Сиренко А.В.
9 декабря 2008 г.

Доклад по теме

«Информационная технология лингвокультурного тезауруса русского языка»

1. Вступление.

Я, Сиренко Александр Викторович, окончил кафедру ИУ5 МГТУ им. Баумана в 2007-м году и в настоящий момент, под руководством Юрия Николаевича Филипповича выполняю исследовательскую работу, называющуюся «Информационная технология лингвокультурного тезауруса русского языка».

В связи с постоянным ростом объемов информации важнейшее значение приобретают возможности сбора, хранения, обработки и анализа данных. При этом большая часть информации, подлежащей обработке и анализу, представлена в естественно-языковой форме. Вопросами, связанными с определением смысловой нагрузки текста, тем, каким образом человек осознает текст (не важно, в зрительной или звуковой форме) и как создает его сам, занимается психолингвистика.

В ранней психолингвистике поведение человека представлялось организованной системой реакций на внешние стимулы. В дальнейшем появилась теория порождающих грамматик Хомского. Теория порождающих грамматик рассматривает формирование и восприятие речи как процесс преобразования смысловых конструкций на основе правил [Слобин, Грин, 2006].

Сложность изучения человеческого мышления состоит в том, что приходится делать выводы по косвенной информации.

На сегодня ни один из подходов полностью не формализовал речевую активность, однако их достижения активно используются в информационном поиске и технологиях обучения.

Вполне понятно, что для обработки текстов автоматизированными системами текст важен скорее не как форма представления знаний, а с точки зрения передаваемого им смысла. Это верно и с позиции фиксации знаний о мире какой-либо исследуемой группы, например носителей русского языка. Поэтому, большое значение приобретает новый тип словарей – идеографический. Идеографические словари группируют элементы не по написанию, а по какому-либо тематическому признаку [Википедия, 2008]. Примером подобного словаря может служить так называемый тезаурус, включающий множество смысловых единиц некоторого языка с заданной на нём системой семантических или смысловых отношений[Яндекс словари, 2008].

Целью проводимого исследования является моделирование языкового сознания носителя русского языка, а также фиксация в лексикографической форме его знаний о мире.

2. Языковое сознание. Ассоциативный и когнитивный эксперимент.

В работе языкового сознания человека условно можно выделить два режима работы – активный, осуществляющий переход от знака к смыслу и обратный ему пассивный. [Караулов, Филиппович 2005, с.5-6]

Информация об этих режимах работы берется из 2-х ранее проведенных экспериментов, соответствующих активному и пассивному режиму работы языкового сознания. Это ассоциативный и когнитивный эксперимент.

Ассоциативный эксперимент проводился путем опроса множества респондентов, в каждом из которых человек должен был назвать первую пришедшую в голову ассоциацию, называемую реакцией, к слову, играющему роль стимула [Черкасова, 2004]. При этом фиксировался пол, возраст и профессия респондента, поскольку эти характеристики значительно влияют на результат. Важнейший параметр – количество повторов ассоциации, отражает ее устойчивость у опрашиваемых. Проведение ассоциативного эксперимента связано со множеством сложностей, так как это длительный и масштабный эксперимент, а ассоциации респондентов со временем меняются.

В когнитивном эксперименте респонденту предоставляется некая формула смысла в словесной форме и предлагается ответить на нее словом, отражающим этот смысл. Это похоже на правила игры «Кроссворд», поэтому данные когнитивного эксперимента были получены из материалов кроссвордов.

Ассоциативный и когнитивный эксперименты показывают разные стороны одного и того же объекта – языковой картины мира носителя языка. Поэтому естественно предположить наличие между ними системных связей.

3. Моделирование языкового сознания.

Моделирование языкового сознания предполагает анализ данных обоих экспериментов. Но, человек, в процессе осмысления текста, пользуется не только той информацией, которая непосредственно в нем присутствует, но и множеством ранее накопленных им знаний, может улавливать скрытый подтекст, иносказания и аналогии. Тезаурус должен иметь возможность расширения, добавления новых словарей и классификаций.

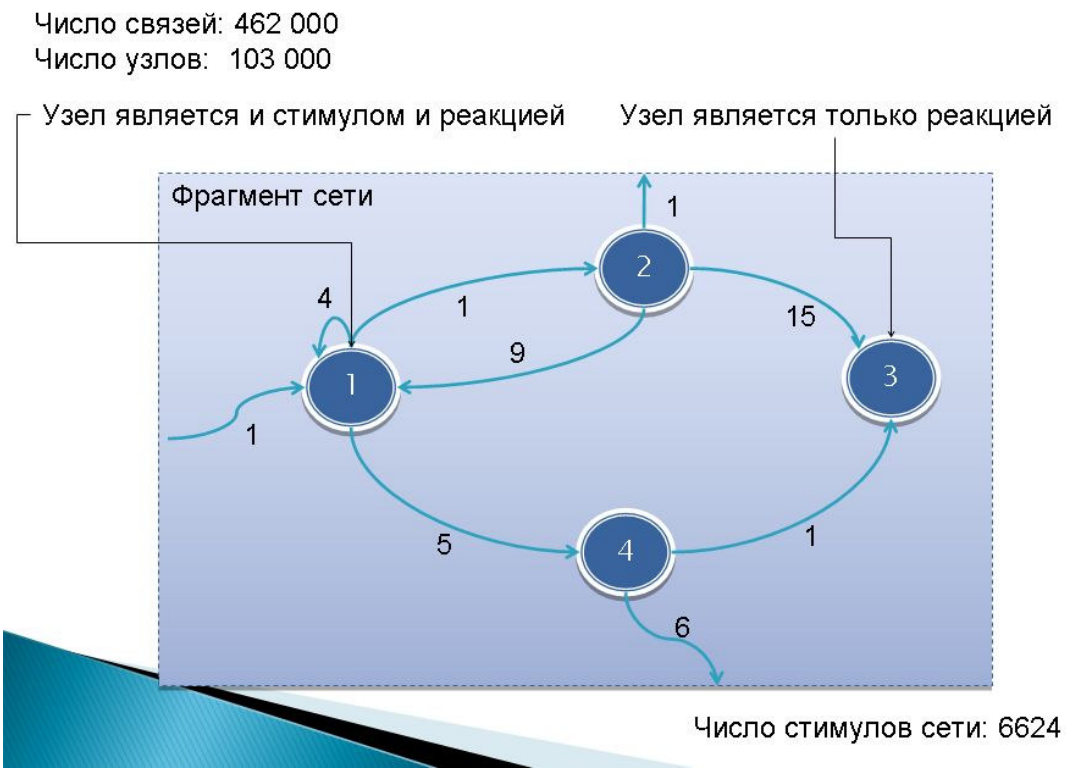

Рисунок 1 Ассоциативная сеть

Центральное место в функционировании тезауруса занимает поиск цепочек в ассоциативной сети. Логично предположить, что в сети присутствует отношение транзитивности, то есть если узел 1 является стимулом для узла 2, а узел 2 является стимулом узла 3, узлы 1 и 3 определенным образом связаны в сознании человека. Поиск цепочек может производиться от формулы смысла к знаку, который эта формула обозначает. Эта задача похожа на задачу отгадывания кроссворда. Причем, если поиск путей при известном конечном знаке является тривиальной задачей, то поиск неизвестного конечного знака достаточно сложен. Ассоциативная сеть содержит порядка 462 тысяч связей и около 103 тысяч узлов, что дает 5 исходящих связей в среднем для каждого узла. Надо учесть, что большинство узлов сети являются только реакциями и не имеют исходящих связей, а значит, не влияют на рост числа возможных путей в процессе поиска. Число стимулов в сети – 6624. Каждый шаг поиска путей, на котором мы переходим от очередного узла к следующему, увеличивает число результатов в 70 раз. Есть предельная длина путей подлежащих поиску. По экспертной оценке, от любого стимульного узла можно дойти до каждого узла сети путем длиной не более 7-ми шагов. Даже при более жестких ограничениях, число возможных конечных знаков велико и составляет тысячи. Формализовать процесс выбора знака из результатов не представляется возможным по следующим причинам:

1) Запрос пользователя как любая лингвистическая информация обладают нечеткостью и не являются полным.
2) То же может быть сказано об ассоциативной сети.

Кроме того, так как разрабатываемая система предполагается как инструмент исследователя, она должна иметь возможность настройки алгоритмов и позволять пользователю-эксперту влиять на процесс поиска.

Тем не менее, ставится задача сокращения результатов внесением ограничений в процесс поиска, либо в виде определенных фильтров после него. Это могут быть:

1) Ограничение длины путей
2) Всевозможные словники
3) Кластеризация сети и ограничение поиска выбранными кластерами.

К тому же возможно дополнение запроса пользователя с помощью словарей синонимов, словоформ, а также результатов когнитивного эксперимента.

Дополнительными возможностями предполагаются:
1) Формирование детализированных отчетов о результатах поиска.
2) Представление топологии найденных путей в графической форме.

4. Средства разработки. Структура программы.

Рисунок 2 Средства разработки

Система предназначена для широкого круга специалистов, обладающих различными аппаратными ресурсами, и не всегда обладающих навыками, необходимыми для установки и администрирования сложных программных систем.

Поэтому, при выборе средств разработки учитывались критерии:

1) Наличие на рынке СУБД, требующей минимум затрат на установку и администрирование с точки зрения конечного пользователя системы;
2) Наличие доступных средств разработки, справочной информации, распространенность технологии;
3) Невысокая стоимость средств разработки.
4) Отсутствие необходимости покупать лицензии конечным потребителем.

В результате выбор был остановлен на наборе средств на базе технологии Java:

Среда разработки: Java SE Development Kit (JDK) 6;
Средство проектирования: Sun Microsystems NetBeans;
СУБД: H2

СУБД работает во встраиваемом режиме, подключаясь к программе в виде дополнительного модуля в процессе выполнения. При этом она достаточно полнофункциональна.

Разработка ведется по следующим принципам:

1) Возможность смены используемой СУБД.
2) Применение модульной структуры с минимальными связями между компонентами системы.

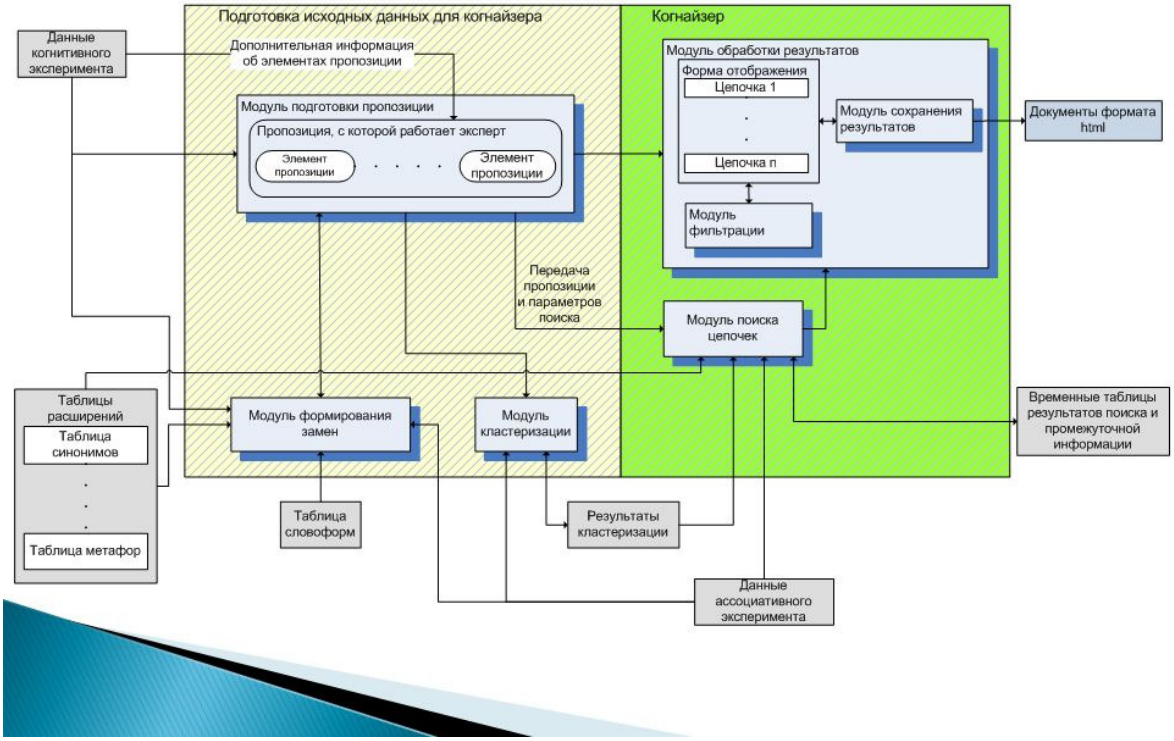

Рисунок 3 Структурная схема тезауруса

На слайде представлено схематичное изображение общей структуры тезауруса. С точки зрения использования тезаурус можно условно разделить на 2 части: подготовку исходных данных – пропозиции и работу когнайзера.
Для простоты, на схеме не представлен модуль взаимодействия с базой данных и модуль ввода данных из внешних файлов.

Модуль подготовки пропозиции позволяет сформировать исходные данные, а именно пропозицию и параметры поиска. В своей работе он использует модуль формирования замен.

Модуль формирования замен привлекает к работе дополнительные словари, а также осуществляет лемматизацию.

На этапе подготовки работы когнайзера, может быть проведена кластеризация сети. Это длительный процесс, поэтому результаты кластеризации сохраняются между сеансами работы и проводить ее каждый раз не обязательно.

После подготовки исходных данных, они передаются в модуль поиска цепочек, работающий без участия пользователя. Результаты передаются в Модуль обработки результатов, а также доступны во временных таблицах базы данных.

Модуль обработки результатов служит для их фильтрации, сортировки и отображения цепочек с учетом их топологии, для чего предполагается показывать их в графической форме.

Модуль сохранения результатов формирует html-документ установленного формата.

5. Подготовка сети

Перед тем, как работать с результатами ассоциативного эксперимента, проводится некоторая их модификация, связанная с количественной характеристикой связей между узлами сети.

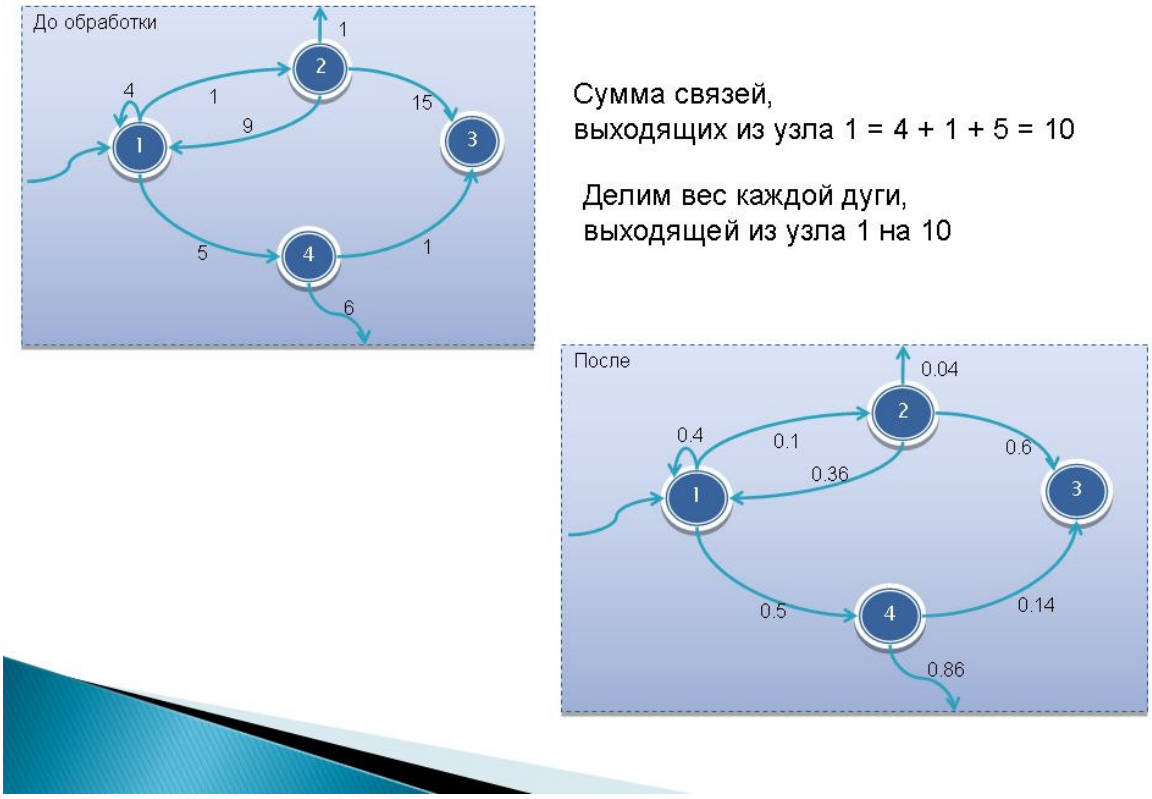

Рисунок 4 Подготовка сети

Число встретившихся ассоциаций узла 1 с узлом 2 недостаточно информативно, если не учитывать общее число ассоциаций, в которых узел 1 является стимулом. Поэтому осуществляется переход от числа повторов ассоциации к ее частоте среди всех ассоциаций узла 1. Таким образом, мы получаем вероятность, с которой, попав в узел 1, в дальнейшем мы перейдем к узлу 2. Будем это считать инвертированным эквивалентом расстояний между узлами. Такая характеристика необходима как для кластеризации, так и для количественной оценки путей.

6. Кластеризация

Кроме использования в когнайзере, кластеризация ассоциативной сети представляет отдельный интерес для лингвистов. Особенностью ассоциативной сети является то ее свойство, что не представляется возможным определение расстояний между всеми элементами сети. Ряд алгоритмов кластеризации требует наличия своего рода координаты узла в общем многомерном пространстве. В ассоциативной сети можно говорить лишь о наличии расстояний между некоторыми узлами, а создание дополнительных связей через свойство транзитивности должно быть обосновано. Был выбран иерархический агрегационный метод кластеризации, в котором на начальном этапе каждый узел сети представляет отдельный кластер. Связи между кластерами повторяют связи между узлами ассоциативной сети.

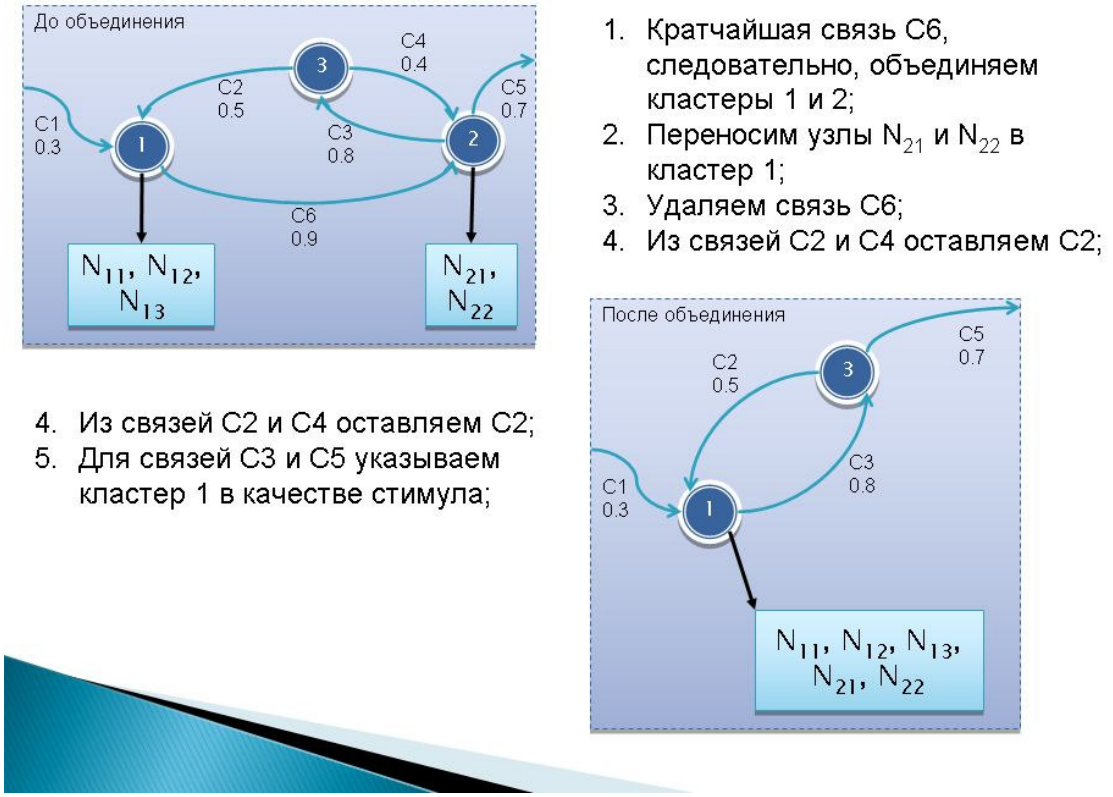

Рисунок 5 Кластеризация

Затем, до выполнения одного из условий окончания: команды пользователя прекратить кластеризацию или достижения определенного числа кластеров, выполняется цикл:

1. Определяются два ближайших кластера по кратчайшей связи между ними. Направленность связи не учитывается. Один из кластеров назовем кластером-стимулом, второй кластером-реакцией.
2. Объединяем содержимое кластеров. Узлы сети, принадлежащие кластеру-реакции, переносятся в кластер-стимул.
3. Объединяем связи кластеров. Связи, относящиеся к кластеру-реакции, переносятся на кластер-стимул. Если существуют две связи, идущие к или от одного внешнего кластера, оставляем связь с минимальным расстоянием. Связи, соединявшие объединяемые кластеры, удаляем.
4. После произведенных действий кластер-реакция перестает существовать. В случае если условия прекращения не выполняются, переходим к следующей итерации.

7. Организация поиска в сети

Что касается алгоритма поиска в сети, к нему предъявляются следующие требования:

1. Максимальная скорость работы.
2. Минимальные затраты памяти.
3. Возможность построения путей после достижения конечной вершины.

В результате был реализован волновой алгоритм. Предположим, нам необходимо найти минимальный путь между узлами 1 и 2.

1. Выполняем поиск всех узлов, непосредственно достижимых из узла 1.
2. Формируем из них множество M[1].
3. Если конечный узел присутствует в M[1] – задача выполнена. Иначе выполняем поиск всех узлов, достижимых из M[1] и формируем из них множество M[2].

Цикл повторяется до достижения одного из условий прекращения работы:

1. Команды пользователя;
2. Получения необходимых результатов;
3. Достижения предельного числа итераций.

Удобством данного алгоритма с точки зрения реализации является возможность создания очередной волны SQL-запросом к СУБД без необходимости обработки данных в программе тезауруса. Это положительно сказывается на простоте и скорости работы. При необходимости, можно исключить кольцевые структуры в найденных путях, сократив число результатов.

8. Заключение

В настоящий момент система находится в стадии создания.

Организован ввод лингвистической информации из внешних файлов, генерация базы словоформ.

Реализованы основные функции модулей: подготовки пропозиции, формирования замен, поиска цепочек, производится отладка модуля кластеризации.

Запуск поиска при полном размере ассоциативной сети показал необходимость совершенствования алгоритма поиска с точки зрения скорости.

Также существует сайт, посвященный проекту, www.tesaurus.ru.

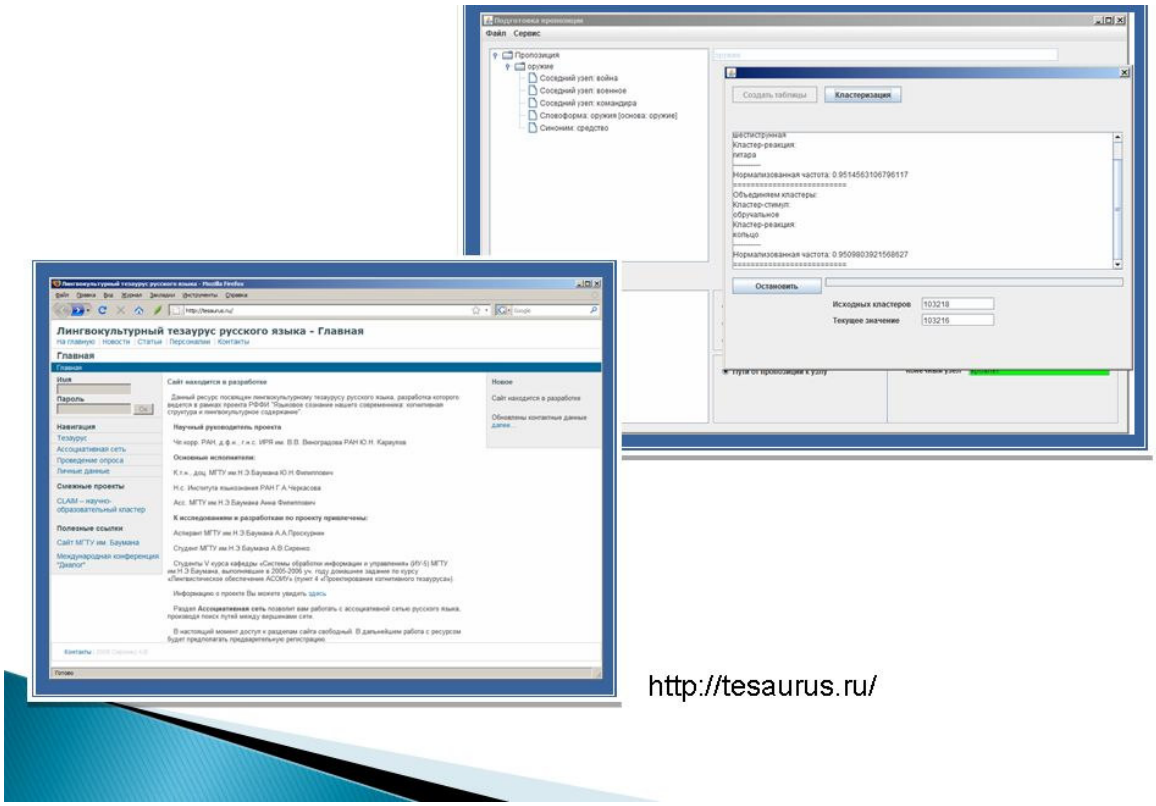

Рисунок 6 Экранные формы

На сайте планируется предоставить возможности по работе с результатами ассоциативного и когнитивного экспериментов, а также размещать материалы по работе.

Список литературы:

Слобин, Грин, 2006	Психолингвистика / Д. Слобин. Психолингвистика. Хомский и психология / Дж. Грин: Пер с англ. / Под общ. ред. и с предисл. А.А. Леонтьева. Изд. 4-е, стереотипное. – М.: КомКнига, 2006. – 352 с.
Википедия, 2008	Википедия [Электронный ресурс] / - Электрон. дан. – М., 2008 – режим доступа: http://ru.wikipedia.org/wiki/Идеографический_словарь
Яндекс словари, 2008	Яндекс словари – Большая Советская Энциклопедия [Электронный ресурс] / - Электрон. дан. – М., 2008 – режим доступа: http://slovari.yandex.ru/dict/bse/article/00078/04200.htm?text=тезаурус
Караулов, Филиппович, 2005	Ю.Н.Караулов, Ю.Н.Филиппович. Лингвокультурологический тезаурус русского языка. –М.: 2005.
Караулов, 2004	Ю.Н.Караулов. Концептография языковой картины мира. Статья 1. Первый этап «восхождения» к образу мира: от элементарных фигур знания к предметно-референтным областям культуры// Проблемы прикладной лингвистики. Выпуск 2. Сборник статей./ Отв. ред. Н.В.Васильева. –М.: «Азбуковник», 2004. – 400с.
Черкасова, 2004	Г.А.Черкасова. Формальная модель ассоциативного исследования.// Проблемы прикладной лингвистики. Выпуск 2. Сборник статей./ Отв. ред. Н.В.Васильева. –М.: «Азбуковник», 2004. – 400с.