А.В.Сиренко

БАЗА ДАННЫХ ЛИНГВОКУЛЬТУРНОГО ТЕЗАУРУСА РУССКОГО ЯЗЫКА

Постановка задачи. Новые типы словарей

Естественный язык долгое время был для лингвистов неприкосновенным объектом изучения, и работы в области языковедения носили описательных характер. Таким образом, язык рассматривался как некий самостоятельно развивающийся объект изучения. Рассматривались различные факты языка, проводилась систематизация, выделение его структур, схем функционирования. С развитием вычислительной техники, постановкой новых задач и накоплением знаний о языках появилась необходимость в лингвистических объектах, имеющих новые свойства. Позволяющих с другой стороны взглянуть на естественный язык. Не только на сложившиеся внешние признаки его функционирования, но и на принципы развития языка. Такими объектами могут быть порождающие синтезаторы, грамматики нового типа и т.д. [Караулов, 1981, с.19]

В чем отличие этих новых лингвистических объектов от прежних? Очевидно, что словари строятся на эмпирической основе. То есть для их составления необходимо обработать некий массив языковых данных, например, в текстовом виде. Если же исходных данных в необходимом количестве нет, то составление словаря по сложившейся методике невозможно. Создание же порождающего механизма, метода лингвистического конструирования могло бы помочь иначе взглянуть на эту задачу. Кроме того, язык не является самостоятельным независимым объектом. Он функционирует и развивается в рамках окружающей его действительности и, с одной стороны, является ее отражением, а с другой, формирует наше представление о ней, так называемое «языковое сознание». Язык в данном случае – инструмент, посредством которого мы можем взглянуть на «сознание», на восприятие мира отдельным субъектом. Очевидно, что язык динамичен, но мы будем рассматривать относительно стабильную форму его существования в виде текстов.

Мы подошли к понятию Языковой Картины Мира. Она состоит из элементов, отражающих знания об окружающем мире (Единицы Знания

о Мире - ЕЗМ). Эти элементы могут представлять собой различные языковые структуры: определения, пословицы, прецедентные тексты. Совокупность таких элементов создает разнородную, подчас противоречивую картину мира. Языковое сознание имеет двойственную структуру, являя собой некий механизм соединения знаний о языке со знанием о мире. Языковое Сознание (далее ЯС) можно представить в виде структуры:

Языковое Сознание = Единица Знаний о Мире + Языковая Единица

В этой зависимости языковое сознание представляет собой когнайзер, в котором происходит постоянное преобразование Единиц Знаний о Мире в Языковые Единицы и наоборот. Задачей, которая ставится при выполнении описываемой работы, в конечном итоге, является моделирование процессов, протекающих в когнайзере, то есть перехода от ЕЗМ к ЯЕ и наоборот. Необходимо выяснить, как этот переход осуществить, возможно ли это и равноценен ли переход в ту, или другую сторону? Работу когнайзера по переходу от слова (знака) к знанию будем называть активным режимом работы, а от знания к знаку – пассивным [Караулов, Филиппович 2005, с.5-6].

Активный и пассивный эксперимент

Ранее был проведен ассоциативный эксперимент. В нем респонденту называлось некое слово (стимул) и его задачей было назвать приходящее при этом в голову слово. Отвечать было необходимо не задумываясь [Черкасова, 2004]. В результате формировалась пара слов «стимул-реакция», которые можно рассматривать как отражение Языкового Сознания, где стимул выступает в роли Языковой Единицы, а реакция в роли Единицы Знаний о Мире. Причем, как показала практика, связь между стимулом и реакцией двунаправлена. Если стимул «слон» вызывает реакцию «хобот», то, как правило, имеет место и обратная связь. Стимул «хобот», скорее всего, имеет среди множества своих слов-реакций «слон». Для этого необходимо проведение достаточно полного опроса. Таким образом, здесь не важно, что мы выберем за Языковую Единицу, а что за Единицу Знания о Мире. Ассоциативный эксперимент отражает активный режим работы когнайзера.

Для рассмотрения пассивного режима работы когнайзера используются материалы кроссвордов. В кроссворде человеку предлагается смысл разгадываемого слова в той или иной форме (дефиниции, синонима и т.д.). Этот смысл, чаще всего в соединении с самим разгадываемым словом, представляет собой не что иное, как элементарное знание о мире, а именно о понятии, которое необходимо разгадать.

Например:

1) Оттенок на фоне какого-нибудь цвета – ОТЛИВ (Дескрипция).
2) Применяют, когда штурм не удался – ОСАДА (Антоним).
3) Очень мелкие цветные бусинки – БИСЕР (Дескрипция).

Таким образом, в материалы кроссвордов представляют в своеобразной форме Языковую Картину Мира. И, если предположения о пассивном и активном режимах работы когнайзера верны, то результаты разгадывания кроссворда должны найти свое подкрепление и в активном режиме работы когнайзера – ассоциативном эксперименте.

Знания об окружающем мире можно представить в виде так называемых Фигур Знания [Караулов, 2004]. Это элементарные когнитивные единицы, которыми мы оперируем в повседневной жизни. Они содержат как интенсиональные характеристики, а именно Знак, Способ, Смысл, описанные выше, так и экстенсиональные, отражающие положение Фигуры знания в общем пространстве знаний. Такими параметрами являются когнитивная область и функция. Схематичное представление фигуры знания можно увидеть на рис.1

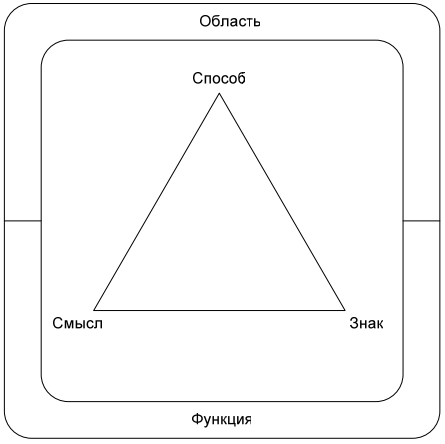

Рис.1 Фигура Знания

Все знания мы можем разбить на категории, отражающие определенные стороны окружающего мира, на области. Знак может быть связан с Областью многообразными ассоциативными связями. К чему относится, с чем связано, частью чего является – вот далеко не полный список возможных вариантов. Строго говоря, составление списка возможных областей и разбиение фигур знания по ним - процесс субъек-

тивный, и каждый исследователь сделает это по-своему. В процессе наполнения базы данных эксперимента постепенно формируется множество возможных областей. При обработке достаточно большого количества понятий это множество перестает расти. Очевидно, одно и то же понятие может принадлежать нескольким областям, что свидетельствует о том, что области могут частично перекрываться, и разбиение на них весьма условно.

В активном режиме когнайзер переходит от Знака к Смыслу. При этом определяется Способ, которым Знак раскрывается через Смысл. В пассивном же режиме изначально имеется Смысл и, таким образом заданный, Способ. Имея эти два параметра, когнайзер переходит к Знаку. В обоих режимах экстенсионал фигуры знания определяется после интенсионала. Казалось бы, связи между экстенсиональными и интенсиональными параметрами быть не должно, однако, некоторые зависимости можно проследить. Например, зависимости между Областью и Способом. В некоторых областях очевидно предпочтение определенным Способам выражения смысла. Например, Способ «Элемент множества» очень популярен в Географии. Также среди фигур знания, принадлежащих области «Авиация», многие имеют Функцию «Ретушь», так как знания в этой области не всегда важны в обыденной жизни. В области «Быт» наблюдаем обратную тенденцию.

Попробуем на примере проследить связь между активным и пассивным режимом в связи знака с областью. Важно заметить, что в ассоциативно-вербальной сети присутствует столь великое множество пар стимулов-реакций, что если мы будем искать связь между словами через сколь угодное множество промежуточных переходов, то на практике можем столкнуться с перемещением в огромном числе ветвлений и цепочек, где практически между любыми словами находится путь. Наша же задача стоит в нахождении минимальных путей. Рассмотрим следующую единицу знания пассивного режима когнайзера:

1. Последняя буква кириллицы. 2. Фрейм. 3. ИЖИЦА. 4. Язык. 5. Рцт.

Рассмотрим обратный ассоциативный словарь, организованный от реакции к стимулу, включающий в себя все содержимое ассоциативного словаря. Ищем пары, в которых ИЖИЦА является реакцией. Таких пар не нашлось, значит, из смысловой формулы подбираем возможный аналог, это знак БУКВА. В обратном ассоциативном словаре БУКВА является реакцией на множество стимулов, среди которых есть стимул «язык», что подтверждает связь знака с областью в активном и пассивном режиме.

Исходные данные. Первичная обработка.

Исходные данные для эксперимента представлены, как уже описывалось выше, данными ассоциативного эксперимента, с одной стороны, и материалами кроссвордов, с другой.

Ассоциативный эксперимент существует в виде базы данных в формате Paradox. Материалы кроссвордов же, было необходимо привести к форме, удобной для последующего использования.

Для этого сначала было необходимо выбрать технологическую базу выполнения эксперимента, то есть выбрать среду разработки программного обеспечения и систему управления базой данных, поскольку работа с информацией предполагала выполнение различных запросов. Выбор пал на использование Borland Delphi , входящий в состав пакета Borland Developer Studio 2006, так как он удобен для быстрой разработки приложений и имеет широкий набор средств для обращения к различным СУБД (Системам Управления Базами Данных). Базу данных предполагалось строить на основе MS Access. Причин для этого несколько:

1) MS Access установлен на подавляющем большинстве персональных компьютеров, что позволит работать с базой данных практически везде, то есть любому исследователю.

2) База данных Access представляет собой отдельный файл формата mdb и при переносе программного обеспечения эксперимента этот файл также может быть перемещен, без необходимости проведения каких-либо наладочных работ. Это дает мобильность.

3) Эта СУБД входит в пакет MS Office, поэтому имеет возможность импорта и экспорта данных в различном виде. Например, импорта данных в виде файла MS Excel, или экспорта базы данных в СУБД MS SQL Server, что важно как предположительный вариант модернизации базы данных в случае необходимости улучшения ее характеристик.

4) Access имеет собственные средства для обработки данных, выполнения запросов, генерации отчетов и выборок.

После выбора СУБД исходная таблица, представленная в текстовом документе MS Word, была сперва переведена в MS Excel, а затем перенесена в базу данных Access.

Таблица имеет следующий вид:

Код	Знак	Смысл	Область	Способ	Функция

Будем именовать эту таблицу далее таблицей фигур знания.

Необходимо внести пояснения по ее составу.

КОД имеет числовой формат и служит для идентификации записей.

ЗНАК содержит в текстовом виде слово, которое в кроссворде требуется отгадать.

СМЫСЛ представляет собой текст, на который ориентируется отгадывающий человек, то есть разъяснение понятия, представленной в той или иной форме.

ОБЛАСТЬ является перечислением тех областей знания, к которым относится данное понятие. Каждое понятие принадлежит как минимум одной области. Области в поле перечислены через запятую, пробелы либо другие небуквенные символы.

СПОСОБ представляет собой структуру смысла, через который загадано слово. Это может быть дефиниция, прецедентный текст, метафора или множество других способов.

ФУНКЦИЯ определяет так называемую «ценность» понятия. С точки зрения ценности различают знание-рецепт и знание-ретушь [Караулов, Филиппович, 2005, с.14]. Знание-рецепт несет в себе важную, существенную для исследователя информацию, необходимую ему в повседневной жизни. Знание-ретушь содержит дополнительные, менее существенные данные, которые в некоторой степени можно не рассматривать при анализе текста. Исследователь, определяя, к какому типу знание относится, руководствуется собственными представлениями о необходимости этого знания для него самого, для его понимания окружающего мира.

Рассмотренная выше таблица имеет ряд недостатков. Например, поле СПОСОБ может принимать одно из некоторого набора возможных значений. На этапе разработки количество вариантов способов не известны. Запись в поле СПОСОБ наименования способа ведет не только к избыточности объемов хранимых в памяти данных, но и является потенциальным источником ошибок. Поэтому возможные области выделяются в отдельную таблицу, а в первоначальной таблице записывается лишь код записи таблицы способов.

Аналогично можно поступить с полем ФУНКЦИЯ.

Особенностью областей с позиции строения базы данных является то, что, во-первых, области образуют некоторое конечное множество, во-вторых, любое понятие может принадлежать одной либо нескольким областям. Поэтому использование той же схемы выделения отдельных таблиц, как в случае с полями СПОСОБ и ФУНКЦИЯ, невозможно. Была смоделирована связь многие-ко-многим с помощью промежуточной

таблицы, в которой каждая запись представляет собой сочетание кода записи таблицы данных и кода записи таблицы областей. Таким образом, одной записи исходной таблицы может быть поставлено в соответствие несколько областей.

Структура таблиц. Их заполнение.

Итоговая схема БД представлена на рисунке 1.

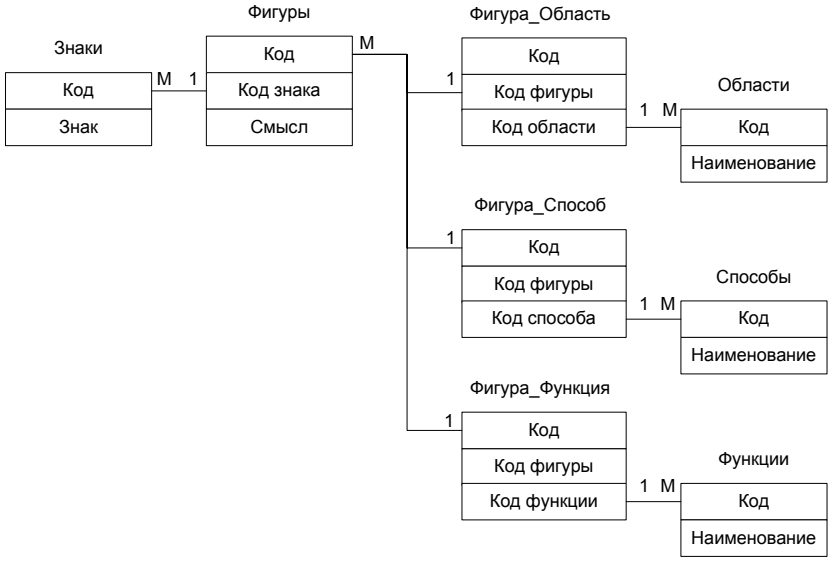

Рис.1

Заполнение таблиц Знаки, Фигуры, Фигура_Область, Фигура_Способ, Фигура_Функция, Области, Способы по исходной таблице не может выполняться полностью в автоматическом режиме. Допустим, в поле Область исходной таблицы мы встречаем обозначение, которое не присутствует среди имеющихся в таблице областей наименований. Либо нам необходимо добавить новую область и создать связь, либо это одна из имеющихся областей, просто иначе обозначенная, в наименовании которой допущена ошибка. Поиск же в БД имеющихся областей и предложение их пользователю можно автоматизировать.

Для автоматизации разбора исходной таблицы была разработана форма, изображенная на рисунке 2.

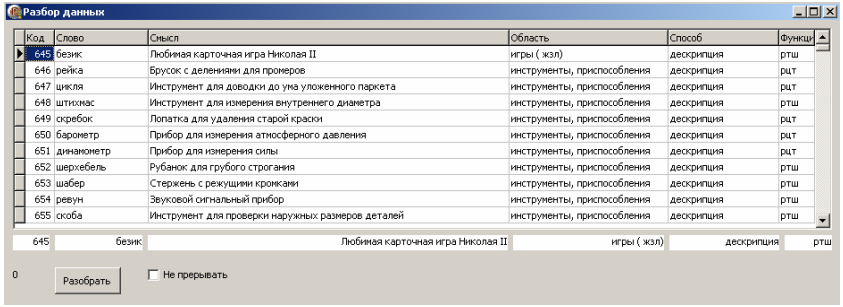

Рис.2

В таблице представлены еще не обработанные записи. Ниже в текстовых полях представлен состав активной записи, которая будет обработана при нажатии кнопки «Разобрать».

Разбор очередной записи исходной таблицы и заполнение результирующих таблиц представляет собой последовательность из следующих этапов:

1) Если среди Знаков уже есть содержимое поля «Знак» исходной таблицы, Фигура знания связывается с имеющимся знаком. Иначе создается новый Знак.

2) Затем среди Областей осуществляется поиск элементов, наименования которых встречаются в поле «Область» исходной таблицы. В случае если в записи исходной таблицы всего одна область, и наблюдается полное совпадение с какой-либо областью таблицы областей, связь устанавливается автоматически, без участия пользователя. Иначе предлагается вручную выбрать необходимые области, в том числе с возможностью создать новые. Пример такого диалога можно увидеть на рисунке 3.

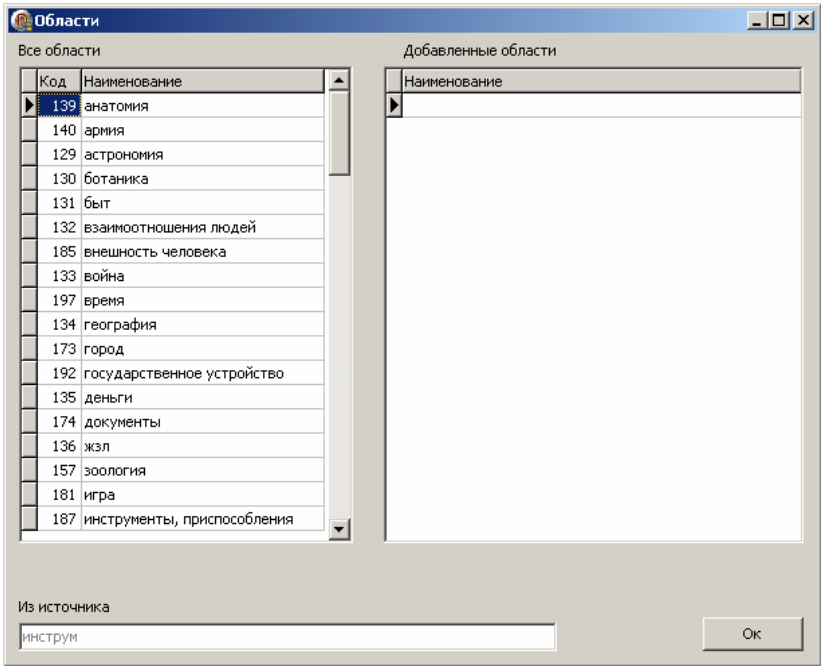

Рис.2

3) Производится разбор поля Способ, аналогично полю Область.

Стоит заметить, что если осуществить разбор поля автоматически не удается, пользователю предлагается выбрать Способ из имеющихся в базе данных, либо ввести новый.

4) Заполнение поля Функция.

Обработав таким образом все записи исходной таблицы, переходим к схеме данных, представленной на рис.1.

Главная форма программного обеспечения позволяет просматривать эти таблицы в удобной форме, а также переходить к редактированию. Ее внешний вид показан на рис.3.

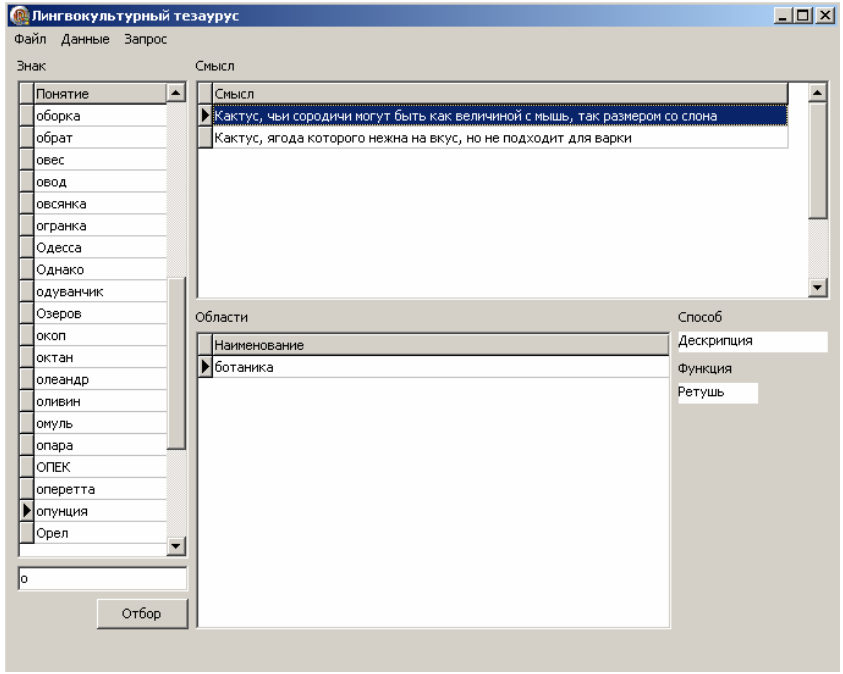

Рис.4

В таблице «Понятие» перечислены наименования всех понятий, представленных в БД, согласно одноименной таблице. Таблица «Смысл» перечисляет все записи таблицы «Фигуры_Знания», которые соответствуют текущей записи в таблице понятий.

Таблица «Области» и текстовое поле «Способ» отображают Области и Способ, связанные с текущей фигурой знания.

Поиск по базе данных

Для поиска информации, как одиночных, так и групп записей в соответствии с различными требованиями была разработана форма, представленная на рисунке 5.

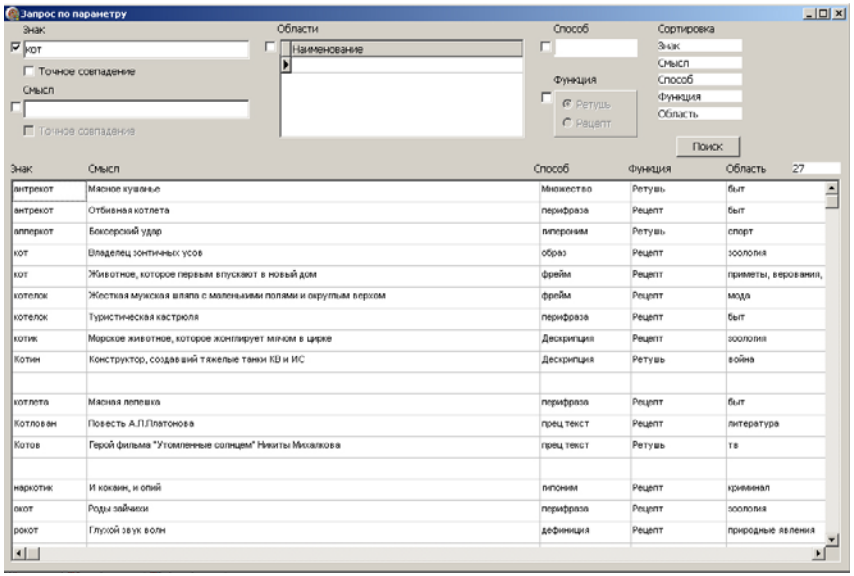

Рис.5

В верхней части формы пользователь задает параметры поиска. Для текстовых полей, таких как Знак и Смысл можно установить положение переключателя «Точное совпадение». В случае если переключатель установлен в положение «Нет», результаты поиска будут содержать фигуры знания, Знак или Смысл которых содержат в качестве подстрок заданные значения. Например, на рисунке 5 в качестве Знака указано значение «кот». В результате среди результатов можно видеть фигуры знания со знаком «Котлован», «рокот» и множество других, в количестве 27.

Рядом с каждым параметром стоит переключатель, который позволяет исключить его из рассмотрения. В результате установки параметров поиска и нажатия кнопки «Поиск» формируется SQL-запрос, включающий также сортировку. Таким образом, результаты запроса не нуждаются в дальнейшей обработке.

Заключение

Полученная схема базы данных представляет собой третью нормальную форму в терминах построения баз данных, что обеспечивает не только минимизацию ресурсов на хранение информации за счет ухода

от избыточности, но и уменьшение риска ошибок [Дейт, 2001]. Она облегчает построение запросов, что позволит выполнять необходимые операции при построении и дальнейшем функционировании лингвокультурного тезауруса. Применение полноценной среды разработки позволяет использовать как работу с данными посредством SQL-запросов, так и самостоятельно описывать обработку записей БД, что может быть полезно в сложных операциях. Программное обеспечение не зависит от одной конкретной базы данных. Можно использовать несколько источников данных, в том числе неоднородных по своему типу. Разработанная база данных позволяет модификацию себя с сохранением работоспособности и расширение за счет добавления новых таблиц либо новых полей в старых таблицах.

Литература

Дейт, 2001 *Дейт, К., Дж.* Введение в системы баз данных, 7-е издание.: Пер. с англ. –М.: Издательский дом «Вильямс», 2001.- 1072 с.

Караулов, 1981 *Ю.Н.Караулов.* Лингвистическое конструирование и тезаурус русского языка. –М.: Издательство «Наука», 1981.

Караулов, 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.