



На правах рукописи

ГОЛОВКО Николай Вячеславович

**ФОРМАЛЬНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ
МНОГОЗНАЧНОЙ ЛЕКСИКИ
КАК СРЕДСТВО ОПТИМИЗАЦИИ
СИСТЕМ АВТОМАТИЗИРОВАННОЙ
ОБРАБОТКИ ТЕКСТОВ**

10.02.19 – Теория языка

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата филологических наук

13 ОКТ 2011


Ставрополь – 2011

Диссертация выполнена в ФГБОУ ВПО
«СТАВРОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»


- Научный руководитель доктор филологических наук, доцент
Гусаренко Сергей Викторович
ФГБОУ ВПО «Ставропольский
государственный университет»
- Официальные оппоненты доктор филологических наук, профессор
Леденев Юрий Юрьевич
ФГОУ ВПО «Ставропольский
государственный аграрный университет»
- доктор филологических наук, профессор
Лебедева Людмила Алексеевна
ФГБОУ ВПО «Кубанский
государственный университет»
- Ведущая организация **ФГБОУ ВПО «Пятигорский
государственный лингвистический
университет»**

Защита состоится 28 октября 2011 г. в 10-00 часов на заседании диссертационного совета ДМ 212.256.02 при ФГБОУ ВПО «Ставропольский государственный университет» по адресу: 355009, г. Ставрополь, ул. Пушкина, 1а, аудитория 416.

С диссертацией можно ознакомиться в научной библиотеке ФГБОУ ВПО «Ставропольский государственный университет» по адресу: 355009, г. Ставрополь, ул. Дзержинского, 120.

Автореферат разослан «» сентября 2011 г.

Ученый секретарь диссертационного совета
доктор филологических наук, профессор


А.А. Фокин

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Первичные исследования, послужившие лингвистической базой для дальнейших разработок в области автоматизированной обработки текстов, 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. 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. Плутников (1981), А.Ф. Лосев (1982), Д.Н. Шмелев (1973), Ю.Д. Апресян (1974), Н.Д. Арутюнова (1988), Е.С. Кубрякова (1992)). Также велись исследования в сфере формализации языковых образцов (А.М. Шахмайкин, 1986).

В настоящее время 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 (в качестве примера может быть упомянут вычислительный кластер IBM Watson (2010)). Из наиболее содержательно близких к данному исследованию диссертаций следует упомянуть работу О.Г. Шевелева «Разработка и исследование алгоритмов сравнения стилей текстовых произведений» (2006).

Актуальность настоящего исследования обусловлена несколькими базовыми факторами. Так, интенсивное развитие информационных технологий, позволяющее накапливать значительные объемы разнообразных данных, автоматически ставит вопросы об их обработке, оперативном поиске требуемой информации, а также об упрощении человеко-машинной коммуникации наряду

с автоматизацией действий, на выполнение которых затрачиваются существенные временные ресурсы. Все это формирует предпосылки для развития и совершенствования автоматизированных систем обработки текстов (АСОТ). Собственно лингвистическое обеспечение упомянутых систем нередко характеризуется недостаточной степенью унифицированности, системности и разработанности, что приводит к выбору далеко не оптимальных аналитических критериев, из чего следует, 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.

Привлечение формально-семантического анализа многозначной лексики для исследования текстовых образцов предоставляет возможность выйти тем самым на формирование у машины определенного эмулированного представления о семантике и о контексте. Анализ разработок в этой области показывает, что в настоящее время 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.

Объектом исследования выступают лексико-семантические аспекты значения текста на естественном языке, в частности – семантический потенциал текста как интегральный показатель многозначности входящих в его состав лексических единиц, определяемый средствами формально-семантического анализа.

Предметом исследования, в свою очередь, являются системные внутриязыковые взаимоотношения, существующие, с одной стороны, между лексической многозначностью и полисемантической текстом, и, с другой стороны, между неоднозначностью текста и его функционально-стилевой отнесенностью. Данные взаимоотношения выражаются в специфических характеристиках естественных языковых текстов, которые, в свою очередь, манифестируются в машиночитаемых формальных маркерах и могут быть подвергнуты автоматизированному анализу.

Цель исследования – выявить, описать и систематизировать те средства формально-семантического анализа многозначной лексики, использование которых позволяет усовершенствовать лингвистическое обеспечение автоматизированных систем обработки текстов и оптимизировать их.

Задачи исследования:

- определить характеристики естественных языковых текстов, обработка которых составляет наиболее эффективные потенциальные пути оптимизации АСОТ;

- на примере русского языка сформировать и описать минимальный набор машиночитаемых формальных маркеров, позволяющих реализовать выбранные пути оптимизации АСОТ посредством классификации естественных языковых текстов на основании упомянутых характеристик;

- изучить классификационные возможности выявленных маркеров, в том числе показателей потенциальной полисемантическойности, находящихся 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 испытание, позволяющее судить о степени эффективности построенного программного обеспечения;

- рассмотреть перспективы повышения эффективности построенного программного обеспечения по результатам проведенных испытаний.

Цель и задачи настоящей работы определяют выбор **методов** исследования. Основными явились общетеоретические методы анализа и синтеза, а также метод автоматизированного анализа текстов и статистический метод. Кроме того, использовались методы моделирования, интроспективного наблюдения, целенаправленной выборки.

В процессе исследования была выдвинута следующая гипотеза: основанием для достоверной и эффективной функционально-стилистической классификации естественных языковых 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 полисемантическойности текста, выявляемой посредством обработки его лексического массива. Потенциальная полисемантическаяность 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 качестве материала исследования выступили словники толковых словарей русского и английского языков («Словарь русского языка» С.И. Ожегова, «Современный толковый словарь русского языка» под редакцией С.А. Кузнецова, Малый академический словарь русского языка в 4-х тт., Merriam-Webster Collegiate Dictionary), а также случайно отобранные электронные копии русскоязычных текстов различной жанровой и функционально-стилевой отнесенности, составившие выборку для итогового тестирования построенного программного обеспечения. Всего было обработано 104 текста.

Научная новизна исследования определяется тем, что в ходе его выполнения были статистически верифицированы классифицирующие особенности формализованных и свободных стилей речи; впервые на основании формально-семантического анализа, результатом которого стали данные об относительной неоднозначности лексических единиц, составляющие естественноречевые тексты, были определены параметры потенциальной полисемантической текстов различной функционально-стилевой принадлежности, а также описан лингвистически обоснованный минимизированный аналитический алгоритм для нужд оптимизации АСОТ, сочетающий в себе универсальность, экономичность, эффективность и высокую скорость работы.

Теоретическая значимость исследования обусловлена тем, что в нем систематизированы, обработаны и сведены в единую дефиницию разрозненные определения формально-семантического анализа, представлен способ оценивания потенциальной полисемантической текстов на основе расчета и сопоставления массовых долей специфических словарных групп (данное терминологическое выражение 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.

Практическая значимость работы состоит в том, что 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.

Методологические основания исследования составляют общепризнанные положения о взаимной связи плана содержания и плана выражения языкового знака, а также фундаментальный принцип асимметрического дуализма языкового знака.

Теоретическими основаниями исследования выступили общетеоретические положения лексико-семантических и контекстуальных концепций и теорий А.Ф. Лосева (1982), Ю.Д. Апресяна (1974), Д.Н. Шмелева (1973), Е.С. Кубряковой (1992). Кроме того, были привлечены теоретические и практические разработки известных специалистов в области компьютерной и квантитативной лингвистики – Р. Шенка (1977), М. Селфридж (1977), А.Н. Баранова (1987), Б.Ю. Городецкого (1989), Ч. Филлмора (1985). Вопросы, связанные с функциональной стилистикой русского языка, потребовали рассмотрения взглядов и идей М.Н. Кожиной (2008), В.В. Виноградова (1963), О.А. Крыловой (2006), Л.Ю. Максимова (1975). Также были изучены работы А.М. Шахмайкина (1986), P. Cimiano, S. Staab, J. Tane (2003), D. Ferrucci (2004), И.В. Азаровой, Е.А. Овчинниковой (2005) в области формально-семантического анализа; проанализированы взгляды П.И. Браславского (1999) и О.Г. Шевелева (2006) на автоматическое определение функционально-стилевой принадлежности русскоязычных текстов.

Положения, выносимые на защиту:

1) 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 естественноразговорных текстов в рамках комплексного подхода к формированию их лингвистического обеспечения.

2) Минимальный набор машиночитаемых языковых маркеров, обработка и анализ которых позволяют достичь успешной оптимизации АСОТ посредством классификации естественноразговорных текстов по признаку их функционально-стилевой отнесенности, состоит из двух диагностических критериев: средней длины слова и потенциальной полисемантической текста.

3) Классификационные возможности минимального набора машиночитаемых языковых маркеров, состоящего из диагностических критериев средней длины слова и потенциальной полисемантической текста,

достаточно для уверенного определения типа функционального стиля русскоязычного текста (свободный тип / формальный тип).

4) Подсчет и сравнение массовых долей слов, принадлежащих к словарным группам, которые характеризуются минимальным и максимальным относительным количеством многозначных слов в соответствии с актуальной словарной нормой, составляют надежный способ оценивания потенциальной полисемантической русскоязычного текста в целом.

5) Существуют устойчивые корреляции между семантическим потенциалом слова и его графической манифестацией, в частности – с инициальной графемой. Минимальной потенциальной полисемантической характеризуются словарные группы «А» (16%...33%), «Б» (17%...37%), «Д» (20%...36%), «И» (23%...38%), «Й» (11%...33%), «Э» (21%...36%), «Ю» (24%...39%); максимальной – «ЕЁ» (29%...39%), «Ж» (22%...41%), «У» (29%...49%), «Ц» (28%...52%), «Ч» (23%...47%), «Ш» (28%...47%), «Я» (33%...43%).

Результаты исследования были апробированы в ходе международных конференций студентов, аспирантов и молодых ученых «Ломоносов» (Москва, 2009-2011 гг.), III Международной научно-методической конференции «Русскоязычие и би (поли) лингвизм в межкультурной коммуникации XXI века: когнитивно-концептуальные аспекты» (Пятигорск, 2010 г.), II Международной научно-практической конференции «Наука и современность» (Новосибирск, 2010 г.). 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.

Основные положения диссертационного исследования отражены в 11 публикациях, в том числе в рецензируемых научных журналах из перечня изданий, рекомендованных ВАК Минобрнауки России для защиты докторских и кандидатских диссертаций – «Вестник Ставропольского государственного университета» (Ставрополь, 2008; Ставрополь, 2011), «Вестник Пятигорского государственного лингвистического университета» (Пятигорск, 2009), «Знание. Понимание. Умение» (Москва, 2011).

Структура исследования. Диссертация состоит из введения, трех глав, заключения, библиографического списка и приложения. Библиографический список включает 135 книг и статей, 15 Интернет-источников, а также 108 источников лингвистического материала.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе «Теоретические основания формально-семантического анализа многозначной лексики как средства оптимизации автоматизированных систем обработки текстов (АСОТ)» подвергаются анализу теоретические исследования по теме диссертации с целью формирования понятийного аппарата исследования и выведения рабочих определений.

Изучение вопроса о природе значения языкового знака позволило акцентировать внимание на прагматическом аспекте языковой единицы, постулируя тем самым контекстную обусловленность ее значения (А.Ф. Лосев, 1982). При этом 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 по-новому взглянуть на вопросы «машинного понимания» текстов на естественном языке.

В исследовании отмечена насущная 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 (Р. Шенк, 1977). Семантика языкового знака принципиально может изучаться в отрыве от сопутствующих ей синтаксического и прагматического аспектов, однако в целях максимально полного понимания значения целесообразно 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 A называется многозначным, если для любых двух его значений a_1 и a_2 найдутся такие значения $a_1, a_2, \dots, a_k, a_1$, что a_1 сходно с $a_1, a_1 - c a_2$ и т.д., a_k с a_1 и $a_1 - c a_j$ ». 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 определения 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 (Р. Шенк, 1977; Ч. Филлмор, 1985). В класс последних могут быть включены информационные языки, что позволяет говорить об использовании в информационной работе всего спектра теоретических и методических средств 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 (А.Н. Баранов, 2001). 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 науке о языке.

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 естественных языковых текстов видится в построении оптимизационных алгоритмов, учитывающих вышеупомянутые потребности.

Во второй главе – «Возможности оптимизации АСОТ посредством функционально-стилистической классификации русскоязычных текстов» – представлены результаты теоретических и эмпирических исследований,

направленных на построение оптимизационных решений для нужд АСОТ на основании надежного и достоверного лингвистического обеспечения.

В ходе исследования установлено, что наиболее целесообразно формировать такой 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 системный подход к организации машинных анализаторов. В связи с этим приоритетной задачей в процессе построения средств оптимизации автоматизированного анализа стал поиск тех параметров естественных языковых 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.

Совершение корректного выбора в пользу того или иного значения полисемантического 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 физическому, 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 полисемантического языкового знака.

В диссертации выдвигается положение, согласно которому в рамках комплексного подхода к оптимизации процедур автоматизированного анализа текстов естественным образом подлежат учету все факторы, оказывающие непосредственное воздействие на интерпретацию текста, что предопределило необходимость изучения и применения в интересах указанной оптимизации как физического, так и идеального контекста. С учетом данной позиции формально-семантический анализ – в полном соответствии с предложенным выше определением – выступил как эффективное средство оптимизации, поскольку позволил сформировать представление о значении как в пределах физического и идеального контекста, так и вне его.

Краткий анализ существующих воззрений на структуру и компоненты идеального контекста (иными словами, ситуации общения в наиболее

широком понимании данного термина, т.е. множества условий и факторов (по преимуществу экстралингвистического характера), в рамках или под воздействием которых осуществляется коммуникация, также как и тех фоновых знаний, которые учитываются коммуникантами в процессе общения) позволил установить, что не существует единого подхода к выделению вышеуказанных компонентов. Распространенной является модель речевой ситуации Якобсона-Хаймса (Р. Якобсон, 1975; Д. Хаймс, 1975), в соответствии с которой 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; данная модель была принята в качестве основной.

В исследовании отмечается, что автоматическое определение функционально-стилевой принадлежности текстов характеризуется равнозначной ценностью для всех типов автоматических анализаторов; в то же время 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. Эти причины обусловили обращение к анализу формально-содержательных признаков функциональных стилей речи, а также изыскание тех их параметров, которые отвечали бы ряду следующих базовых требований: 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, а также наличие эвристического и оптимизационного потенциала, позволяющего реализовать высокопроизводительные алгоритмы, не требующие организации и ведения объемных БД.

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. Очевидно, что решение вопроса о классификации функциональных стилей непосредственным образом зависит от оснований классификации; надежным основанием могут послужить разнообразные формы общественного сознания, в соответствии с которыми выделены 6 функциональных стилей – научный, официально-деловой, публицистический, художественный, религиозный и разговор-

но-обиходный (М.Н. Кожина, 2008). 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 соответствии с их сущностью, характерными особенностями и решаемыми задачами данные стили могут быть подвергнуты типизации и разделены на свободные (художественный, публицистический) и формализованные (научный, официально-деловой).

В ходе дальнейшего исследования были выявлены те исчисляемые параметры текстов на естественных языках, которые позволяют составить представление о его функционально-стилевой принадлежности, а также определены возможности автоматического определения функционального стиля текста на основании анализа указанных параметров. Определение функционально-стилевой принадлежности текста, в свою очередь, является инструментом реализации комплексного подхода 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.

В результате исследования установлено, что на основании типических черт различных функциональных стилей и их групп могут быть выделены два базовых параметра, которые обладают достаточным потенциалом для реализации на их основе оптимизированного алгоритма стилового 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 не является новым в вопросах анализа стиловой принадлежности текста, однако типические характеристики функциональных стилей предоставили основания для утверждения, что параметр средней длины слова может являться (и, как было выяснено в итоге, действительно является) единственным количественным параметром текста, достаточным для определения его функционально-стилевой принадлежности.

С другой стороны, это качественный показатель, описывающий степень полисемантической текста, или, иначе, его семантический потенциал. В силу определенной ценности подобного показателя для процесса выявления функционально-стилевой отнесенности текста было признано необходимым выявить такой формальный критерий анализа, который мог бы обеспечить его достоверное оценивание.

В результате исследования внешних признаков потенциальной полисемантической текста было установлено, что низкая степень полисемантической текста может свидетельствовать о высокой доле в нем заимствованных слов, в том числе терминологического характера, и наоборот; на основании данной установки решение поставленной задачи было сведено к изысканию внешних признаков заимствованных и не заимствованных слов, а получаемые данные – соотнесены с вопросами потенциальной полисемантической текста.

Анализ в данном направлении производился с опорой на лексикографические источники. В качестве наиболее очевидного и просто реализуемого приема, который позволил бы оценить роль характерных и не характерных для языка графических символов в оценивании лексической многозначности, было избрано изучение словарных групп, так как 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.

В ходе дальнейшего исследования было 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. Данный вывод базируется на положении о том, что, поскольку те или иные звуко-сочетания, 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 полисемантической. Потенциальная же полисеманτικότητα 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 позволяет составлять представление о степени потенциальной полисемантической текста, т.е. о плане его содержания, посредством формального анализа – изучения плана выражения текста.

Было проведено практическое исследование словарного материала, отдельные результаты которого¹ представлены ниже в качестве примера.

¹ Современный толковый словарь русского языка / Под ред. С.А. Кузнецова. – М., 2004.

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				

Условные обозначения в таблице: 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

Изучение словарного материала и проведение требуемых расчетов в совокупности с последующим анализом полученных результатов позволили выделить пять типов характеристических словарных групп, характеризующихся значимо высокой / низкой массовой долей неоднозначных слов, значимо высокой / низкой массовой долей слов среднего уровня полисемантизации, значимо высокой / низкой массовой долей сильно полисемантизованных слов, после чего определить 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 «ЕЁ» (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 и подлежит специфической обработке с приданием повышенного веса коннотативным значениям слов и подключением дополнительных тематических словарей, 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 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 словарях имеют наиболее высокие / низкие коэффициенты общей степени полисемантизации, среднего уровня многозначности, высокого уровня многозначности и суммарного количества слов, имеющих более трех значений.

В третьей главе – «Реализация оптимизационного потенциала функционально-стилистической классификации русскоязычных текстов» – описаны 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 количественные значения средней длины слова и предоставило возможность наиболее общим образом оценить перспективы параметра потенциальной полисемантичности текста как детектирующего критерия. С этой целью были изучены средняя длина слова и относительная степень

полисемантизованности текстов различных функциональных стилей, а также исследована возможность применения тех или иных полученных показателей в целях идентификации стилевой принадлежности текста.

Определено, 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 МФСП-алгоритма, основанного на оценивании средней длины слова и потенциальной полисемантичности, достаточны для уверенной классификации текстов по признаку функционально-стилевой отнесенности на свободные и формальные.

Был разработан аналитический алгоритм, состоящий из двух этапов: подготовительного и основного. Этап предварительной обработки решает задачи очистки текста от потенциальных помех и подготовки его к формально-семантическому исследованию; он включает удаление знаков препинания и других особых знаков, создающих помехи, имен собственных и аббревиатур, а также служебных слов, у которых дейктический или функционально-служебный элемент значения доминирует над собственно лексическим. Необходимо заметить, что указанные операции производятся в оперативной памяти ЭВМ и не отражаются на состоянии исходного текста, который постоянно хранится в особой строковой переменной и в любой момент может быть подан на вход другого информационного процесса, следующего за МФСП-анализатором в последовательности машинных действий.

Основной этап анализа включает расчет данных по первому детектирующему критерию (средняя длина слова), по второму детектирующему критерию (потенциальная полисемантичность), а также анализ данных по обоим детектирующим критериям и вспомогательные верификационные процедуры, необходимые для уточнения вердикта в случае противоречия основных параметров. Одной из подобных процедур является отсеивание недостоверных словарных групп на основании данных о тематике текста; для этих целей при написании программного обес-

печения было предусмотрено особое текстовое поле, предназначенное для указания темы либо ключевых слов; словарные группы, к которым принадлежат введенные в поле слова, проверяются программой на соответствие выявленным отклонениям и при необходимости помечаются как характеризующиеся сниженной либо нулевой достоверностью.

Предварительные результаты исследований, позволяющие первоначально задать критерии и диапазоны для проверки исследуемых текстов, механизм предварительной очистки текстового материала от потенциальных помех и основная логическая последовательность действий, которые машина будет исполнять в процессе анализа, составили полноценный набор данных и логических построений, необходимых для разработки программного обеспечения, реализующего МФСП-алгоритм, с его последующим детализированным описанием и практическим применением на более масштабной выборке текстов, позволяющим уточнить границы заданных аналитических диапазонов и снизить количество вероятных ошибок, недочетов и ложных срабатываний.

Аналитическое приложение, реализующее МФСП-алгоритм, построено 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 программном обеспечении реализованы все этапы и промежуточные шаги, запланированные в ходе построения аналитического алгоритма.

При построении первого действующего образца анализатора его база данных и внутренние инструкции основывались на предварительных результатах, полученных при проверочном изучении четырех эталонных текстов, представляющих каждый из отобранных функциональных стилей. 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. По итогам верификационного анализа на материале текстов, взятых за эталон, 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.

Были внесены соответствующие модификации в собственную базу данных МФСП-анализатора, равно как и в его внутренние алгоритмы, обеспечивающие проведение автоматического исследования; вследствие этого детектирующие механизмы продукта стали функционировать более четко и точно, позволяя достичь уверенного разделения вводимых в анализатор текстов на типы в соответствии с признаками их функционально-стилевой принадлежности.

Опытная проверка МФСП-анализатора на широкой выборке 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.

Материал общим объемом в 100 текстов, где в равных пропорциях были представлены образцы четырех функциональных стилей, был отобран в произвольном порядке без явных требований к объему, жанру и преобладающей тематике, в то же время были предприняты усилия по обеспечению многообразия выборки по упомянутым выше аспектам, что позволило повысить репрезентативность корпуса текстов и получить более достоверные данные, слабо зависящие от конкретных особенностей тех или иных подвергаемых изучению образцов.

Результаты исследования показали, что МФСП-анализатор успешно классифицировал в общей сложности 93 из 100 предъявленных ему текстов, а в оставшихся 7 случаях завершил анализ с нечетким вердиктом, не допустив фактов 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 полисемантизированных словарных групп в текстах свободных и формальных функциональных стилей.

Изучение результатов первого тестового прогона указало на необходимость сокращения списка показательных словарных групп: было отбраковано в общей сложности 4 группы из 14. Это обусловило пот-

ребность в проведении вторичной проверки на основании 10 словарных групп; в результате были получены новые данные, которые позволили повысить четкость работы программного обеспечения: общая точность его работы возросла до 97%, причем 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, так что уточняющие процедуры успешно отработали в 8 случаях.

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%.

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 полисемантической, определяемой по преобладанию в тексте слов, относящихся к определенным словарным группам. Испытания разработанного программного обеспечения на широкой выборке текстов свидетельствуют о том, что выявленные закономерности в целом имеют универсальный характер, и случаи отклонений весьма редки. Алгоритм, реализованный в виде демонстрационного образца, 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.

С целью проиллюстрировать работу алгоритма предлагается фрагмент одного из текстов, вошедших в тестовую выборку, с соответствующими отметками и пояснениями.

С наступлением XXI века определился новый этап развития потребностей общества. Информационный бум, формирование рыночных отношений в мире труда, сложные экономические условия требуют подготовки человека к активному самостоятельному решению многих жизненных ситуаций, в том числе к подготовке для своего будущего трудоустройства, в выборе профессии, к формированию способности самостоятельно ориентироваться в мире информации. В центре внимания такой подготовки должна находиться личность человека, а в центре внимания педагогической науки – создание личностно – ориентированных новых технологий обучения.

<...>

Корпоративный тренинг максимально конкретен и функционален, а главной целью его является поддержка и повышение конкурентоспособности организации и персонала, в ней работающих. Недавние исследования, проведенные в 3200 американских компаниях, показали, что 10%-ное увеличение расходов на тренинг персонала дает прирост производительности труда в 8,5 %, в то время как такое же увеличение капиталовложений дает прирост производительности труда только на 3,8%.

Кроме перечисленных навыков технология социально-психологического тренинга помогает расширить рамки профессионального мастерства, ликвидировать дефицит управленческих и коммуникативных навыков руководителей и специалистов ювелирной отрасли. Данные тренинги стимулируют изменения в промышленной и коммерческой деятельности организаций и несут инновационный характер в современной системе образования, а также способствуют позитивным изменениям в данных организациях при переходе на новый уровень развития в рыночной экономике.

(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.ХI. –Уфа, 2010.)

В этом фрагменте комбинацией полужирного шрифта и подчеркивания отмечены формальные элементы, на которые реагирует аналитический алгоритм в процессе оценивания потенциальной полисемантической, а курсивом – элементы, им принципиально игнорируемые. Согласно подсчетам анализатора, в полном варианте данного текста 6307 символов, из них 1008 служебных и 5299 неслужебных (т.е. букв); признаков начала слова обнаружено 746. На основании этих данных определяется средняя длина слова, которая составляет в данном случае 7,1032. Так как полученная величина больше заданной пороговой величины в 6 знаков, по первому критерию текст определяется как формализованный.

Далее рассчитываются показатели и массовые доли характеристических словарных групп: «А» – 10 (1,3405%) слов, «И» – 24 (3,2172%), «Й» – 0 (0%), «Э» – 11 (1,4745%), «Ю» – 9 (1,2064%); «ЕЁ» – 10 (1,3405%), «Ж» – 2 (0,2681%), «Ч» – 8 (1,0724%), «Щ» – 0 (0%), «Я» – 3 (0,4021%). В результате суммирования находятся итоговые данные по второму критерию: слабо полисемантизированные словарные группы – 54 (7,2386%), сильно полисемантизированные – 23 (3,0831%). Разность массовых долей положительна, из чего делается вывод о низкой потенциальной полисемантической текста в целом.

Сравнение данных по первому и второму диагностическим критериям не обнаруживает различий между предварительными субвердиктами, на основании чего анализатор выносит окончательный вердикт: текст при-

надлежит к одному из формализованных функциональных стилей (в данном случае, как можно наблюдать, – к научному). Вспомогательные процедуры верификации не запускаются в силу отсутствия потребности в них. Процесс анализа вышеприведенного текста занял 1 секунду.

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%. 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.

В **Заключении** подводятся основные итоги работы и намечаются перспективы дальнейшего исследования.

Основные положения диссертации изложены в следующих публикациях:

1. Головкин Н.В. Логико-количественный аспект теории фиксации типов языковой информации [Текст] / Н.В. Головкин // Вестник Ставропольского государственного университета. – 2008. – № 3. – Вып. 56. – С. 72-79. – [Статья – 0,6 п.л.]. – **Издание из перечня ВАК РФ.**

2. Головкин Н.В. К вопросу о формальной идентификации функционального стиля в русском языке [Текст] / Н.В. Головкин // Вестник Пятигорского государственного лингвистического университета. – 2009. – №2. – С. 16-19. – [Статья – 0,25 п.л.]. – **Издание из перечня ВАК РФ.**

3. Головкин Н.В. Алгоритм автоматической классификации русскоязычных текстов по признаку функционально-стилевой отнесенности [Текст] / Н.В. Головкин // Вестник Ставропольского государственного университета. – 2011. – № 4. – Вып. 75. – С. 54-59. – [Статья – 0,5 п.л.]. – **Издание из перечня ВАК РФ.**

4. Головкин Н.В. Расширенный механизм формальной идентификации функциональных стилей русскоязычных текстов [Текст] / Н.В. Головкин // Знание. Понимание. Умение. – 2011. – № 3. – С. 201-203. – [Статья – 0,2 п.л.]. – **Издание из перечня ВАК РФ.**

5. Головкин Н.В. Концепция коэффициентов полисемизации как квантитативное основание теории фиксаторной идентичности языковых явлений [Текст] / Н.В. Головкин // Информационные технологии в науке и образовании: Материалы Международной научн.-практ. интернет-конференции. – Шахты: Изд-во ЮРГУЭС, 2006. – С. 3-5. – [Статья – 0,25 п.л.].

6. Головкин Н.В. Длина слова как характеризующий показатель замкнутости и степени полисемизации [Текст] / Н.В. Головкин // Язык как воплощение культуры: лингвистическая, переводческая и дидактическая рефлексия: В 2-х ч. Ч. I. – Ставрополь: Ставро. кн. изд-во, 2006. – С. 72-76. – [Статья – 0,25 п.л.].

7. Головкин Н.В. Статистическое исследование количественных показателей лексической многозначности в русском языке [Электронный ресурс] / Н.В. Головкин // Материалы докладов XVI Международной конференции студентов, аспирантов и молодых ученых «Ломоносов – 2009». – М.: МГУ, 2009. – 1 CD-ROM. – Международная конференция «Ломоносов – 2009». – [Статья – 0,2 п.л.].

8. Головкин Н.В. Опыт разработки алгоритма формальной идентификации функционально-стилевой принадлежности русскоязычных текстов. [Электронный ресурс] / Н.В. Головкин // Материалы докладов XVII Международной конференции студентов, аспирантов и молодых ученых «Ломоносов – 2010». – М.: МГУ, 2010. – 1 CD-ROM. – Международная конференция «Ломоносов – 2010». – ISBN 978-5-317-03197-8. – [Статья – 0,2 п.л.].

9. Головкин Н.В. Функциональный стиль текста как фактор оптимизации автоматизированного анализа [Текст] / Н.В. Головкин // Материалы II Международной научно-практической конференции «Наука и современность-2010»: В 3-х ч. Ч. 3. – Новосибирск: Изд-во «СИБПРИНТ», 2010. – С. 104-110. – [Статья – 0,5 п.л.].

10. Головкин Н.В. Формально-семантический анализ как компонент комплексного подхода к автоматизированной обработке текстов [Текст] / Н.В. Головкин // Материалы III Международной научно-методической конференции «Русскоязычные и би(поли)лингвизм в межкультурной коммуникации XXI века: когнитивно-концептуальные аспекты». – Петрозаводск: Изд-во ПГЛУ, 2010. – С. 59-66. – [Статья – 0,44 п.л.].

11. Головкин Н.В. Формальные маркеры функционально-стилевой принадлежности русскоязычных текстов [Электронный ресурс] / Н.В. Головкин // Материалы докладов XVIII Международной конференции студентов, аспирантов и молодых ученых «Ломоносов – 2011». – М.: МГУ, 2011. – 1 CD-ROM. – Международная конференция «Ломоносов – 2011». – ISBN 978-5-317-03634-8. – [Статья – 0,2 п.л.].

24

Подписано в печать 22.09.11

Формат 60x84 1/16	Усл.печ.л. 1,4	Уч.-изд.л. 1,31
Бумага офсетная	Тираж 100 экз.	Заказ 584

Отпечатано в Издательско-полиграфическом комплексе
Ставропольского государственного университета.
355009, Ставрополь, ул.Пушкина, 1.