845
П	6287	4685	1602	1135	429	38	74,52%	25,48%	0,3419	18,05%	0,2423	70,85%	92,57%	6,82%	0,0916	26,78%	0,3780	0,60%	0,0081	2,37%	0,0335	7,43%	0,0997	29,15%	0,4115	0,0802
Р	2102	1444	658	458	187	13	68,70%	31,30%	0,4557	21,79%	0,3172	69,60%	90,49%	8,90%	0,1295	28,42%	0,4083	0,62%	0,0090	1,98%	0,0284	9,51%	0,1385	30,40%	0,4367	0,1052
С	3607	2691	916	624	258	34	74,60%	25,40%	0,3404	17,30%	0,2319	68,12%	91,90%	7,15%	0,0959	28,17%	0,4135	0,94%	0,0126	3,71%	0,0545	8,10%	0,1085	31,88%	0,4679	0,0881
Т	1272	936	336	212	113	11	73,58%	26,42%	0,3590	16,67%	0,2265	63,10%	90,25%	8,88%	0,1207	33,63%	0,5330	0,86%	0,0118	3,27%	0,0519	9,75%	0,1325	36,90%	0,5849	0,1080
У	1056	752	304	218	80	6	71,21%	28,79%	0,4043	20,64%	0,2899	71,71%	91,86%	7,58%	0,1064	26,32%	0,3670	0,57%	0,0080	1,97%	0,0275	8,14%	0,1144	28,29%	0,3945	0,0887
Ф	490	358	132	97	32	3	73,06%	26,94%	0,3687	19,80%	0,2709	73,48%	92,86%	6,53%	0,0894	24,24%	0,3299	0,61%	0,0084	2,27%	0,0309	7,14%	0,0978	26,52%	0,3608	0,0769
Х	441	321	120	82	29	9	72,79%	27,21%	0,3738	18,59%	0,2555	68,33%	91,38%	6,58%	0,0903	24,17%	0,3537	2,04%	0,0280	7,50%	0,1098	8,62%	0,1184	31,67%	0,4634	0,0943
Ц	196	142	54	38	16	0	72,45%	27,55%	0,3803	19,39%	0,2676	70,37%	91,84%	8,16%	0,1127	29,63%	0,4211	0,00%	0,0000	0,00%	0,0000	8,16%	0,1127	29,63%	0,4211	0,0889
Ч	488	374	114	77	32	5	76,64%	23,36%	0,3048	15,78%	0,2059	67,54%	92,42%	6,56%	0,0856	28,07%	0,4156	1,02%	0,0134	4,39%	0,0649	7,58%	0,0989	32,46%	0,4805	0,0820
Ш	509	391	118	94	23	1	76,82%	23,18%	0,3018	18,47%	0,2404	79,66%	95,28%	4,52%	0,0588	19,49%	0,2447	0,20%	0,0026	0,85%	0,0106	4,72%	0,0614	20,34%	0,2553	0,0495
Щ	72	46	26	17	8	1	63,89%	36,11%	0,5652	23,61%	0,3696	65,38%	87,50%	11,11%	0,1739	30,77%	0,4706	1,39%	0,0217	3,85%	0,0588	12,50%	0,1957	34,62%	0,5294	0,1429
Э	323	255	68	53	15	0	78,95%	21,05%	0,2667	16,41%	0,2078	77,94%	95,36%	4,64%	0,0588	22,06%	0,2830	0,00%	0,0000	0,00%	0,0000	4,64%	0,0588	22,06%	0,2830	0,0487
Ю	45	34	11	11	0	0	75,56%	24,44%	0,3235	24,44%	0,3235	100,00%	100,00%	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,0000
Я	135	90	45	30	14	1	66,67%	33,33%	0,5000	22,22%	0,3333	66,67%	88,89%	10,37%	0,1556	31,11%	0,4667	0,74%	0,0111	2,22%	0,0333	11,11%	0,1667	33,33%	0,5000	0,1250
Итого	36732	27367	9365	6690	2448	227	74,50%	25,50%	0,3422	18,21%	0,2445	71,44%	92,72%	6,66%	0,0895	26,14%	0,3659	0,62%	0,0083	2,42%	0,0339	7,28%	0,0977	28,56%	0,3999	0,0785

Таблица 2. Полная таблица результатов исследования по словарю Кузнецова

Буква	Слов	О	М	СП	СрП	СиП	% О	% М	М / О	% СП	СП / О	% СП (М)	% О+СП	% СрП	СрП / О	% СрП (М)	СрП / СП	% СиП	СиП / О	% СиП (М)	СиП / СП	% СрП+ СиП	СрП+ СиП / О	%СрП+ СиП (М)	СрП+ СиП / СП	СрП+ СиП / О+СП
А	884	593	291	224	67	0	67,08%	32,92%	0,4907	25,34%	0,3777	76,98%	92,42%	7,58%	0,1130	23,02%	0,2991	0,00%	0,0000	0,00%	0,0000	7,58%	0,1130	23,02%	0,2991	0,0820
Б	1623	1023	600	414	171	15	63,03%	36,97%	0,5865	25,51%	0,4047	69,00%	88,54%	10,54%	0,1672	28,50%	0,4130	0,92%	0,0147	2,50%	0,0362	11,46%	0,1818	31,00%	0,4493	0,1294
В	2654	1662	992	630	317	45	62,62%	37,38%	0,5969	23,74%	0,3791	63,51%	86,36%	11,94%	0,1907	31,96%	0,5032	1,70%	0,0271	4,54%	0,0714	13,64%	0,2178	36,49%	0,5746	0,1579
Г	1083	656	427	278	129	20	60,57%	39,43%	0,6509	25,67%	0,4238	65,11%	86,24%	11,91%	0,1966	30,21%	0,4640	1,85%	0,0305	4,68%	0,0719	13,76%	0,2271	34,89%	0,5360	0,1595
Д	1506	968	538	335	184	19	64,28%	35,72%	0,5558	22,24%	0,3461	62,27%	86,52%	12,22%	0,1901	34,20%	0,5493	1,26%	0,0196	3,53%	0,0567	13,48%	0,2097	37,73%	0,6060	0,1558
ЕЁ	136	83	53	31	19	3	61,03%	38,97%	0,6386	22,79%	0,3735	58,49%	83,82%	13,97%	0,2289	35,85%	0,6129	2,21%	0,0361	5,66%	0,0968	16,18%	0,2651	41,51%	0,7097	0,1930
Ж	281	166	115	76	34	5	59,07%	40,93%	0,6928	27,05%	0,4578	66,09%	86,12%	12,10%	0,2048	29,57%	0,4474	1,78%	0,0301	4,35%	0,0658	13,88%	0,2349	33,91%	0,5132	0,1612
З	1971	1190	781	490	265	26	60,38%	39,62%	0,6563	24,86%	0,4118	62,74%	85,24%	13,44%	0,2227	33,93%	0,5408	1,32%	0,0218	3,33%	0,0531	14,76%	0,2445	37,26%	0,5939	0,1732
И	1064	664	400	274	119	7	62,41%	37,59%	0,6024	25,75%	0,4127	68,50%	88,16%	11,18%	0,1792	29,75%	0,4343	0,66%	0,0105	1,75%	0,0255	11,84%	0,1898	31,50%	0,4599	0,1343
Й	6	4	2	2	0	0	66,67%	33,33%	0,5000	33,33%	0,5000	100,00%	100,00%	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,0000
К	2548	1527	1021	614	361	46	59,93%	40,07%	0,6686	24,10%	0,4021	60,14%	84,03%	14,17%	0,2364	35,36%	0,5879	1,81%	0,0301	4,51%	0,0749	15,97%	0,2665	39,86%	0,6629	0,1901
Л	841	509	332	200	120	12	60,52%	39,48%	0,6523	23,78%	0,3929	60,24%	84,30%	14,27%	0,2358	36,14%	0,6000	1,43%	0,0236	3,61%	0,0600	15,70%	0,2593	39,76%	0,6600	0,1862
М	1564	874	690	427	241	22	55,88%	44,12%	0,7895	27,30%	0,4886	61,88%	83,18%	15,41%	0,2757	34,93%	0,5644	1,41%	0,0252	3,19%	0,0515	16,82%	0,3009	38,12%	0,6159	0,2022
Н	2457	1474	983	636	313	34	59,99%	40,01%	0,6669	25,89%	0,4315	64,70%	85,88%	12,74%	0,2123	31,84%	0,4921	1,38%	0,0231	3,46%	0,0535	14,12%	0,2354	35,30%	0,5456	0,1645
О	2960	1685	1275	769	455	51	56,93%	43,07%	0,7567	25,98%	0,4564	60,31%	82,91%	15,37%	0,2700	35,69%	0,5917	1,72%	0,0303	4,00%	0,0663	17,09%	0,3003	39,69%	0,6580	0,2062
П	6903	3846	3057	1795	1106	156	55,71%	44,29%	0,7949	26,00%	0,4667	58,72%	81,72%	16,02%	0,2876	36,18%	0,6162	2,26%	0,0406	5,10%	0,0869	18,28%	0,3281	41,28%	0,7031	0,2237
Р	2299	1324	975	593	341	41	57,59%	42,41%	0,7364	25,79%	0,4479	60,82%	83,38%	14,83%	0,2576	34,97%	0,5750	1,78%	0,0310	4,21%	0,0691	16,62%	0,2885	39,18%	0,6442	0,1993
С	4060	2271	1789	1005	654	130	55,94%	44,06%	0,7878	24,75%	0,4425	56,18%	80,69%	16,11%	0,2880	36,56%	0,6507	3,20%	0,0572	7,27%	0,1294	19,31%	0,3452	43,82%	0,7801	0,2393
Т	1390	769	621	344	232	45	55,32%	44,68%	0,8075	24,75%	0,4473	55,39%	80,07%	16,69%	0,3017	37,36%	0,6744	3,24%	0,0585	7,25%	0,1308	19,93%	0,3602	44,61%	0,8052	0,2489
У	1160	587	573	318	220	35	50,60%	49,40%	0,9761	27,41%	0,5417	55,50%	78,02%	18,97%	0,3748	38,39%	0,6918	3,02%	0,0596	6,11%	0,1101	21,98%	0,4344	44,50%	0,8019	0,2818
Ф	574	339	235	146	82	7	59,06%	40,94%	0,6932	25,44%	0,4307	62,13%	84,49%	14,29%	0,2419	34,89%	0,5616	1,22%	0,0206	2,98%	0,0479	15,51%	0,2625	37,87%	0,6096	0,1835
Х	454	242	212	122	67	23	53,30%	46,70%	0,8760	26,87%	0,5041	57,55%	80,18%	14,76%	0,2769	31,60%	0,5492	5,07%	0,0950	10,85%	0,1885	19,82%	0,3719	42,45%	0,7377	0,2473
Ц	208	99	109	55	47	7	47,60%	52,40%	1,1010	26,44%	0,5556	50,46%	74,04%	22,60%	0,4747	43,12%	0,8545	3,37%	0,0707	6,42%	0,1273	25,96%	0,5455	49,54%	0,9818	0,3506
Ч	469	248	221	123	86	12	52,88%	47,12%	0,8911	26,23%	0,4960	55,66%	79,10%	18,34%	0,3468	38,91%	0,6992	2,56%	0,0484	5,43%	0,0976	20,90%	0,3952	44,34%	0,7967	0,2642
Ш	571	324	247	168	74	5	56,74%	43,26%	0,7623	29,42%	0,5185	68,02%	86,16%	12,96%	0,2284	29,96%	0,4405	0,88%	0,0154	2,02%	0,0298	13,84%	0,2438	31,98%	0,4702	0,1606
Щ	74	39	35	16	17	2	52,70%	47,30%	0,8974	21,62%	0,4103	45,71%	74,32%	22,97%	0,4359	48,57%	1,0625	2,70%	0,0513	5,71%	0,1250	25,68%	0,4872	54,29%	1,1875	0,3455
Э	410	261	149	104	42	3	63,66%	36,34%	0,5709	25,37%	0,3985	69,80%	89,02%	10,24%	0,1609	28,19%	0,4038	0,73%	0,0115	2,01%	0,0288	10,98%	0,1724	30,20%	0,4327	0,1233
Ю	54	33	21	13	8	0	61,11%	38,89%	0,6364	24,07%	0,3939	61,90%	85,19%	14,81%	0,2424	38,10%	0,6154	0,00%	0,0000	0,00%	0,0000	14,81%	0,2424	38,10%	0,6154	0,1739
Я	145	85	60	15	42	3	58,62%	41,38%	0,7059	10,34%	0,1765	25,00%	68,97%	28,97%	0,4941	70,00%	2,8000	2,07%	0,0353	5,00%	0,2000	31,03%	0,5294	75,00%	3,0000	0,4500
Итого	40349	23545	16804	10217	5813	774	58,35%	41,65%	0,7137	25,32%	0,4339	60,80%	83,67%	14,41%	0,2469	34,59%	56,90%	1,92%	0,0329	4,61%	0,0758	16,33%	0,2798	39,20%	0,6447	0,1951

Таблица 3. Полная таблица результатов исследования по МАС

Буква	Слов	О	М	СП	СрП	СиП	% О	% М	М / О	% СП	СП / О	% СП (М)	% О+СП	% СрП	СрП / О	% СрП (М)	СрП / СП	% СиП	СиП / О	% СиП (М)	СиП / СП	% СрП+ СиП	СрП+ СиП / О	%СрП+ СиП (М)	СрП+ СиП / СП	СрП+ СиП / О+СП
А	1600	1338	262	222	40	0	83,63%	16,38%	0,1958	13,88%	0,1659	84,73%	97,50%	2,50%	0,0299	15,27%	0,1802	0,00%	0,0000	0,00%	0,0000	2,50%	0,0299	15,27%	0,1802	0,0256
Б	3122	2577	545	409	128	8	82,54%	17,46%	0,2115	13,10%	0,1587	75,05%	95,64%	4,10%	0,0497	23,49%	0,3130	0,26%	0,0031	1,47%	0,0196	4,36%	0,0528	24,95%	0,3325	0,0455
В	6049	4586	1463	1095	337	31	75,81%	24,19%	0,3190	18,10%	0,2388	74,85%	93,92%	5,57%	0,0735	23,03%	0,3078	0,51%	0,0068	2,12%	0,0283	6,08%	0,0802	25,15%	0,3361	0,0648
Г	2274	1821	453	306	137	10	80,08%	19,92%	0,2488	13,46%	0,1680	67,55%	93,54%	6,02%	0,0752	30,24%	0,4477	0,44%	0,0055	2,21%	0,0327	6,46%	0,0807	32,45%	0,4804	0,0691
Д	3528	2838	690	525	149	16	80,44%	19,56%	0,2431	14,88%	0,1850	76,09%	95,32%	4,22%	0,0525	21,59%	0,2838	0,45%	0,0056	2,32%	0,0305	4,68%	0,0581	23,91%	0,3143	0,0491
ЕЁ	226	165	61	43	16	2	73,01%	26,99%	0,3697	19,03%	0,2606	70,49%	92,04%	7,08%	0,0970	26,23%	0,3721	0,88%	0,0121	3,28%	0,0465	7,96%	0,1091	29,51%	0,4186	0,0865
Ж	630	489	141	91	46	4	77,62%	22,38%	0,2883	14,44%	0,1861	64,54%	92,06%	7,30%	0,0941	32,62%	0,5055	0,63%	0,0082	2,84%	0,0440	7,94%	0,1022	35,46%	0,5495	0,0862
З	4828	3673	1155	846	284	25	76,08%	23,92%	0,3145	17,52%	0,2303	73,25%	93,60%	5,88%	0,0773	24,59%	0,3357	0,52%	0,0068	2,16%	0,0296	6,40%	0,0841	26,75%	0,3652	0,0684
И	2250	1654	596	455	131	10	73,51%	26,49%	0,3603	20,22%	0,2751	76,34%	93,73%	5,82%	0,0792	21,98%	0,2879	0,44%	0,0060	1,68%	0,0220	6,27%	0,0852	23,66%	0,3099	0,0669
Й	18	16	2	2	0	0	88,89%	11,11%	0,1250	11,11%	0,1250	100,00%	100,00%	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,00%	0,0000	0,0000
К	5056	3870	1186	800	350	36	76,54%	23,46%	0,3065	15,82%	0,2067	67,45%	92,37%	6,92%	0,0904	29,51%	0,4375	0,71%	0,0093	3,04%	0,0450	7,63%	0,0997	32,55%	0,4825	0,0827
Л	1768	1357	411	298	100	13	76,75%	23,25%	0,3029	16,86%	0,2196	72,51%	93,61%	5,66%	0,0737	24,33%	0,3356	0,74%	0,0096	3,16%	0,0436	6,39%	0,0833	27,49%	0,3792	0,0683
М	3473	2661	812	557	242	13	76,62%	23,38%	0,3051	16,04%	0,2093	68,60%	92,66%	6,97%	0,0909	29,80%	0,4345	0,37%	0,0049	1,60%	0,0233	7,34%	0,0958	31,40%	0,4578	0,0792
Н	5846	4386	1460	1047	379	34	75,03%	24,97%	0,3329	17,91%	0,2387	71,71%	92,94%	6,48%	0,0864	25,96%	0,3620	0,58%	0,0078	2,33%	0,0325	7,06%	0,0942	28,29%	0,3945	0,0760
О	6646	4794	1852	1347	459	46	72,13%	27,87%	0,3863	20,27%	0,2810	72,73%	92,40%	6,91%	0,0957	24,78%	0,3408	0,69%	0,0096	2,48%	0,0341	7,60%	0,1053	27,27%	0,3749	0,0822
П	16594	12415	4179	3199	890	90	74,82%	25,18%	0,3366	19,28%	0,2577	76,55%	94,09%	5,36%	0,0717	21,30%	0,2782	0,54%	0,0072	2,15%	0,0281	5,91%	0,0789	23,45%	0,3063	0,0628
Р	4725	3284	1441	980	418	43	69,50%	30,50%	0,4388	20,74%	0,2984	68,01%	90,24%	8,85%	0,1273	29,01%	0,4265	0,91%	0,0131	2,98%	0,0439	9,76%	0,1404	31,99%	0,4704	0,1081
С	8611	6317	2294	1481	701	112	73,36%	26,64%	0,3631	17,20%	0,2344	64,56%	90,56%	8,14%	0,1110	30,56%	0,4733	1,30%	0,0177	4,88%	0,0756	9,44%	0,1287	35,44%	0,5490	0,1043
Т	2969	2188	781	491	252	38	73,69%	26,31%	0,3569	16,54%	0,2244	62,87%	90,23%	8,49%	0,1152	32,27%	0,5132	1,28%	0,0174	4,87%	0,0774	9,77%	0,1325	37,13%	0,5906	0,1082
У	2582	1686	896	637	238	21	65,30%	34,70%	0,5314	24,67%	0,3778	71,09%	89,97%	9,22%	0,1412	26,56%	0,3736	0,81%	0,0125	2,34%	0,0330	10,03%	0,1536	28,91%	0,4066	0,1115
Ф	1347	1053	294	214	74	6	78,17%	21,83%	0,2792	15,89%	0,2032	72,79%	94,06%	5,49%	0,0703	25,17%	0,3458	0,45%	0,0057	2,04%	0,0280	5,94%	0,0760	27,21%	0,3738	0,0631
Х	1008	776	232	141	74	17	76,98%	23,02%	0,2990	13,99%	0,1817	60,78%	90,97%	7,34%	0,0954	31,90%	0,5248	1,69%	0,0219	7,33%	0,1206	9,03%	0,1173	39,22%	0,6454	0,0992
Ц	491	349	142	91	48	3	71,08%	28,92%	0,4069	18,53%	0,2607	64,08%	89,61%	9,78%	0,1375	33,80%	0,5275	0,61%	0,0086	2,11%	0,0330	10,39%	0,1461	35,92%	0,5604	0,1159
Ч	1134	861	273	174	87	12	75,93%	24,07%	0,3171	15,34%	0,2021	63,74%	91,27%	7,67%	0,1010	31,87%	0,5000	1,06%	0,0139	4,40%	0,0690	8,73%	0,1150	36,26%	0,5690	0,0957
Ш	1212	904	308	213	91	4	74,59%	25,41%	0,3407	17,57%	0,2356	69,16%	92,16%	7,51%	0,1007	29,55%	0,4272	0,33%	0,0044	1,30%	0,0188	7,84%	0,1051	30,84%	0,4460	0,0850
Щ	156	113	43	24	17	2	72,44%	27,56%	0,3805	15,38%	0,2124	55,81%	87,82%	10,90%	0,1504	39,53%	0,7083	1,28%	0,0177	4,65%	0,0833	12,18%	0,1681	44,19%	0,7917	0,1387
Э	881	686	195	148	43	4	77,87%	22,13%	0,2843	16,80%	0,2157	75,90%	94,67%	4,88%	0,0627	22,05%	0,2905	0,45%	0,0058	2,05%	0,0270	5,33%	0,0685	24,10%	0,3176	0,0564
Ю	74	50	24	18	6	0	67,57%	32,43%	0,4800	24,32%	0,3600	75,00%	91,89%	8,11%	0,1200	25,00%	0,3333	0,00%	0,0000	0,00%	0,0000	8,11%	0,1200	25,00%	0,3333	0,0882
Я	193	111	82	54	25	3	57,51%	42,49%	0,7387	27,98%	0,4865	65,85%	85,49%	12,95%	0,2252	30,49%	0,4630	1,55%	0,0270	3,66%	0,0556	14,51%	0,2523	34,15%	0,5185	0,1697
Итого	89291	67018	22273	15908	5762	603	75,06%	24,94%	0,3323	17,82%	0,2374	71,42%	92,87%	6,45%	0,0860	25,87%	36,22%	0,68%	0,0090	2,71%	0,0379	7,13%	0,0950	28,58%	0,4001	0,0768

Таблица 4. Полная таблица результатов исследования выборки (художественный стиль)

Текст	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
А	0,3339	0,6327	0,2991	0,2187	0,2519	0,1365	0,2549	0,2251	0,2626	0,1389	0,2889	0,6341	0,3903	0,3088	0,2223	0,1050	0,2094	0,7848	0,3498	0,3622	0,7884	1,3114	0,4024	0,1862	0,4478
Б	2,9529	2,7974	3,4378	3,1016	3,2866	2,4565	2,4773	3,7985	2,3232	2,9735	2,6004	2,1515	2,8436	4,5430	3,6953	3,2826	3,2050	4,7646	2,4484	3,0684	3,0274	4,5871	3,5298	3,0502	3,6542
В	9,1518	9,0880	9,1285	9,1625	7,3464	6,9260	6,9936	8,0191	6,2071	11,0185	8,9569	8,2400	7,3042	7,9637	8,4828	8,0095	7,6937	9,1368	6,7506	8,5233	9,4607	7,7232	8,3643	7,1467	7,4397
Г	1,4969	1,4169	1,5038	2,2360	2,5048	1,6035	2,1307	2,7856	1,3485	2,0981	1,1305	4,1642	1,8121	3,1593	2,3204	1,9170	2,9318	2,4103	2,3785	2,3226	2,2390	2,0982	2,5239	1,4364	2,6809
Д	3,3257	3,7946	3,9020	3,7087	2,6148	2,2177	2,2428	3,4328	2,7929	4,7311	4,1823	3,8424	3,0109	3,0821	4,9105	4,5168	4,0517	3,0269	2,7632	2,6848	2,9328	3,5770	3,6396	3,2453	4,5438
Е	2,2777	2,5734	2,1618	2,0952	0,9092	0,3412	0,7034	2,4479	0,8611	1,4728	1,9936	2,6815	1,5612	1,8766	1,8196	1,4443	1,2747	1,0090	0,6296	0,8097	1,3876	1,3895	1,4205	1,6936	1,5465
Ж	0,9126	0,4891	0,5611	0,7234	1,0106	0,6141	0,8869	0,5065	0,7955	1,0699	0,7007	0,7287	0,7527	0,8433	0,6432	1,1555	1,0015	0,5045	0,4547	0,5540	1,5453	1,2779	1,4022	0,7093	1,2181
З	2,7415	2,4487	2,4694	2,6351	2,2094	2,1153	2,2224	2,8700	1,9066	2,4038	2,0767	2,5458	2,6206	2,3873	3,0694	2,4947	3,9242	2,1861	3,9175	1,9817	2,4913	2,7846	2,4569	2,7487	2,6929
И	1,7547	1,6393	1,8940	1,9958	1,5955	1,7059	1,1826	1,3787	0,9444	1,3756	1,5133	1,1956	1,2824	1,3184	1,5627	1,9170	1,4386	2,0740	0,9094	1,4277	1,4822	1,6127	1,3717	1,2059	0,9673
Й	0,0000	0,0079	0,0028	0,0000	0,0000	0,0341	0,0000	0,0000	0,0000	0,0000	0,0000	0,3218	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
К	4,2847	5,1420	5,1923	4,8573	3,4024	2,5589	3,6803	3,5453	2,9773	3,5084	3,9151	4,2998	4,9345	4,5193	4,9084	5,0945	5,5358	5,7735	5,6663	3,6863	4,1943	5,1172	3,7188	4,8324	6,6515
Л	1,4579	1,1013	1,4212	1,4101	1,8330	1,6718	2,0186	3,4609	1,0682	1,6674	1,8456	2,0568	1,4218	1,6866	1,6340	1,3655	1,3384	1,9058	2,5883	1,7686	2,0183	2,6786	1,3778	1,0640	1,7196
М	4,9802	3,0278	5,0271	3,5497	3,2779	4,4695	3,0584	3,8548	2,9975	5,1202	4,9516	4,5396	3,9866	3,1356	3,0931	4,4905	2,9864	3,6996	3,0080	4,2190	3,0274	3,9007	4,7796	3,8659	4,4901
Н	9,2241	10,0994	9,7009	9,3934	5,9767	6,6189	5,3624	5,8807	5,5354	7,8991	7,3823	9,5366	7,5829	7,9340	9,0872	8,0095	8,0488	8,6883	4,2322	5,8172	5,8657	8,9453	7,6449	7,5279	9,3026
О	7,1077	7,5387	6,7018	7,3884	3,2606	2,5930	3,2215	6,0214	3,2601	5,5996	4,2184	4,4891	4,9624	4,6083	4,9818	5,8561	4,2520	4,5964	2,3435	3,6011	4,1943	3,6551	5,0357	4,1231	2,9615
П	8,2466	9,4841	8,8579	9,1855	5,8030	5,7318	6,3615	6,2746	7,9192	7,6490	6,6238	6,7195	7,8617	9,0504	8,4763	9,1124	10,3524	8,2399	8,7793	7,4579	6,9063	7,2768	7,3218	6,8984	6,1918
Р	1,6879	2,2121	2,0934	2,6275	1,9083	1,7059	1,6312	2,2228	1,3207	1,5632	1,8853	2,3597	2,2582	2,5952	2,6031	2,6261	3,8241	2,4664	3,1829	2,5783	2,2390	1,6797	2,4081	4,5487	1,5465
С	7,5084	10,3108	8,7696	8,5477	7,3898	8,1201	6,6673	10,7485	6,6490	8,5591	9,7190	9,5650	9,2835	10,3866	8,1871	8,3771	8,8774	7,7915	9,0591	7,2874	8,1047	7,2935	8,2180	7,4304	6,7112
Т	3,3331	4,7081	3,7169	4,0360	3,9903	3,0365	3,7619	4,7833	3,5581	3,5501	2,8677	3,8802	5,4921	3,8245	5,0228	5,0158	4,3886	7,9596	2,7632	4,4108	2,9013	6,0268	5,3893	5,6127	6,4724
У	2,3371	1,9344	2,1276	2,1824	1,3060	1,1941	1,4069	1,4350	1,0833	1,7855	1,7083	1,7035	1,7285	1,5025	1,9750	2,0746	1,9120	1,5135	1,0843	1,6407	1,7660	1,8527	1,8533	1,6315	2,1495
Ф	0,2003	0,4339	0,3019	0,3456	0,1564	0,0682	0,2855	0,0000	0,1717	0,0486	0,1517	0,3659	0,2230	0,0950	0,1899	0,2363	0,7375	0,2242	0,0700	0,4688	0,5361	0,0725	0,1707	0,1951	0,1552
Х	0,6288	1,0871	0,9200	0,6622	0,6660	0,4094	0,3772	0,4221	0,6995	0,6600	0,8957	1,8928	0,5576	0,7601	1,1202	0,8929	1,9303	0,6166	0,6995	0,5966	0,6623	1,2835	1,2437	0,6561	1,1345
Ц	0,0853	0,1104	0,1481	0,3135	0,4025	0,1365	0,3466	0,1407	0,1843	0,0903	0,1084	0,5458	0,2788	0,5523	0,0799	0,1576	0,1821	0,4484	0,3498	0,3409	0,2208	0,0949	0,1829	0,2039	0,1433
Ч	3,4444	3,0467	3,1843	2,8294	1,7780	1,8424	1,6108	2,0259	1,5278	2,5149	3,7923	2,9591	2,2582	1,7816	2,2340	3,3613	1,9940	2,6345	0,7345	1,8112	1,5453	2,7679	2,5910	2,3674	3,5407
Ш	0,3098	0,2808	0,3987	0,6270	0,3764	0,2388	0,3772	0,3095	0,3535	0,2015	0,2420	0,3849	0,5576	0,4276	0,6756	0,6303	0,7830	0,7848	0,5247	0,7671	0,5676	0,3404	0,3719	0,4167	0,3463
Щ	0,0260	0,0426	0,0399	0,0627	0,0492	0,0341	0,0306	0,1688	0,0429	0,0486	0,0361	0,0221	0,0558	0,0475	0,0691	0,0263	0,0910	0,2242	0,1049	0,1278	0,0631	0,0223	0,0366	0,0266	0,0179
Э	0,2096	0,0805	0,1481	0,1805	0,0666	0,0000	0,0612	0,0844	0,1035	0,5627	0,1445	0,7066	0,2230	0,0653	0,1662	0,4202	0,1457	0,1682	0,0000	0,1278	0,1261	0,4967	0,2134	0,1773	0,2329
Ю	0,0204	0,0221	0,0256	0,1071	0,1071	0,0000	0,0714	0,1125	0,0480	0,0347	0,0433	0,0126	0,1115	0,0534	0,0130	0,0000	0,0091	0,0561	0,0000	0,0426	0,0946	0,0000	0,0610	0,0355	0,0000
Я	1,9958	1,2354	2,5776	0,7280	0,6139	1,3306	0,9175	0,8441	1,1869	2,5705	2,1526	1,9559	1,6170	0,5226	0,7749	2,2059	0,7648	1,3453	1,0493	1,6407	0,8199	2,1652	1,6338	1,9152	2,0002

Таблица 5. Полная таблица результатов исследования выборки (публицистический стиль)

Текст	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
А	0,3112	1,9716	1,0218	0,4119	0,7966	0,7979	1,0613	0,5687	1,5356	1,0209	0,7565	1,3431	0,6609	0,7135	0,2191	0,3989	0,7946	0,3984	1,0676	0,7353	0,2137	0,4149	0,8929	0,6393	0,7347
Б	2,9046	2,2082	2,2212	3,3773	3,1863	3,5904	2,7123	2,9384	3,3039	1,9398	3,8514	3,1746	2,8420	4,5005	2,1358	3,5461	3,4052	2,1248	2,6097	2,8361	3,5256	2,8216	2,4107	2,2376	1,2245
В	9,0249	7,7287	8,4407	7,9077	9,4363	8,2447	6,6038	9,9526	7,7711	7,8101	9,5598	10,0122	8,6583	9,6597	12,6506	9,6631	9,8751	9,0305	9,3713	6,6176	7,6923	7,2199	10,9821	6,6063	7,7551
Г	2,1784	1,8139	2,2212	2,5947	2,9412	2,1277	3,3019	1,8009	2,2336	2,2971	2,2008	1,7094	2,3794	1,7563	2,7382	2,3936	2,2701	2,7888	2,4911	1,3655	2,2436	2,4896	1,1607	1,2786	3,6735
Д	5,7054	2,2082	3,0209	3,5008	3,3088	3,5904	3,1840	3,5071	3,3969	3,5222	2,8198	2,1978	3,1725	2,9089	3,6692	3,3688	3,7457	3,8513	3,0842	4,6218	2,0299	2,4066	2,6786	2,8237	2,8571
Е	1,4523	1,8927	1,9547	1,4415	1,3480	1,4628	1,8868	1,1374	1,5821	1,4293	1,8226	1,5873	0,9253	0,9879	1,6429	1,1968	2,3837	0,6640	2,0166	0,7353	0,7479	1,5768	0,8929	2,4507	1,4694
Ж	1,5560	0,4732	0,1333	0,3913	0,7353	0,3989	1,2972	2,0853	0,6980	1,0209	0,5502	0,2442	0,5288	1,0977	0,1095	0,4433	0,6810	0,3984	0,1186	0,9454	0,7479	0,5809	0,0893	0,1598	0,4898
З	1,7635	3,0757	2,4434	2,8213	2,2672	2,3936	2,9481	2,0853	3,3504	2,1950	3,1293	2,5641	3,7673	2,9638	2,0811	2,3050	2,9512	1,7264	1,3049	2,6261	2,0299	2,8216	1,9643	2,2909	2,0408
И	1,5560	2,0505	2,0880	1,7710	2,7574	3,1915	2,7123	2,3697	2,7455	1,9908	2,4072	2,6862	1,5863	2,0856	2,7930	2,4379	2,0431	2,5232	2,0166	1,2605	3,3120	2,5726	1,6964	2,6105	2,0408
Й	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
К	4,4606	4,4953	4,6202	4,5717	5,3309	6,6489	2,8302	4,6445	4,1415	6,1256	5,7772	3,6630	4,6927	4,5554	4,7097	4,6986	5,4484	3,8513	5,4567	4,8319	5,1282	5,6432	4,2857	4,0490	6,2041
Л	2,2822	0,8675	0,8885	0,8443	1,7157	0,5319	0,9434	0,8531	1,3495	1,3783	1,4099	2,0757	1,1897	2,1405	1,3691	0,9309	2,1566	2,2576	0,9490	2,2059	1,9231	1,4938	1,2500	0,8524	0,8980
М	3,8382	2,2871	3,8649	3,9333	2,8186	3,7234	6,9575	4,4550	3,2108	3,6243	3,6107	2,1978	5,4197	1,7563	2,9573	4,1667	2,0431	4,1169	2,7284	3,5714	3,4188	2,0747	2,4107	2,2376	2,3673
Н	7,2614	5,5205	7,2857	7,8254	9,0074	7,8457	7,1934	7,7725	7,8176	9,8009	8,7689	6,8376	8,5261	8,6718	6,6265	10,0621	11,0102	10,6242	9,3713	10,7143	8,9744	8,6307	5,0000	8,3111	7,6735
О	4,8755	4,4164	4,3092	4,6540	5,0858	5,9840	4,9528	4,4550	6,1889	4,9005	6,8776	4,6398	4,6927	5,2141	2,1906	5,7181	3,6322	4,1169	5,1008	6,5126	5,5556	3,9004	5,1786	7,3522	4,3265
П	10,6846	9,0694	8,9294	9,0815	9,9265	8,5106	8,7264	6,5403	10,3769	10,7708	11,4512	10,1343	8,9888	7,1899	9,5838	10,4610	11,1237	8,3665	7,9478	10,0840	9,0812	9,4606	10,4464	9,8029	9,5510
Р	2,5934	1,8139	2,8876	2,8830	3,2475	3,9894	2,8302	2,7488	3,2108	2,3481	2,5103	2,3199	2,7759	3,0187	2,1358	2,8369	3,0647	3,4529	3,3215	2,7311	3,3120	3,0705	3,7500	2,0245	3,5102
С	7,2614	7,1767	8,5740	6,5280	7,9044	9,7074	5,3066	6,7299	10,3769	9,1373	9,0096	7,2039	7,7330	8,0681	6,6265	7,8014	7,8320	10,8898	6,9988	7,4580	8,4402	7,7178	8,2143	8,6841	8,8980
Т	4,5643	2,4448	2,8432	4,2216	4,2892	2,3936	3,7736	3,4123	3,5831	4,2879	3,9202	3,1746	3,1725	4,8847	2,9025	5,1862	4,9943	7,5697	4,6263	3,4664	3,2051	4,6473	2,6786	3,6761	4,4082
У	1,5560	1,7350	1,7326	1,9563	1,5319	2,6596	2,3585	1,8957	1,7683	1,4803	1,8226	1,9536	1,9167	1,5368	1,5882	1,8174	1,8161	2,3904	1,7794	0,8403	2,1368	2,5726	0,8929	1,7048	1,6327
Ф	0,5187	0,6309	0,9773	0,2471	0,7353	0,9309	0,3538	0,5687	0,8376	1,4293	0,1376	0,1221	0,1322	0,3293	0,9858	0,7979	0,3405	0,5312	0,1186	0,7353	0,6410	0,5809	0,2679	0,3197	0,4898
Х	0,4149	0,6309	0,3110	0,6178	0,7353	1,1968	1,6509	1,0427	0,5119	1,1230	0,2063	0,2442	1,3880	0,7684	0,3286	0,6649	1,3621	0,5312	0,3559	0,4202	0,7479	0,8299	0,3571	1,0655	0,1633
Ц	0,6224	0,3943	0,3554	0,3913	0,7353	0,1330	0,0000	0,1896	0,1396	0,6126	0,1376	0,1221	0,5948	0,1098	0,1095	0,2216	0,3405	0,5312	0,1186	0,2101	0,3205	1,3278	0,2679	0,4262	0,1633
Ч	2,8008	1,9716	2,1768	2,6359	2,3897	2,3936	2,9481	1,8009	2,7920	2,1950	1,7538	0,7326	1,9828	2,6345	2,9573	3,1472	3,6322	2,9216	4,0332	2,1008	2,3504	1,4108	2,2321	4,0490	1,4694
Ш	0,0000	0,1577	0,2221	0,2059	0,1838	0,5319	0,8255	0,1896	0,1396	0,5105	0,2751	0,0000	0,2644	0,7684	0,0000	0,3546	0,4540	0,1328	0,1186	0,9454	0,2137	0,4149	0,0000	0,1066	0,0816
Щ	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0465	0,0000	0,0000	0,0000	0,0000	0,3293	0,0000	0,0000	0,0000	0,1328	0,0000	0,0000	0,0000	0,0830	0,0000	0,0000	0,0000
Э	0,2075	0,5521	0,6219	0,3089	0,4289	1,1968	1,1792	0,1896	0,0931	0,7147	0,0688	0,2442	0,3966	1,3172	0,0000	0,3103	0,0000	0,5312	0,1186	0,4202	0,4274	0,4979	0,4464	0,2131	0,4898
Ю	0,0000	0,0789	0,0000	0,0000	0,0613	0,1330	0,0000	0,0948	0,0931	0,1021	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,1068	0,1660	0,0000	0,2664	0,0816
Я	0,3112	0,3155	0,1777	2,4300	0,3676	0,3989	1,7689	1,8009	0,3257	0,6636	0,4127	0,1221	0,9253	0,7684	0,6572	0,7092	0,3405	0,5312	1,0676	0,3151	0,7479	0,6639	0,2679	0,0533	0,5714

Таблица 6. Полная таблица результатов исследования выборки (научный стиль)

Текст	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
А	1,8701	0,9091	1,8111	1,3405	1,3867	2,3969	1,2548	1,4166	2,3553	1,2195	0,7139	2,5304	2,0331	1,3840	0,8663	1,1440	1,3639	1,1037	0,1691	0,4010	1,3274	0,6999	0,6233	0,8996	0,6116
Б	2,6575	1,5152	2,8461	1,3405	1,3867	1,5317	1,4852	1,3517	1,4907	1,8722	1,2857	4,4534	4,3225	1,2539	1,5607	2,5635	1,9427	2,3545	2,0287	1,4435	1,5929	1,7498	1,3158	0,7496	2,2936
В	10,3346	8,9394	7,6326	8,4450	11,8644	7,9003	6,8079	7,2220	9,3267	8,5243	8,8682	10,6275	9,0047	8,7437	8,3826	11,7730	7,3757	9,3676	7,5232	8,3400	10,2655	8,3115	9,1413	7,6462	11,4679
Г	0,7874	0,6061	1,0349	1,4745	2,1572	2,6119	0,9292	2,2443	1,5614	1,6025	1,6748	2,2267	1,7279	0,5073	0,6839	1,9441	1,2784	2,9809	1,4370	1,3633	1,7699	4,8119	1,1080	1,4993	0,6116
Д	1,9685	4,3182	4,0103	4,1555	5,2388	2,0207	4,2474	2,8019	2,5871	2,2491	3,8437	0,7085	3,2269	3,5043	3,2173	2,6954	2,9354	2,4310	2,7050	2,0850	3,2743	2,2747	3,4626	4,6477	3,9755
Е	0,6890	0,4545	0,0000	1,3405	0,6163	0,6449	0,6878	0,9946	0,7283	0,3340	0,8391	0,4049	0,4415	0,7050	1,3494	0,7750	0,9550	0,5881	0,5917	1,4435	1,5044	1,3123	0,9695	1,1994	1,6820
Ж	0,0000	0,3788	0,2587	0,2681	0,0000	0,7524	0,1096	0,2388	0,2698	0,2819	0,3248	0,1012	0,7849	0,2133	0,3413	0,4964	0,2919	0,1430	0,0845	0,2406	0,0885	0,1750	0,3463	0,2999	0,1529
З	1,8701	4,0909	3,7516	1,7426	1,2327	0,8868	3,6518	1,8455	2,1732	1,2256	1,6545	1,6194	1,4826	1,8263	2,1908	2,4450	1,6294	1,7744	3,2967	2,2454	1,5929	1,3123	2,9086	2,0990	2,1407
И	4,0354	3,7121	3,1048	3,2172	4,0062	2,1229	5,4197	3,5635	3,3691	3,1468	3,0993	2,0243	3,3740	3,7566	3,7673	3,3756	3,4312	3,0836	4,1420	3,6889	4,0708	2,3622	3,6704	2,0990	1,6820
Й	0,0000	0,0000	0,0000	0,0000	0,0000	0,0752	0,0206	0,0174	0,0039	0,3922	0,0000	0,0000	0,0327	0,0000	0,0000	0,0015	0,0000	0,0060	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
К	4,1339	3,1818	5,1746	5,2279	5,8552	6,4761	5,5817	4,0330	5,4703	4,2744	6,4964	2,9352	4,2625	4,0194	4,2149	6,0547	4,7158	3,8348	4,8183	5,7739	4,3363	3,7620	5,7479	4,7976	5,8104
Л	1,5748	1,2121	2,3286	1,3405	0,6163	1,2952	1,5995	0,7767	0,6864	0,9254	1,5970	0,7085	1,8315	1,3580	0,6629	1,0284	0,8657	0,7190	0,1691	0,4010	0,8850	0,7874	0,9003	0,7496	1,2232
М	2,6575	2,8788	2,0699	3,8874	2,6194	2,5098	2,8639	3,4962	3,8158	4,6084	3,4749	3,2389	3,3250	2,1333	3,8027	3,5223	4,6278	4,7311	3,2967	2,8067	3,1858	2,0122	2,9778	1,0495	2,5994
Н	6,0039	6,6667	4,2691	6,0322	3,2357	3,3858	5,5388	5,2965	4,5599	4,5502	4,6794	7,5911	6,0940	5,0131	6,3375	5,1611	5,3424	5,5408	7,6078	8,6608	6,9912	5,5993	6,2327	5,6972	5,6575
О	7,3819	5,5303	6,2096	8,7131	10,0154	5,3367	7,5894	5,5632	7,2806	5,9811	8,8919	9,8178	4,9166	7,5991	7,1277	5,8280	6,2219	4,9970	7,6078	6,9767	8,7611	7,3491	8,5873	13,7931	6,7278
П	13,4843	13,1061	13,5834	12,1984	11,2481	7,7820	12,3515	10,6174	10,2214	10,8285	10,1336	11,8421	13,1309	14,5165	10,4211	11,4203	9,2341	9,2709	11,4117	17,5621	10,1770	10,0612	13,6427	16,6417	9,9388
Р	2,6575	3,7879	6,5977	4,2895	2,7735	2,5797	4,1187	3,4406	4,7341	4,4000	3,0486	4,4534	3,9900	3,0386	2,2236	5,6190	2,6347	4,6546	6,6779	3,5285	4,4248	3,6745	5,5402	4,0480	4,4343
С	12,1063	11,8182	15,3946	8,8472	11,2481	7,6745	9,9625	8,1169	8,5368	7,3201	10,9355	8,7045	7,6801	11,9644	9,2830	8,4626	7,8551	8,9507	9,3829	10,4250	12,3009	11,2861	11,0111	9,4453	13,3028
Т	4,0354	3,8636	1,9405	8,0429	2,1572	2,4776	4,6525	3,7849	3,6691	2,5064	2,9538	3,9474	4,3933	2,6171	4,3107	4,0705	3,1946	4,0544	4,9873	4,6512	6,0177	5,5118	4,2936	2,9985	3,6697
У	1,2795	4,9242	3,3635	2,2788	5,2388	0,7793	2,4604	1,4224	2,0382	1,5137	2,3752	5,8704	1,3572	1,9251	1,4190	2,3457	1,5766	2,3605	3,8039	3,7690	1,5929	2,3622	2,9778	5,9970	2,4465
Ф	1,8701	2,6515	4,6572	1,0724	2,6194	1,4833	0,9943	1,1152	1,2588	0,7813	1,1402	2,1255	1,3409	1,1135	1,6316	1,0758	0,7411	0,7472	0,3381	1,2831	1,0619	0,5249	0,6233	1,6492	5,5046
Х	0,5906	0,6061	0,0000	0,1340	0,0000	1,2522	1,0325	1,4676	0,6379	0,8089	0,4838	0,7085	0,4088	0,5515	0,3610	0,5246	0,4026	0,5881	0,3381	0,5613	0,3540	1,4873	0,0693	0,0000	0,3058
Ц	0,1969	0,5303	0,2587	0,9383	0,6163	0,6610	0,4797	0,2886	0,1493	0,2758	0,7511	0,1012	0,4143	0,3096	0,2665	0,4327	0,3611	0,4693	0,5917	1,6038	0,0885	0,5249	0,5540	0,2999	1,0703
Ч	2,5591	1,6667	0,6468	1,0724	1,2327	0,9781	1,1961	1,7736	3,3298	1,0357	1,7560	0,4049	1,6080	1,3190	3,1714	1,2729	2,0698	1,3635	0,9298	1,2831	1,4159	3,0621	1,4543	0,7496	3,0581
Ш	0,8858	1,2121	0,1294	0,1340	0,0000	0,1397	0,3590	0,1403	0,1729	0,0552	0,2368	0,0000	0,1145	0,0728	0,1260	0,2741	0,2642	0,2014	0,5072	0,2406	0,0000	1,7498	1,7313	0,0000	0,1529
Щ	0,0000	0,0000	0,0000	0,0000	0,0000	0,1720	0,0286	0,0394	0,0118	0,0827	0,0000	0,0000	0,0273	0,0026	0,0013	0,0445	0,0025	0,0222	0,0000	0,0000	0,0000	0,0875	0,0000	0,0000	0,0000
Э	0,4921	0,3788	0,2587	1,4745	0,6163	1,1394	1,1039	0,7848	0,8829	1,4126	0,6158	0,6073	0,7958	2,4636	0,3833	0,8135	0,5033	3,8268	2,1978	1,0425	0,6195	0,2625	0,6925	0,0000	0,3058
Ю	0,0000	0,0000	0,0000	1,2064	0,0000	0,0161	0,0286	0,0081	0,0026	0,0153	0,0271	0,0000	0,0055	0,5723	0,0118	0,0089	0,0063	0,0483	0,0000	0,0000	0,0000	0,0000	0,0693	0,0000	0,0000
Я	0,7874	0,2273	0,6468	0,4021	0,6163	0,6825	0,4257	0,5008	0,5515	0,2329	0,4669	0,6073	0,4252	0,4605	0,7718	0,2223	0,6794	0,4109	0,5917	0,4812	1,7699	0,6999	0,4155	0,7496	0,1529

Таблица 7. Полная таблица результатов исследования выборки (официально-деловой стиль)

Текст	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
А	3,3682	2,3377	3,0957	0,2933	1,6873	0,4926	0,4722	0,4091	0,1610	4,2863	0,3322	0,7456	1,5172	1,0601	0,9254	1,2433	9,8031	1,3672	0,9957	0,6027	0,6273	0,5920	0,5798	0,8933	0,5914
Б	1,5002	1,7272	1,7824	2,2287	0,8041	2,4015	1,0088	0,8183	2,4155	1,6516	3,7209	0,6980	0,1339	0,4836	0,3599	2,4688	0,8840	0,7465	0,9293	0,7367	1,3243	1,3048	0,8610	1,5299	1,2382
В	10,9127	7,6980	6,5666	8,0645	7,4347	9,9138	10,3456	5,7791	8,9372	6,2918	6,3123	7,5825	4,2392	3,1693	11,1568	7,1618	6,9322	7,5491	8,5629	6,9917	5,2100	5,8234	7,0692	11,2948	9,0002
Г	3,6053	2,8142	4,8780	1,7009	2,6496	1,7241	2,6186	0,5114	1,7713	1,2977	0,7973	1,5387	1,4726	1,0118	2,4165	1,9567	2,9847	2,4912	0,7302	6,3622	1,0106	3,6245	1,2827	2,3103	1,6633
Д	3,0875	3,8565	4,2214	3,0792	3,3549	6,5271	3,7562	3,0003	5,4750	3,2245	4,3189	4,1402	3,1682	2,8457	3,7275	3,7554	2,4595	3,3551	4,5470	3,6432	10,2980	2,9721	4,3516	3,1112	2,7352
Е	0,6291	1,1167	0,5629	0,9091	0,5207	0,5952	1,0732	0,3921	0,0805	1,2190	0,7309	0,8407	0,1339	0,2381	0,7712	0,6166	0,5864	0,6794	0,7302	0,5626	1,3940	0,4712	1,5404	0,4621	0,5544
Ж	0,1355	0,1042	0,1876	0,5865	0,1582	0,1642	0,7298	0,0000	0,0000	0,0786	0,0664	0,1110	0,0000	0,0298	0,0000	0,2038	0,1663	0,1510	0,0000	0,0402	0,2439	0,2779	0,3690	0,0411	0,4620
З	2,0664	2,1292	2,8143	2,8739	4,5874	4,9261	2,3825	1,4490	3,0596	3,3032	3,1894	7,2176	2,0080	2,3844	2,6221	1,7096	3,5624	3,7410	2,4228	4,0986	3,9554	2,6821	8,7150	3,5733	2,6982
И	3,3440	5,1221	1,3133	3,3138	4,3897	2,2167	3,9493	2,1309	2,4960	1,9269	6,8439	3,2202	2,8559	2,5555	4,0617	2,6955	3,8775	4,2107	5,0448	3,0271	3,6940	4,0111	7,7018	5,5755	1,6818
Й	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
К	3,7311	3,8267	5,6285	3,7830	2,0960	3,4278	3,0908	2,9662	5,6361	10,2635	4,5183	1,8718	3,2575	2,7824	2,5193	5,4522	1,4880	1,5685	6,6379	2,6118	1,9167	3,1412	3,4438	2,4232	5,4703
Л	0,9824	0,5211	1,7824	1,7009	8,1136	0,5542	1,8244	0,2557	0,7246	0,1180	1,1296	1,1897	0,0000	1,8934	0,3856	0,7389	0,3501	0,4529	1,1948	1,2323	0,9758	2,0176	1,1772	1,5197	0,9056
М	1,2631	3,2162	1,4071	2,5513	1,4830	4,8235	1,5239	1,0569	0,9662	1,1011	2,3256	2,6808	2,1865	1,4470	1,1568	1,0981	1,9344	1,7363	1,0621	1,3796	1,3243	1,5706	1,1304	6,3148	1,8850
Н	3,9586	5,5688	5,5347	8,0645	5,8199	6,0961	4,7864	5,8814	4,7504	5,3480	7,3754	5,4569	2,9451	4,7911	3,6504	4,5146	4,3589	3,9255	3,7504	4,0048	4,7395	6,2462	7,3738	5,5447	6,2835
О	7,2784	7,8916	4,6904	7,1554	8,4959	6,7529	8,0489	8,0123	9,2593	15,0216	8,5050	5,7265	8,7461	3,2511	8,3548	8,2369	16,4026	14,9052	5,7750	10,0857	11,0995	12,1783	8,3870	6,0068	8,0207
П	12,5387	17,2275	8,5366	11,2903	9,5044	10,7964	11,8266	11,0297	10,8696	12,8982	11,2957	20,0666	15,4842	6,5990	11,2853	9,2841	7,8600	8,6982	14,6034	8,4784	15,7867	11,2480	11,8777	9,4260	10,9961
Р	2,4051	1,2210	1,6886	2,4047	3,6646	2,9762	3,1981	2,0798	3,4622	1,7696	4,1196	4,8065	3,5252	4,7837	3,3676	2,9019	2,3545	2,3822	1,2944	3,3351	1,4288	2,3680	2,3310	2,8340	1,9959
С	10,6514	12,3437	11,4447	10,5865	8,1004	8,3128	12,7066	7,2622	8,0515	6,6457	5,5150	5,9803	7,3628	6,1079	8,0720	8,3949	7,5361	7,8259	14,4042	11,2108	11,1866	8,3364	9,1836	8,1425	8,5751
Т	2,3035	3,1864	2,1576	2,4047	1,5357	3,0788	1,7600	1,3979	5,2335	2,1628	1,3953	2,4270	0,6247	1,4396	1,8766	1,8471	1,6368	1,7698	0,9625	4,2593	2,6311	2,6217	2,4365	3,4911	3,0309
У	2,6520	3,3204	2,7205	2,3460	3,7899	3,6330	3,3913	2,2503	2,9791	3,7751	5,3156	3,3788	4,5515	1,6516	2,9563	2,1936	2,0481	1,8537	5,8745	2,4243	2,6485	2,2230	3,1393	1,9407	3,6962
Ф	0,8130	0,7296	0,1876	0,1466	1,2128	1,7857	0,3649	2,2844	1,4493	1,5729	3,0565	0,5235	2,2311	1,0453	2,0566	1,8064	1,4267	1,2162	1,3940	0,5759	1,2197	0,5678	0,5271	1,3040	1,2197
Х	0,0581	0,2829	1,6886	0,2053	0,2768	1,0878	0,1717	0,0852	0,2415	0,2359	0,2658	0,6662	0,0000	0,0744	0,0771	0,1758	0,1663	0,0923	0,3651	0,0938	0,1917	1,0874	0,0703	0,2054	0,0554
Ц	0,2274	0,1638	0,0938	0,4692	0,5471	0,1232	0,1502	0,2216	0,4831	0,1180	0,0664	0,0000	1,3387	0,1637	0,4113	1,0828	0,3501	0,1174	4,7129	0,1473	0,8015	0,3141	0,2753	0,3388	0,1294
Ч	0,5323	1,5634	0,3752	1,4956	1,0018	1,8883	1,4810	0,2046	0,5636	0,6292	0,1329	0,3807	0,2231	0,5691	0,7712	0,7134	0,7265	0,7214	0,1659	0,5090	0,9584	0,5195	0,6091	0,6880	4,1952
Ш	0,0290	0,1042	0,2814	0,0587	0,0264	0,4721	0,0000	0,0511	0,0000	0,0786	0,0000	0,1428	0,0000	0,1376	0,2314	0,1656	0,0088	0,0168	0,0332	0,0938	0,0174	0,0362	0,0117	0,1746	0,0554
Щ	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0446	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Э	0,8711	0,1936	0,0938	0,5865	0,4350	0,5337	0,4722	0,1705	0,3221	0,1573	0,0664	0,5235	0,0446	0,6287	0,4884	0,3134	0,3151	0,1090	3,5513	0,2545	0,0871	0,4108	0,0000	0,5545	0,1109
Ю	0,1645	0,1489	0,0000	0,0587	0,2768	0,0000	0,0215	0,0682	0,6441	0,0000	0,2658	0,1110	0,4909	0,8072	0,0514	0,2650	0,0963	0,1007	0,2323	0,0000	0,1917	0,5678	0,1054	0,1848	0,0000
Я	0,3000	0,4318	0,0938	0,6158	0,2834	0,3284	0,2576	0,0000	0,4026	0,2359	0,3987	0,2697	0,0892	0,0409	0,1285	0,3159	0,2363	0,2684	0,0996	0,2545	0,3311	0,2658	0,5564	0,6161	0,1848

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 3
1. ФОРМАЛЬНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ МНОГОЗНАЧНОЙ ЛЕКСИКИ И АКТУАЛЬНЫЕ ПРОБЛЕМЫ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ОБРАБОТКИ ТЕКСТОВ (АСОТ) 9
1.1. Структурно-семантические аспекты лексического значения слова 9
1.2. Лексическая полисемия и ее разновидности 20
1.3. Современные продукты и решения в сфере анализа текстов 26
1.4. Формально-семантический анализ в АСОТ 40
1.5. Лингвистическое обеспечение АСОТ 46
1.6. Выводы 63
2. ВОЗМОЖНОСТИ ОПТИМИЗАЦИИ АСОТ ПОСРЕДСТВОМ ФУНКЦИОНАЛЬНО-СТИЛИСТИЧЕСКОЙ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ 65
2.1. Функциональный стиль как фактор оптимизации автоматизированного анализа текста 65
2.2. Формальные маркеры функционально-стилевой принадлежности (МФСП) текста 79
2.3. Исследование потенциальной полисемантичности слов, принадлежащих к различным словарным группам 91
2.4. Выводы 103
3. РЕАЛИЗАЦИЯ ОПТИМИЗАЦИОННОГО ПОТЕНЦИАЛА ФУНКЦИОНАЛЬНО-СТИЛИСТИЧЕСКОЙ КЛАССИФИКАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ 105
3.1. Алгоритм детектирования функционального стиля при помощи МФСП 105
3.2. Программное обеспечение, реализующее МФСП-алгоритм 121
3.3. Результаты опытного использования программного обеспечения, реализующего МФСП-алгоритм 133
3.4. Выводы 171
Заключение 174
Библиографический список 178
Приложение 195