МОДЕЛИРОВАНИЕ ПРИМЕНЕНИЯ КОРПУСНЫХ МЕТОДОВ ДЛЯ ЛОКАЛЬНЫХ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

Баранов А.Г. *(baranov@ttrans.ru), ООО «Технотранс»*

Описываются материалы, облегчающие выполнение некоторых задач лингвистического анализа художественных и специальных текстов на русском языке в учебных или научных целях: частотного, морфологического, семантического и дистрибутивного.

1. Предварительные замечания

В настоящее время имеется большое количество различных ресурсов и программ, предназначенных для решения тех или иных задач лингвистического анализа, в том числе и применительно к русскому языку. В практике же лингвистических исследований, особенно выполняемых студентами и аспирантами гуманитарных специальностей или отдельными лингвистами, не имеющими специальной подготовки в области информационных технологий, «рутинные» операции лингвистического анализа автоматизированы достаточно слабо. К таким операциям могут быть отнесены:

– создание картотек / конкордансов словоупотреблений с указанием частотности;

– морфологический разбор словоупотреблений и построение частотного списка лексем (а не словоформ) для исследуемого текста;

– выделение грамматически связанных сочетаний слов, расположенных как контактно, так и дистантно.

Каждая из этих задач так или иначе решается при создании корпусов, именно поэтому мы говорим о моделировании применения корпусных методов без создания представительного корпуса текстов языка (или подъязыка).

Для решения упомянутых задач существует разнообразный инструментарий, достаточно указать две коллекции ссылок на соответствующие ресурсы – С.В. Логичева (http://www.rvb.ru/soft/catalogue/index.html) или ассоциации ACL (http://registry.dfki.de). В то же самое время при использовании распространяемого программного обеспечения филологом, не имеющим специальной подготовки в области информационных технологий, возникает ряд проблем:

1) сложность использования результатов обработки данных одной программой в качестве входных данных для другой (используя бесплатные ресурсы, достаточно просто получить частотный список словоформ и таблицу с вариантами морфологического разбора отдельных словоформ, однако построение на их основе частотного списка лексем потребует некоторых навыков работы, например, в реляционных базах данных);

2) неочевидность для филолога технологии работы с морфологическими анализаторами, в частности с парсером mystem компании «Яндекс» (http://company.yandex.ru/technology/mystem) или продуктами группы АОТ (http://aot.ru);

3) знакомство с возможностями поиска, предлагаемого, в частности, на сайте Национального корпуса русского языка (далее – НКРЯ), вызывает естественное желание применить соответствующие методы к своему материалу, заведомо не вошедшему в корпус. Отсутствие инструментов для аналогичной разметки самостоятельно исследуемых текстов не позволяет этого сделать (как не позволяет в полной мере сопоставить результаты анализа своего материала с данными корпуса).

Представляется, однако, возможным применение методов, которые условно могут быть названы корпусными, для локальных лингвистических исследований на материале текстов, не вошедших пока ни в один из разрабатываемых корпусов русского языка. Под локальными исследованиями имеются в виду курсовые и дипломные работы, диссертационные и иные исследования, выполняемые лингвистами самостоятельно, без привлечения специалистов по информационным технологиям. Для этого необходимо выбрать некоторую среду, обеспечивающую целостность данных и не требующую специальных навыков программирования.

Опыт решения описанных задач подтверждает, что в качестве такой среды может быть использован продукт MS Access, входящий в состав пакета MS Office. Для его освоения

достаточно навыков работы в MS Excel, кроме того, изучение MS Access входит в состав курсов по информационным технологиям на некоторых гуманитарных факультетах.

2. Общая характеристика предлагаемой методики

В MS Access реализованы широкие возможности обмена данными с внешними приложениями, анализа и преобразований строковых (текстовых) данных, поддержания единства данных, наследования свойств и т.д. Для анализа текста требуется (если говорить упрощенно) только представление текста в виде двух таблиц с двумя колонками (полями):

1) <*словоупотребление*> – <*идентификатор контекста словоупотребления*>;
2) <*словоупотребление*> – <*вариант морфологического разбора*>.

Таблица (вернее, набор таблиц) первого типа формируется с помощью программы xMarkup С.В. Логичева (http://www.rvb.ru/soft/index.html). Источником формирования таблицы второго типа служат результаты морфологического разбора, выполняемые парсером mystem.

Использование MS Access и указанных программ, распространяемых их авторами бесплатно (на основе пользовательского соглашения) позволяет в рамках одной программной среды выполнять с различной степенью автоматизации ряд операций, регулярно повторяющихся при лингвистическом анализе текста:

– автоматическое разделение текста на абзацы, абзацев – на предложения, а предложений – на слова (словоупотребления);
– автоматическое получение частотного списка словоупотреблений;
– автоматизированная лемматизация словоупотреблений (получение списков <*словоупотребление*> – <*лексема*>; <*лексема*> – <*часть речи*>);
– автоматизируемое присвоение семантических категорий лексемам или выделение лексико-семантических вариантов одной лексемы;
– автоматическое построение картотек и конкордансов для словоупотреблений, лексем, слов одной части речи или семантической категории;
– автоматизированное выделение грамматически связанных сочетаний слов, расположенных как контактно, так и дистантно.

Определение *автоматический* означает, что программа реализует указанную функцию без участия человека, определение *автоматизированный* подразумевает бо́льшую или меньшую степень ручной работы, а *автоматизируемый* указывает на то, что эта функция может быть автоматизирована (при наличии базы данных, в которой лексемам уже приписаны те или иные семантические категории).

Подготовленные автором статьи материалы для выполнения указанной работы выложены на сайте http://minicorpus.narod.ru для свободного использования.

3. Материалы, предлагаемые на сайте

На указанном сайте представлены следующие материалы:

– описание методики работы;
– файлы правил обработки (с расширением .par) для использования программы xMarkup;
– исполняемые (пакетные) файлы (с расширением .bat) и файл спецификации экспорта (export.ini);
– файл базы данных (БД) для работы (minicorpus.mdb);
– файл БД с образцом разбора короткого рассказа (minicorpus_1.mdb);

3.1. Описание методики работы

В описании методики работы указываются ссылки для скачивания программ xMarkup и mystem, даются указания по использованию рабочих файлов комплекса (.par, .bat и .ini), подробно описывается структура базы данных MS Access и способы работы с ней.

3.2. Файлы правил обработки

Все файлы правил обработки предусматривают работу с файлами в формате .txt с кодировкой ANSI и предназначены для преобразования информации в исходном файле в таблицы, записываемые в выходные текстовые файлы с разделителем в виде табуляции или точки запятой. Для достижения поставленных задач на сайте выложены следующие файлы:

– файл для предварительной подготовки к работе файлов в формате .txt, размещаемых в библиотеке Мошкова (www.lib.ru): удаление лишних пробелов между словами, удаление знаков возврата каретки, разрывающих строку в пределах абзаца, удаление лишних пробелов, указывающих на абзацный доступ;

– файл для нумерации абзацев в исходном тексте (для получения таблицы *<порядковый номер абзаца в тексте> – <текст абзаца>*);

– файл для нумерации предложений в исходном тексте (для получения таблицы *<порядковый номер абзаца в тексте> – <порядковый номер предложения в тексте> – <текст предложения>*);

– файл для нумерации слов (словоупотреблений) в исходном тексте (для получения таблицы *<порядковый номер предложения в тексте> – <порядковый номер слова в предложении> – <слово>*);

– два файла для последовательной обработки файла с результатами морфологического разбора, выполненного программой mystem с учетом всех вариантов разбора омографов, предлагаемых mystem (для получения таблицы *<уникальный номер словоформы в БД> – <лексема со словоклассифицирующими характеристиками>$_1$ – <лексема со словоклассифицирующими характеристиками>$_2$ – <лексема со словоклассифицирующими характеристиками>$_n$*).

3.3. Исполняемые файлы:

– файл segment_text.bat, запускающий программу xMarkup и последовательно выполняющий нумерацию абзацев, предложений и слов в исходном тексте с созданием соответствующих выходных текстовых файлов, готовых для импорта информации в БД;

– файл lemmatize.bat, последовательно запускающий программу mystem для морфологического разбора слов текста из списка уникальных словоупотреблений, сформированного в БД, а затем программу xMarkup, для преобразования файла с результатами морфологического разбора в вид, удобный для загрузки обратно в БД.

3.4. Файл базы данных

Файл базы данных содержит таблицы, запросы, макросы, несколько подчиненных и одну главную форму (загружаемую при открытии базы) для выполнения всех этапов лингвистического анализа по предлагаемой методике. Назначение таблиц, запросов и макросов дано в описании методики работы.

В базе использованы таблицы следующих типов:

– таблицы связи с несколькими источниками данных – текстовыми файлами;

– обновляемые таблицы, в которые записываются данные об анализируемом тексте;

– перезаписываемые таблицы, служащие для обновления данных о частоте словоформ, лексем и количестве вариантов морфологического разбора;

– таблицы-справочники, в которых хранятся текстовые метаданные, морфологические и семантические категории.

Использование в базе практически всех видов запросов, возможных в MS Access (запросов на создание таблиц, на выборку, на поиск повторений, на добавление, обновление и удаление записей, а также на объединение) делает возможным использование базы для самостоятельного изучения MS Access или в преподавании филологам информационных технологий.

4. Использование комплекса minicorpus для лингвистических исследований

4.1. Вкладка «Кнопки управления»

На вкладке размещены кнопки управления для вызова исполняемых файлов, запросов и макросов, объединяющих несколько запросов, описана типовая последовательность работы с базой.

4.2. Вкладка «Лемматизация»

Вкладка предназначена для лемматизации тех словоформ, для которых парсер mystem предложил более одного варианта морфологического разбора. Имеет окно с перечнем словоформ, представляющее собой таблицу *(<словоформа> – < количество вариантов морфологического разбора> – <частота употребления словоформы в тексте>*) с индикатором

развертывания таблицы с вариантами морфологического разбора (вида “лексема = индекс части речи”) для конкретной лексемы, каждый из которых, в свою очередь, имеет индикатор развертывания для вызова контекстов употребления словоформы.

Таблица с контекстами служит для следующих целей:

– просмотр контекстов для выбора правильного варианта лемматизации;

– указание правильного варианта лемматизации для каждого словоупотребления в случае, если в пределах одного текста встречаются грамматические омонимы.

Например, в тексте встретилось два глагола *стекла* и четыре существительных *стекла*. В этом случае во вложенной таблице контекстов достаточно указать правильный вариант разбора для обоих глаголов (“стекать=V”), и один раз поставить отметку напротив варианта разбора “стекло=S,муж,неод” во вложенной таблице вариантов разбора.

Словоформа в поле контекстов выделена верхним регистром; это не совсем удобно при просматривании контекстов употребления коротких слов, например, союза “и” – верхним регистром будут выделены все буквы “и” в предложении, однако при просмотре контекстов полнозначных слов длиной более четырех букв это неудобство пропадает.

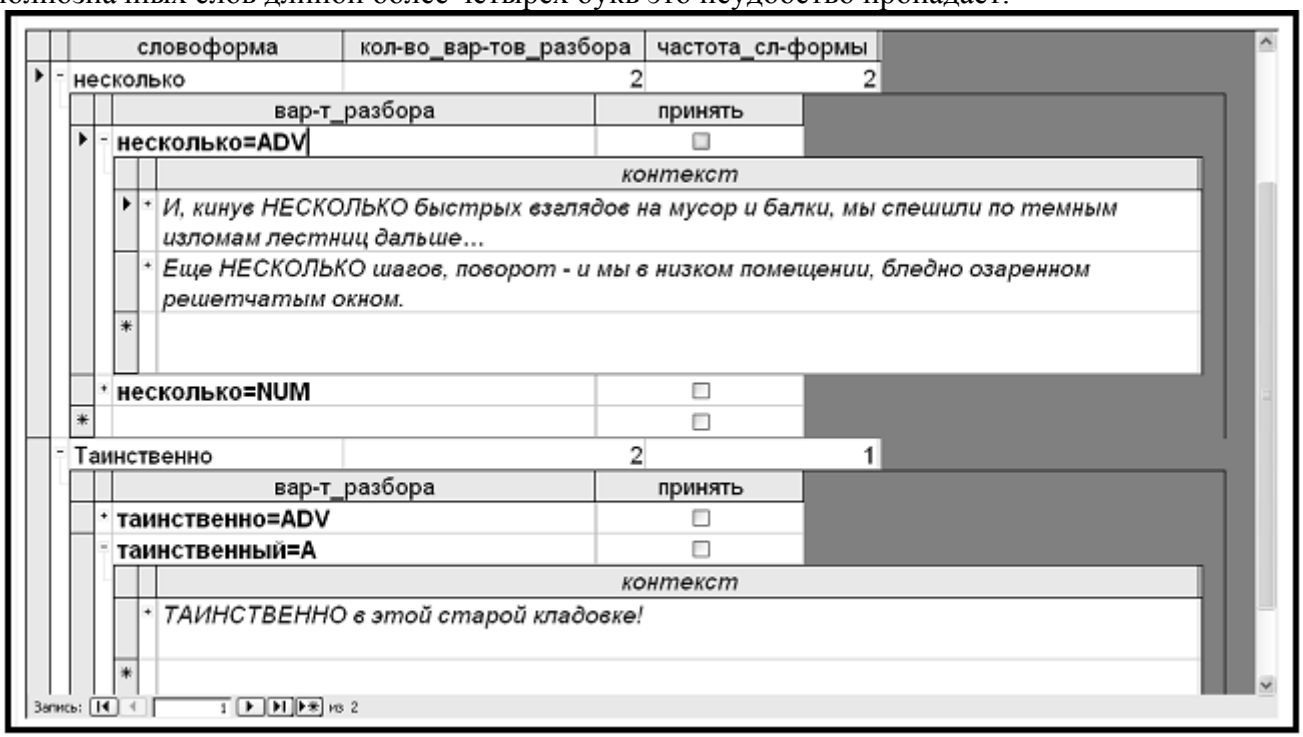

Рис. 1. Окно вкладки «Лемматизация»

4.3. Вкладка «Категоризация»

Вкладка предназначена для присвоения произвольных категорий лексемам, составляющим словник текста. Характер присваиваемых категорий зависит от целей исследования – это может быть гипероним, условное наименование тематической группы, лексико-семантической группы, функционально-семантического поля, интегральной семы / семантического множителя, концепта и т.д. При необходимости присвоения лексико-семантической информации всем лексемам текста можно ориентироваться на перечень лексико-семантических помет, представленных на странице, описывающей семантическую разметку НКРЯ. Вкладка имеет два окна:

– окно со списком категорий (состав которых можно пополнять в процессе работы) – список категорий имеет индикатор развертывания для просмотра слов, которым присвоена та или иная категория;

– окно со списком лексем для присвоения им категорий с возможностью просмотра контекстов употребления лексемы.

4.4. Вкладка «Контексты по категориям»

Вкладка позволяет увидеть выполненную группировку лексем по категориям с возможностью просмотра контекстов употребления лексем. Она может служить для

верификации исследовательской гипотезы о распределении лексем той или иной категории в тексте, а также быть источником формирования рядов слов, соотнесенных по тому или иному критерию.

4.5. Вкладка «Карточка слова»

Вкладка сводит воедино всю хранящуюся в базе информацию о конкретном словоупотреблении:

- контексты употребления с указанием источника каждого контекста; количество употреблений;
- лемма с ее частеречной принадлежностью и приписанной ей семантической категорией;
- количество употреблений всех словоформ данной лексемы.

4.6. Вкладка «Сочетаемость»

Вкладка предназначена для моделирования поиска сочетаний слов в пределах предложений текста по нескольким категориям (ограниченно имитируя функциональность соответствующего поиска в НКРЯ). Использование стандартных средств фильтрации и сортировки, предусмотренных в MS Access, позволяет искать пары слов по следующим параметрам (по отдельности или одновременно):

- лексема первого и / или второго слова;
- часть речи первого и / или второго слова;
- семантическая категория первого и / или второго слова;
- расстояние между словами в пределах предложения.

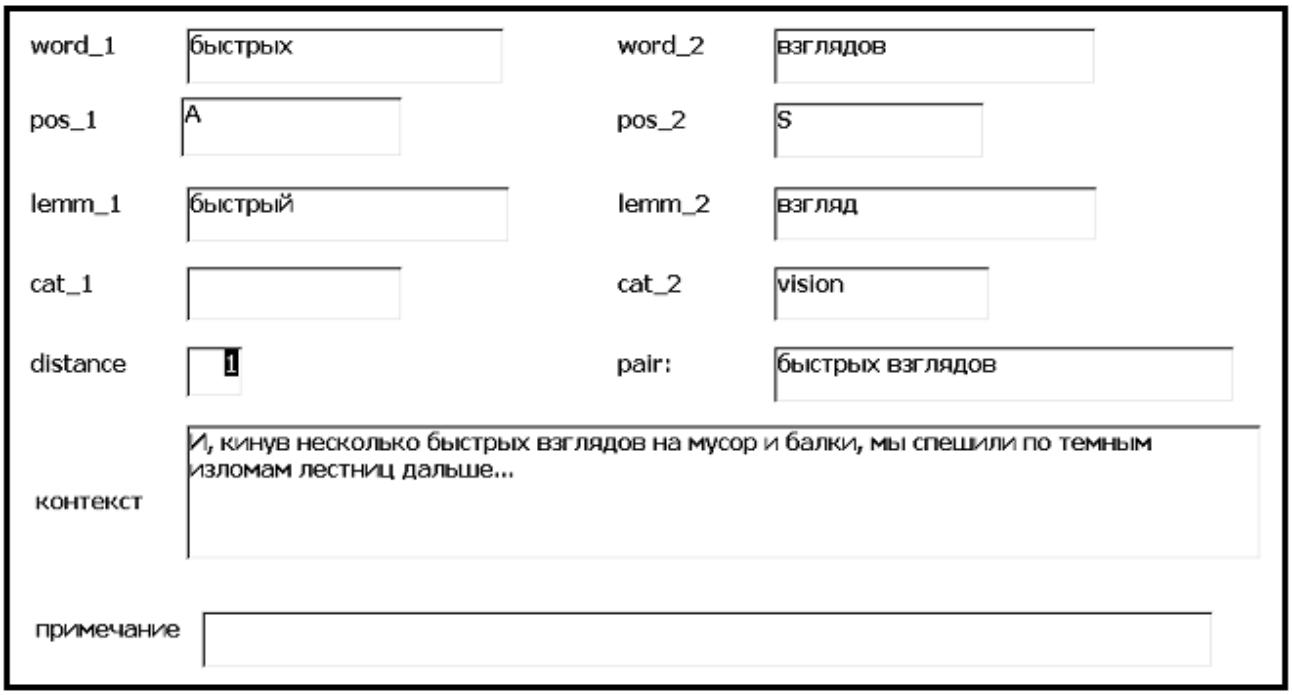

Рис. 2. Окно вкладки «Сочетаемость»

5. Выводы

Представляется, что описанный программно-методический комплекс может быть использован по меньшей мере в трех направлениях:

- автоматизация решения ряда задач лингвистического анализа в рамках одной программной среды;
- демонстрация принципов (теоретических и технических) формирования корпусов национальных языков при изучении филологами информационных технологий;
- демонстрация возможностей MS Access как для лингвистического анализа, так и для организации результатов своих исследований в рамках базы данных.