Компьютерные программы обработки русскоязычных текстов

Настоящий обзор выполнен в рамках исследовательского проекта «Современные русские профессиональные социолекты», реализуемого по гранту РГНФ 12-04-00381/12.

Компьютерная обработка текстов, в первую очередь текстов, созданных на флективных языках, основывается на морфологическом и синтаксическом анализе синтагм, предложений и СФЕ в соответствии с правилами формальной грамматики. Работа программ опирается на статистическую основу – корпус авторитетных текстов, которые предварительно аннотированы разработчиками и использованы для «обучения» программы, а также алгоритмическое индексирование той или иной словарной базы, обычно – словаря, каждый элемент словника которого снабжен морфологическим модификатором (модификаторами). Широко используется вероятностный подход. Существенными задачами такого анализа является машинная обработка значительных объемов информации и обобщенное представление ее основного смысла в сжатой форме, вычленение смысловых доминант и тематической структуры, определение формальных характеристик стиля и жанра.

Безусловно, никакая программная обработка текста не может заменить собой анализ, который может осуществить человек – особенно эксперт в той или иной области. Однако программы, о которых здесь идет речь, позволяют специалисту прийти к заключениям о тенденциях, потратив на проведение исследования меньшее количество времени. Кроме того, эти программы позволяют апробировать гипотезы на большем объеме материала и с большей долей уверенности в объективности полученных данных. Именно с этих позиций и будут рассмотрены имеющиеся сегодня программы обработки русскоязычных текстов.

Разумеется, предложенные разными коллективами программы могут быть использованы в различных областях знания и с различными целями, мы сосредоточимся на выявлении тех позиций, которые могут дать позитивные результаты для описания современных профессиональных социолектов как составляющей живой речи в социальном аспекте.

1. Первая группа компьютерных программ предназначена для **синтаксического** и **морфологического** анализа русскоязычных текстов. Грамматический срез – один из важнейших при формировании целостного представления о системе языка, так что эти программы могут быть полезны в нашем исследовании.
   Russian Morphological Dictionary (http://faqproject.ru/morphological-dictionary-russian-download.html, автор – Сергей Сикорский) работает с входным ASCII-текстом. Используется морфологический словарь А.Зализняка, включающий 120.000 слов. Реализована на SWI-Prolog для Windows. Основан на сайте www.ruscorpora.ru. Программа быстро и с опорой на авторитет указанного словаря определяет грамматические признаки слов. При обращении к текстам социолектной принадлежности это может обеспечить доказательную атрибуцию морфов, используемых в речи пользователей социальных сетей. С другой стороны, необходимо иметь в виду проблему *ограниченности словника* словаря А.Зализняка, в котором отсутствуют имена собственные, некоторые неологизмы последнего времени, сравнительные формы вроде *постарше,* наречия вида *по-детски,* многие сложные слова, пишущиеся через дефис, многие наречия на *-о* и *-е* (последняя задача не снимается введением синкретического класса “наречие/краткая форма прилагательного”). Соответственно, мы прогнозируем затруднения при определении грамматической принадлежности новых для системы литературного языка слов.
   Mystem (авторы - Илья Сегалович, Виталий Титов, компания Яndex) – это компактный, очень быстрый и бесплатный морфологический парсер русскоязычных текстов, реализованный также на основе словаря А.Зализняка. Доступны для загрузки версии для

Windows и Linux. Работает как консольное приложение и имеет различные режимы представления результатов. В общем, программа Mystem производит морфологический анализ литературного нормативного текста на русском языке. Для слов, отсутствующих в словаре, порождаются гипотезы на основании частотности суффиксов. Следовательно, неологизмы и окказионализмы, появление которых легко прогнозировать в социолекте, не получат при использовании этой программы достаточно аргументированного опознания, однако факт наличия гипотезы, сформированной на основе имеющихся в программе сведений о функционировании литературного языка, не может не порадовать. К сожалению, в подавляющем количестве отзывов, которые оставляют пользователи этой программы, отмечается сложность установки программы и введения нужных параметров исследования.

Рабочее Место Лингвиста (www.aot.ru, компания Dialing (Москва)), предлагает анализ текстов для построения систем автоматического перевода с русского на английский язык (и наоборот). Включает ряд автономных компонентов:

- синтаксический анализатор текстов на русском языке;
- морфологический анализатор текстов на русском и английском языках;
- построение конкордансов для заданной совокупности текстов.

Система написана на языке C++ и работает в среде Windows 9x/2000/NT. Программа имеет множество позитивных откликов от исследователей различных научных областей, однако создается впечатление, что проект приостановлен, т.к. по указанному адресу программы нет, не предлагается и никакого альтернативного пути знакомства с продуктом.

Морфологический анализатор (http://www.keva.ru/ling/rus/help.htm, автор - С.А.Старостин) – это онлайн-версия программы морофлогического анализа слов русского/английского языков. Позволяет получить для вводимого слова базовую форму и морфологическую информацию. Программа реализована на основе словарей А.А. Зализняка и В.К. Мюллера (английский язык). В Морфологический анализатор может быть введено любое русское или английское слово в произвольной грамматической форме. Программой анализа выдаются следующие сведения для русского слова:

a) исходная слоформа (по Зализняку);
b) словарная информация, то есть морфологический индекс русского слова и имеющиеся комментарии из Грамматического Словаря Зализняка;
c) перевод, то есть набор словарных статей из словаря Мюллера, в которых содержится соответствующее русское слово, с готовыми отсылками на соответствующие словарные статьи;
d) морфологическая характеристика введенного русского слова.

В случае многозначности введенной формы выводятся все варианты анализа. Именно возможность получить варианты анализа введенной в программу формы представляется привлекательной с точки зрения нашего проекта возможностью, т.к. эти варианты дадут почву для объективного определения места морфа в системе языка.

2. Вторая группа программ автоматической обработки текстов объединяет продукты, которые позволяют прийти к обобщенному представлению о частоте выявленных **лексических единиц**, об их группировке в текстах, а также дают основания для исследования **семантических процессов** в изучаемых речевых продуктах.

TextAnalyst 2.0 (http://www.analyst.ru/index.php?lang=eng&dir=content/products/&id=ta) произведен научно-производственным инновационным центром"МикроСистемы" как инструмент анализа символьных текстов. Позволяет построить семантическую сеть понятий, выделенных в обрабатываемом тексте, со ссылками на контекст. Имеется возможность

смыслового поиска фрагментов текста с учетом скрытых в тексте смысловых связей со словами запроса. Позволяет анализировать текст путем построения иерархического дерева тем/подтем, затрагиваемых в тексте. Также имеется возможность реферирования текста.

Кроме отдельного продукта TextAnalyst , также предлагается инструментарий разработчика TextAnalyst SDK, включающий функции лемматизации (приведения слов к нормальной форме) для русского и английского языков, построения частотных списков понятий, поиска слов в контексте и т.д.

Еще одна компонента, TextAnalyst Lib, может использоваться для построения гипертекстовых электронных книг.

Все компоненты реализованы для Windows 95 и выше и доступны для бесплатной загрузки. На американском рынке технологию TextAnalyst продвигает фирма Megaputer Intelligence Inc.

Основные возможности:

- анализ содержания текста с автоматическим формированием семантической сети с гиперссылками - получение смыслового портрета текста в терминах основных понятий и их смысловых связей;
- анализ содержания текста с автоматическим формированием тематического древа с гиперссылками - выявление семантической структуры текста в виде иерархии тем и подтем;
- смысловой поиск с учетом скрытых смысловых связей слов запроса со словами текста;
- автоматическое реферирование текста - формирование его смыслового портрета в терминах наиболее информативных фраз;
- кластеризация информации - анализ распределения материала текстов по тематическим классам;
- автоматическая индексация текста с преобразованием в гипертекст;
- ранжирование всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования;
- автоматическое формирование полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации.

Очевидно, что программный продукт с такими широкими возможностями найдет свое применение в процессе нашего исследования. Особенно хочется подчеркнуть, что результаты подобной автоматической обработки текста позволят выявить семантические связи, которые мозг и языковое сознание каждого отдельного человека выявляет достаточно субъективно, основываясь на личном языковом и культурном опыте.

Galaktika-ZOOM (www.galaktika-zoom.ru) произведена корпорацией Галактика (Москва) и представляет собой автоматизированную систему поиска и аналитической обработки информации. Это мощный инструмент анализа и обработки текста (Text Mining), позволяющий извлекать необходимые сведения из огромного объема данных.

Программа позволяет вести профессиональный поиск информации, основанный на принципиально ином по сравнению с другими поисковыми системами подходе, – внимании к анализу и уточнению найденной информации. При обработке запроса «Галактика ZOOM», кроме списка документов, где содержится информация по тому объекту, который ищет пользователь, формирует еще и информационный портрет объекта – список значимых для полученной по запросу выборки слов и словосочетаний, которые и следует уточнить. При работе с информационным портретом пользователь может получить общее представление об объекте (флэш-репорт), уточнять запрос по отдельным словам, составляющим информационный портрет объекта, отсекать лишнюю информацию, определять связи между отдельными словами, составляющими информационный портрет. Знаменательно, что комплекс «Галактика ZOOM» широко применяется на телеканале «Россия» для решения актуальных задач информационной службы программы «Вести». Используя комплекс, журналисты находят нужную информацию – факты, о которых будет рассказано в программах канала; определяют «бэкграунд» основных новостей (предыстория событий, комментарии и оценка экспертов, прогнозы), что необходимо для их быстрого и качественного освещения. Возможности «Галактики ZOOM» помогают заранее формировать список предстоящих значимых событий, которые затем будут отражены в новостных выпусках канала, планировать телевизионные съемки. Тем самым «Галактика ZOOM» позволяет получать качественный информационный результат в кратчайшие сроки. Все эти особенности и

характерные черты компьютерного продукты могут быть полезны и в нашем исследовании – например, при проверке гипотезы о наличии особый «слов-сигналов», которые позволяют профессионалам опознавать того или иного человека как члена своей корпорации, своей социо-профессиональной группы.

Система Пропись 4.0 (АО «Агама») предназначена, в общем, для иных целей. Это набор средств для лингвистической обработки русскоязычных текстов:

- проверка орфографии;
- расстановка переносов;
- построение списка синонимов и антонимов слова;
- грамматическая и стилистическая проверка текста;
- толкование слова (по Толковому словарю);
- поиск и замена слов в тексте с учетом их форм;
- статистический анализ текстов.
  Можно сказать, что эта система хороша для коррекции и обучения, однако наличие таких функции, как соотнесение с Толковым словарем и статистический анализ текстов может сделать этот комплекс полезным для исследования социолектов.

NetXtract (http://download.cnet.com/NetXtract-Personal/3000-12512_4-10073214.html) произведен Relevant Software Inc. Это замечательная компонента, подключаемая к Microsoft Internet Explorer (версии 5.0 и выше), которая позволяет быстро получить упорядоченный индекс слов в загруженном HTML документе. Индекс может быть упорядочен по алфавиту или частоте. Для каждого слова в индексе можно исследовать контекст, в котором это слово встречается. Выбранные слова по желанию заносятся в персональную базу знаний, которая позволяет систематизировать найденные документы удобным образом. С помощью этой программы можно быстро найти нужную информацию на веб-страницах и в документах, а потом сохранить её в собственной базе данных. NetXtract автоматически индексирует каждый документ, отображаемый в IE, выделяет все контексты для любого ключевого термина по вашему выбору и позволяет вам выбирать наиболее интересный контекст. Таким образом, для систематизации лексических единиц и выявления закономерностей их использования, для понимания среды формирования новых лексических значение использование этой программы может быть очень эффективным средством.

WordStat (Яндех, автор - А.Г. Дубинский) предлагает подсчет частоты встречаемости различных слов в текстовых или html-файлах. Понимает основные русские кодировки, игнорирует html-разметку. WordStat может быть использован для быстрого извлечения и анализа информации из большого количества документов.

Используется для:

- контент-анализа: открытые ответы, интервью или фокус-группы, стенограммы;
- бизнес-аналитики и конкурентного анализа веб-сайтов;
- извлечения информации и знаний из отчетов об инцидентах, жалоб клиентов;
- контент-анализа новостей или научной литературы;
- автоматической маркировки и классификации документов;
- выявления случаев мошенничества, авторства, патентного анализа; таксономической разработки и валидации.
- является наиболее используемым сервисом, который показывает статистику ключевых слов и помогает в прогнозировании трафика.
  К сожалению, этот программный продукт требует большей компьютерной грамотности, чем вышеуказанные аналоги, однако при определенной подготовке может быть привлечен для апробации гипотез и для получения максимально объективных данных.

3. В третьей группе программных продуктов собраны те системы, которые позволяют собирать данные, необходимые для определения стилевой принадлежности текстов, а также степени оригинальности текстов и/или приверженности авторов текстов той или иной стилистической манере.

Свежий взгляд/Fresh Eye версия 1.21, 1995 (http://quittance.ru/tautology.php#effectus), продукт Д.Кирсанова, это DOS-утилита, реализующая стилистическую проверку русскоязычных текстов. Программа отыскивает в тексте места, где фонетически и морфологически схожие слова расположены в непосредственной близости, что порождает паронимию (например: «минимально возможное количество информации, которое можно...». В процессе работы над нашим предметом исследования подобная точка зрения характеризации речевого продукта может оказаться весьма продуктивной.

Технологии поиска и анализа текстовой информации (продукт компании «Гарант-Парк-Интернет») – это сайт, на котором представлены разработки известной компании Гарант-Парк-Интернет. Среди представленных технологий:

- анализ и классификация текстов, автоматическое реферирование;
- различные варианты поиска текста;
- морфологический, синтаксический и семантический анализ текста;
- средства навигации по большим массивам текстов.
  Безусловно, автоматическое реферирование не может быть целью при углубленном филологическом анализе, однако использование этой программы позволит делать некоторые выводы относительно тем, волнующих профессиональные сообщества, т.к. автоматическая обработка текстов даст возможность получить факты относительно очень большого объема контентов.

Худломер (продукт Л. Делицына) связан с задачей автоматической классификации стиля русскоязычных текстов. Автором были собраны и проанализированы 4 корпуса текстов, взятых из русской сети. Сюда вошли художественные произведения, публицистика, научные статьи и протоколы диалогов через ICQ и IRC. В результате были получены эмпирические кривые распределения длин слов в текстах, в зависимости от стиля. Эти кривые используются в качестве эталонов при классификации. Программа классифицирует стиль входного текста как: разговорная речи, художественная литература (худло), газетная статья или научная статья. Представляется, что использование этой программы позволит сделать наблюдения над стилистической принадлежностью исследуемых текстов вне зависимости от обсуждаемых тем, в то время как при чтении текста человеком обсуждаемая тема часто играет решающую роль при определении функционального стиля.

Лингвоанализатор (программа Д.В.Хмелева) – это онлайн-версия программы математического анализа структуры текста. Целью анализа является определение близости любого из предлагаемых пользователем текстов к одному из авторских эталонов, определенных заранее. (*Авторский эталон* - это набор текстов данного автора, взятый из ресурсов «Русской Фантастики»). Программа анализирует входной текст и выдает имена трех писателей, которые могли бы быть его наиболее вероятными авторами. Кроме этого, программа находит три произведения каждого из авторов, которые наиболее близки данному тексту. Целью анализа является определение близости любого из предлагаемых пользователем Интернета текста к одному из авторских эталонов, определенных заранее. Разумеется, мы не полагаем, что каждый текст профессионалов в социальной сети имеет смысл соотносить с текстами русских фантастов, однако в силу некоторой единой технократической заинтересованности и общего технократического дискурса вполне может случиться так, что именно русские фантастические тексты служат одной из основ формирования прецедентной базы речевого продукта пользователей профессиональной социальной сети.

Обзор подготовлен Д.В. Колесовой, А.В.Голубевой