Программа семантической классификации лексики — ПроСеКа: теоретические и прикладные аспекты

On the semantic classification program ProSeCa: theoretical and practical aspects

Кретов А. А. (a_a_kretov@rambler.ru)
Воронежский государственный университет, Воронеж

Рафаева А. В. (anna_raf@rambler.ru)
Московский государственный университет им. М.В. Ломоносова, Москва

Для семантической классификации лексем предлагается использовать модифицированный метод словарной идентификации Э.В.Кузнецовой, ориентированный на лексическую, а не грамматическую семантику. Описывается компьютерная программа ПроСеКа, облегчающая процесс семантической классификации лексики.

Исследование лексико-семантических процессов и состояний лексико-семантической системы, предполагает знание существа, скорости и направления этих процессов, что предполагает хронологическую и семантическую привязку слов и их значений. Хронологическая привязка осуществляется через датировку обрабатываемых текстов. Для осуществления семантической привязки необходимо явным образом описать семиосферу и определить в ней место каждого значения. Наиболее близкими к предлагаемому подходу являются работы по построению ресурсов типа WordNet для разных языков, в которых лексические значения описываются в виде семантической сети [Fellbaum 1998; Азарова и др. 2004]. Но этот подход отличен от нашего. Одно дело — когда каждая дефиниция (и соответствующее ей значение слова) является траекторией в пространстве метаслов. Другое — когда значения лексем представляют собой не цепочку, а узел графа. Узлы графа объединяются отношением «толкуемое-толкующее» или «частное-общее». Таким образом, значение у нас — не семантическая цепочка, а одно из её звеньев — единица классификации.

Синопсисы идеографических словарей отражают как строение семиосфер, так и мировоззрение составителей. Стремлением уменьшить произвол в идеографии было продиктовано алгоритмизованное обращение к дефинициям толковых словарей Э. В. Кузнецовой [Кузнецова 1969], Ю. Н. Караулова и других [Караулов 1982].

Опыт коллектива, созданного Э. В. Кузнецовой [Кузнецова 1988] и руководимого Л. Г. Бабенко [Бабенко 1999], показал, что глагольная семантика по природе своей грамматична, и чем далее мы идём по цепочке глаголов-идентификаторов, тем меньше лексического остаётся в глаголе и тем очевиднее «грамматичность» его семантики.

Например, *плестись* — «*идти* медленно, устало, с трудом передвигая ноги» > идти «*двигаться*, передвигаться ступая ногами» > двигаться «*совершать* движение» > совершать «*делать*, осуществлять, производить» > делать «*совершать*, выполнять, производить». Круг замкнулся. *Производить* > осуществлять «приводить в исполнение, *воплощать* в действительность» > воплощать «*делать* реальностью, осуществлять». Таким образом, от *идти* остаётся лишь ‘каузировать быть’ (осуществлять перемещение ног) или на языке семантических функций «Caus перемещение ног», соответственно, *плестись* — «медленно, устало, с трудом Caus перемещение ног». Поскольку ноги используют преимущественно для перемещения, формулу можно переписать как «Caus Func ноги». Как видим, лексическая семантика сконцентрирована в существительном *ноги*, а ‘функционировать’ и ‘каузировать’ — значения грамматические.

Итак, метод словарной идентификации Э. В. Кузнецовой необходимо переориентировать с метаслов-глаголов на метаслова-существительные.

Для этого надо организовать привязку лексических значений к семантическому пространству. Сложность задачи состоит в том, что имеющиеся описания семантического пространства по разным причинам непригодны, а «опробованный в экспе-

рименте Ю.Н. Караулова метод автоматического извлечения тезауруса из толкового словаря может использоваться только при условии дальнейшего редактирования полученных данных человеком» [Кобозева 2000: 134] или при условии предварительного редактирования дефиниций: ведь лексикографы не предполагали, что их дефиниции послужат семантической классификации лексики.

Для описания семиосферы с опорой на дефиниции нужна специальная компьютерная программа, избавляющая исследователя от повторения уже выполненной работы по семантизации ЛСЕ (значений слов и фразеологизмов) в произвольно выбранном тексте.

Такая потребность возникает в силу ряда причин: 1) открытости словаря, 2) необходимости описывать синхронное состояние лексико-семантической системы, 3) необходимости определять тематическую принадлежность текстов, 4) необходимости проекции лексической семантики текста на систему семантических координат.

Осуществление полной семантизации ЛСЕ текста или корпуса текстов открывает целый ряд новых возможностей: 1) создание частотно-семантических словарей, названное П. М. Алексеевым «оценкой толкового словаря по тексту» [Алексеев 1973], 2) создание исторической лексикологии русского (а в перспективе — любого другого языка, имеющего письменную традицию), 3) типологические исследования лексической семантики и т.д.

Предлагаемое решение задачи базируется на следующих положениях:

1) лексическая семантика не зависит от грамматики данного языка («части речи» — разные по форме сосуды, наполняемые одной и той же лексико-семантической субстанцией);
2) в индоевропейских языках лексическая семантика концентрируется в именах (существительном и прилагательном) и глаголе (в отличие от WordNet'а, мы не работаем с наречиями), а в общем случае — в лексических морфемах — корнях.
3) лексическая семантика глагола и прилагательного в конечном итоге сводится к семантике существительного и может быть описана через неё (например, *белый* — 'цвета *снега, мела* или *молока*');
4) собственно глагольная семантика при ближайшем рассмотрении оказывается грамматической: процессуальной (действие, отношение, состояние, погружённые во время), инхоативной-каузативной (*ослепнуть* — 'начать не иметь зрения' и *ослепить* — 'каузировать начать не иметь зрения'), фазовой (*расцвести* 'начать быть — о цветке', *увянуть* 'кончить быть — о цветке), утвердительной-отрицательной (*ослепить* — каузировать кого-л. начать не иметь зрения, *воочесить* — 'каузировать кого-л. начать иметь зрение'), акционсъартной (*петь — попеть, запеть, пропеть, допеть, распеться, отпеть, отпеться* и т. д.);
5) семантические функции И.А.Мельчука — А.К.Жолковского — Ю.Д.Апресяна являются грамматической надстройкой над лексической семантикой; аналогичный статус имеют и «семантические примитивы» А.Вежбицкой.
6) лексическая семантика не выразима вне частеречного оформления; следовательно, при анализе лексической семантики следует ориентироваться на наименее маркированную часть речи — существительное и те значения, которые им выражаются. Наименьшая маркированность существительного как части речи обоснована В.Г.Руделёвым [Руделёв 1995], а также выводится из сближения маркированности с рецессивностью, а немаркированности с доминантностью, предложенного Вяч. Вс. Ивановым и Т.В. Гамкрелидзе [Гамкрелидзе, 1984]. Наиболее многочисленный член оппозиции является доминантным и немаркированным, а наиболее многочисленная часть речи в известных нам словарях — имя существительное;
7) специфика лексической семантики может быть выявлена и описана только в результате последовательного снятия грамматических надстроек и напластований, составляющих «грамматику семантики»;
8) всё регулярное должно выноситься из словаря в грамматику [Щерба, 1974; Морковкин 1990];
9) графически лексико-семантическое пространство может быть представлено в виде ориентированного графа, исходными (при развёртывании) и конечными (при классификации значений) узлами которого являются базовые понятия человеческого языка.

Проектами, подобными данному, являются: [Паллас 1787–1789; Шишков 1832; Roget 1986; Lorge, Thorndike, 1938; Прокопов 1945; Караулов 1982; Морковкин 1984; Кузнецова 1988; Баранов 1995; Шведова 1998–2007, Fellbaum 1998; Бабенко 1999; Лукашевич, Добров 2002].

Идеографические словари отражают языковую реальность на нижних уровнях обобщения (синонимические ряды, гипо-гиперонимические отношения), а на высших уровнях обобщения количество и качество выделяемых таксонов зависит от исследователя, на что указывали [Задорожный 1983] и [Караулов 1976, 1981]. Особенно хотелось бы отметить значительную частеречную независимость «Тезауруса» [Roget 1986], хотя и обусловленную морфологической бедностью английского языка, но принципиально верную.

«Русский семантический словарь» [Шведова 1998–2007] — в соответствии со взглядами Н. Ю. Шведовой — ориентирован на частеречную

семантику, поэтому *белый, белеть-белить* и *белизна* в нём оказываются в разных местах и разных томах, а сбор этих лексически тождественных значений воедино по трудозатратам близок к созданию нового идеографического словаря.

Словарь [Lorge, Thorndike 1938] свидетельствует о принципиальной возможности тотальной семантизации больших корпусов текстов, правда, опыт, накопленный при создании этого словаря, практически недоступен, равно, как и проверка обоснованности решений, принятых его составителями. В цифры, полученные исследователями, остаётся только верить или оценивать их достоверность, исходя из соображений общего плана.

В этом отношении содержательнее «Русский семантический словарь» [Караулов 1982]. К сожалению, слова-аттракторы в нем задавались a priori, а не получались в ходе исследования. Кроме того, этот словарь, как и словарь екатеринбургского коллектива [Кузнецова 1988], показал, что *словарные дефиниции — независимо от их качества — лишь полуфабрикат для семантизации лексики*.

Опыт уральских лингвистов [Кузнецова 1988; Бабенко 1999] позволил увидеть: ориентация на толкующие глаголы при семантизации глагольной лексики порой приводит к созданию чисто грамматических (фазовых или каузативных) группировок глаголов, весьма разнородных по своей лексической семантике.

Поскольку слово или словосочетание текста на любом языке может быть семантизировано по-русски, мы в перспективе получаем инструмент семантического анализа текстов на любых языках и соответственно — анализа лексико-семантических систем любого языка.

Автоматическая семантизация иноязычного текста существенно облегчается, если он входит в корпус параллельных текстов, одним из которых является русский. В таком случае исследователю останется установить соответствие между входным и русским словом и связать это слово со словарём.

Выделение в дефинициях метаслов-идентификаторов может быть частично формализовано. Так, в сочетании «глагол+существительное» идентификатором, как правило, является существительное. В конструкциях типа «*Молочные железы <u>женщины</u>*», «*Физиологическое состояние <u>человека</u>*», «*Непрерывное движение <u>крови</u>*» надо выбирать идентификатором последнее слово — «женщина», «человек», «кровь». В конструкции «*Совокупность жизненных отправлений <u>организма</u>*» — «организм».

Опыт [Караулов 1982] свидетельствует также о необходимости создания (с опорой на имеющиеся) особого типа дефиниций и особого метаязыка (на базе русского), ориентированных на компьютерный анализ и приспособленных для него: предполагается снятие неоднозначности (асимметрии) единиц — как в виде омонимии-полисемии, так и в виде синонимии. Этот метаязык должен быть ориентирован скорее на компьютер, чем на человека. Хорошая дефиниция — та, что автоматически приводит к верной классификации входной ЛСЕ.

В качестве базовых предполагается взять дефиниции МАС-2 с корректировкой по БТС и БАС-3. многоступенчатые толкования МАС предполагается обрабатывать следующим образом: например, в случае глагола *идти* (первое значение: «Передвигаться, перемещаться в пространстве»): а) каждая отдельная дефиниция принимается за отдельное значение; б) идентификаторами оказываются существительные: *ноги, средства передвижения, почтовые отправления/грузы, облака, вода, воздух, льдины, плоты, бревна*; в) для значения «г) Перемещаться массой, потоком, вереницей. 1) О движении облаков, воды, воздуха и т. п. 2) О движении льдин, плотов, бревен и т. п. по воде. || Передвигаться стаей, косяком и т. п. (о рыбе, мелком пушном звере)» интерпретация двуступенчата: рыба -> косяк -> движение косяка; лемминг -> стая леммингов -> движение стаи леммингов. Исходное — льдина, плот, бревно, рыба или мышь, к которой применяется семантическая функция «множество», а затем к данному конкретному множеству применяется семантическая функция «движение».

Для компьютеризованного построения лексико-семантического пространства русского языка (а в перспективе и других языков) в виде ориентированного графа необходимо:

- снять неоднозначность единиц метаязыка, представив их в виде: *1лемма1 <номер омонима, лемма, номер значения>* (синонимия метаязыка может быть устранена целенаправленной редукцией синонимических рядов до их доминанты);
- каждой ЛСЕ дать дефиницию;
- в дефиниции выделить метаслово (по сложно формализуемому принципу — сохранению важнейшего лексического значения). Именно это метаслово будет служить определяемым следующего уровня;
- повторять процесс до тех пор, пока последовательность (цепочка) подобных пар вида ЛСЕ — словарная дефиниция не дойдёт до одного из финальных узлов или пока не возникнет противоречие с уже введёнными данными. Одним из признаков, позволяющих определить, что конкретная ЛСЕ является финальной, служит появление цикла в толковании: ЛСЕ в конечном итоге толкуется сама через себя или через ЛСЕ, отличающуюся лишь номером значения (например, *1существо2 (то, что существует как живой организм, **животное**) — 1животное1 (живое **существо**, способное чувствовать и передвигаться) — 1существо1)*;
- для этого последнего узла словарная дефиниция может и не быть заданной, что приближает концы цепочек к неопределяемым понятиям математики.

Каждая такая последовательность (цепочка) ЛСЕ является путём в ориентированном графе, представляющем лексико-семантическое пространство языка, а само ЛСП строится как объединение таких путей, заданных пользователем на основе анализа словарных дефиниций.

Исходная задача программы ПроСеКа (ПРОграмма СЕмантической Классификации) — служить инструментом создания, проверки и сохранения цепочек ЛСЕ, заданных пользователем, т.е. фактически быть редактором этих цепочек. Программа написана на языке C++ в среде Borland C++ Builder. Основная задача программы — сохранять цепочки и текстовые примеры к ним, введённые пользователем, и создавать словарь всех встретившихся в этих цепочках ЛСЕ. Кроме того, в функции программы входит контроль за согласованностью данных: каждая ЛСЕ может иметь не более одной дефиниции (финальные узлы цепочек могут не иметь дефиниции), и последовательность ЛСЕ во всех цепочках должна сохраняться неизменной. Кроме того, на множество цепочек могут накладываться дополнительные ограничения, например, пользователь может ограничить длину цепочек.

Цепочки могут создаваться как в программе (в Мастере цепочек или в редакторе, позволяющем вводить несколько цепочек одновременно), так и в текстовом редакторе или электронной таблице, в виде текстовых файлов с разделителями. На Рис. 1 приведён пример цепочки, построенной пользователем в Мастере цепочек для ЛСЕ *ездить*:

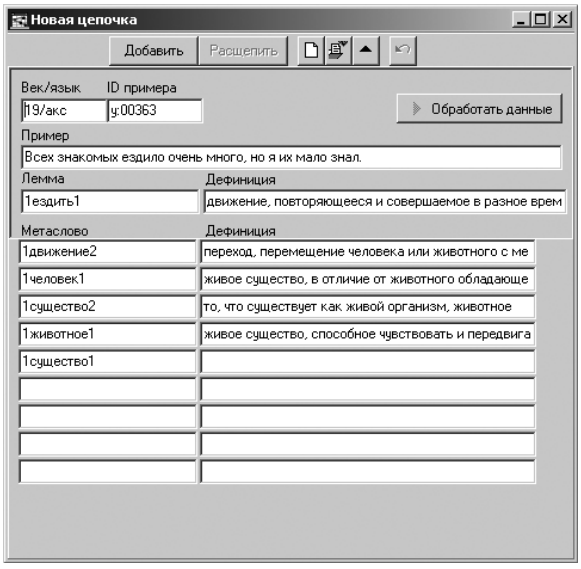

Рис. 1. Режим редактирования цепочки в Мастере цепочек

Создание цепочек частично автоматизировано. Так, если в цепочке используется слово, уже внесённое в словарь, продолжение цепочки достраивается автоматически.» В том случае, когда пользователь вводит многозначное слово, одно или несколько значений которого уже содержатся в словаре, программа предлагает пользователю достроить цепочку по уже имеющимся данным или ввести новое значение рассматриваемого слова. Выбор нужного значения — дело пользователя, а не программы, задачи программы сводятся к следующему: 1) предложить варианты автоматического продолжения цепочек для тех слов (значений), которые уже содержатся в словаре, и 2) проверить соответствие каждой новой цепочки установленным правилам (т.е. выполнение наложенных ограничений). Помимо существенной экономии сил и времени, такой режим исключает ошибки ручного ввода, в том числе несогласованность различных цепочек, ошибок и опечаток в дефинициях и т.п.

На множество цепочек накладываются следующие ограничения:

- Ограничение на длину. Цепочка должна содержать хотя бы два узла, максимальная длина цепочки может быть ограничена пользователем (по умолчанию ограничение отсутствует);
- Ограничение на единственность цикла. В цепочке не может присутствовать более двух одинаковых узлов, т. е. допустимо появление не более одного слова, толкующегося через себя само (в том же или другом значении). При этом второе вхождение данного слова служит сигналом конца цепочки;
- Ограничение на единственность толкования. Каждое значение слова или словосочетания может иметь не более одного толкования (финальные узлы могут не иметь толкования вовсе). Поскольку используемые словарные толкования редактируются, представляется достаточно важным, чтобы в дальнейшем не возникло расхождений между данными, внесёнными в различное время.
- Ограничение на согласованность. Множество цепочек должно быть согласованным, т. е. от каждого значения слова существует ограниченное число путей к финальному узлу или узлам, за исключением случаев толкования его через себя или другое значение этого же слова.

Первые два ограничения проверяются для каждой конкретной цепочки отдельно; третье и четвертое, очевидно, требуют сравнения новой цепочки с уже введёнными данными.

На Рис. 2 приведён пример одного из вариантов автоматического дополнения, предложенного программой. Исходная цепочка, созданная при помощи программы MS Excel, противоречила введённым ранее данным (см. Рис. 3).

Результат обработки введённых пользователем данных помещается в файлы данных (в настоящее время это текстовые файлы с разделителями, что позволяет просматривать и анализировать результат работы программы в электронной таблице, в дальнейшем возможен переход на другой или другие форматы хранения данных). Кроме того, результат выводится на экран (Рис. 4).

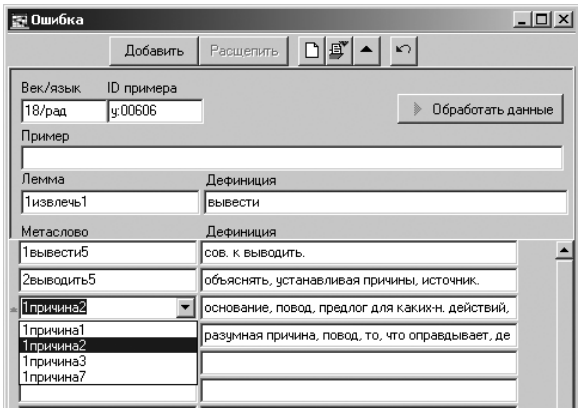

Рис. 2. Автоматическая проверка и автоматическое дополнение цепочки

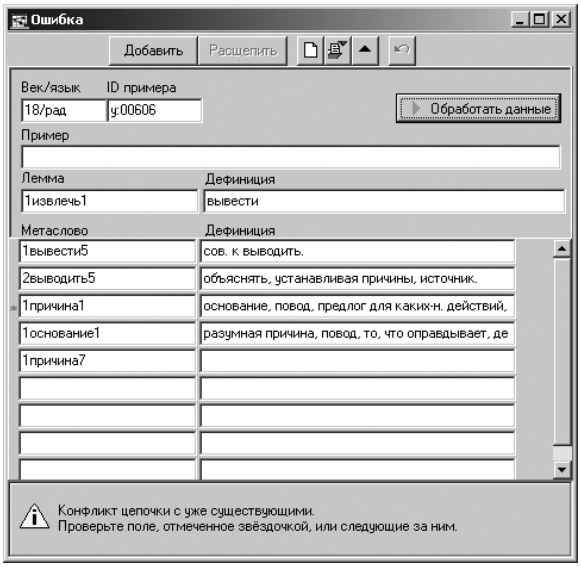

Рис. 3. Результат автоматической проверки цепочки: сообщение об ошибке

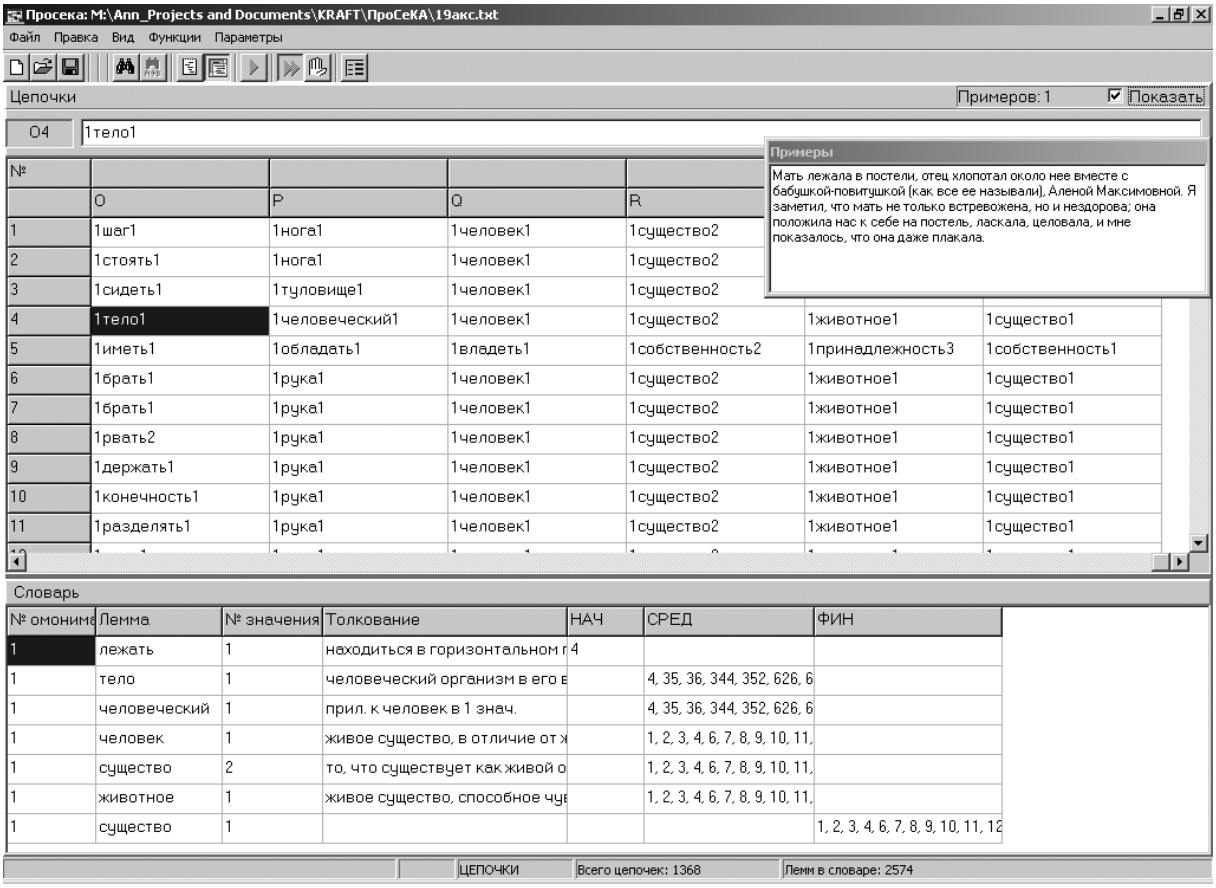

Рис. 4. Просмотр цепочек и словаря

В верхней части экрана даны все существующие в настоящее время цепочки, выровненные по правому краю, а в нижней — все ЛСЕ, входящие в выбранную цепочку, причём показаны не только сами ЛСЕ, но и часть словарной статьи (дефиниция, а также идентификаторы цепочек, в которые входит заданная ЛСЕ). В отдельном окне показываются все примеры, приписанные к выбранной цепочке. Пользователь может включать и отключать просмотр примеров.

Работа с программой выявила необходимость внесения следующих дополнений:

1) Возможность задавать произвольное количество наборов правил для проверки цепочек. Как уже говорилось, ЛСП строится эмпирически, как множество допустимых «путей» в этом

пространстве. При этом некоторые первоначальные предположения о виде ЛСП не подтвердились, другие требуют дальнейшей проверки. Эта возможность в настоящее время реализована на уровне абстрактного класса правил и ряда конкретных правил;

2) Возможность изменять вид и представление данных (абстракция данных). В настоящее время программа фактически хранит не только сами созданные пользователем цепочки, но и порядок их построения. На первоначальном этапе построения ЛСП эта возможность, безусловно, полезна, однако в дальнейшем, особенно при переходе к изучению вида ЛСП, она едва ли окажется нужной;
3) Дополнительные возможности по классификации, сортировке и обработке данных. В частности, сейчас цепочки располагаются в том порядке, в каком они введены в программу, что является не самым удачным способом хранения с точки зрения быстродействия программы и удобства просмотра.
4) Возможно, целесообразным окажется просмотр ЛСП не только в виде цепочек, построенных пользователем, но и в виде графа, описывающего допустимые переходы между узлами.

Реализация этих возможностей — дело будущего.

Литература

1. *Азарова И. В., Синопальникова А. А., Яворская М. В.* Принципы построения wordnet-тезаруса RussNet. — http://www.dialog-21.ru/Archive/2004/Sinopalnikova.htm.
2. *Алексеев П. М.* Семантические частотные словари // Статистика речи и автоматический анализ текста. 1972. — Л., 1973, с.20-36.
3. *Бабенко Л. Г.* (ред.) Толковый словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы. — М., 1999.
4. *Баранов О. С.* Идеографический словарь русского языка / О. С. Баранов. — М.,1995.
5. *БАС-3* Большой академический словарь русского языка, тт. 1–10. — М.-СПб, 2004–2008.
6. *БТС-1998* Большой толковый словарь русского языка / под ред. С. А. Кузнецова, — СПб, 1998.
7. *Гамкрелидзе Т. В.* Индоевропейский язык и индоевропейцы: Реконструкция и историко-типологический анализ праязыка и протокультуры / Гамкрелидзе Т. В., Иванов Вяч. Вс. — Тбилиси, 1984.
8. *Задорожный М. И.* Два подхода к построению идеографического словаря // О преподавании русского языка и литературы в киргизской школе. Вып. 10. — Фрунзе, 1983.
9. *Караулов Ю. Н.* Общая и русская идеография. М., 1976.
10. *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка, М., 1981.
11. *Караулов Ю. Н.* Русский семантический словарь. Опыт автоматического построения тезауруса: от понятия к слову / Ю. Н. Караулов, В. И. Молчанов, В. А. Афанасьев, Н. В. Михалев. — М.: Наука, 1982.
12. *Кобозева И. М.* Лингвистическая семантика. — М.: Удиториал УРСС, 2000.
13. *Кузнецова Э. В.* Метод ступенчатой идентификации в описании лексико-семантических групп слов. //Учен. зап. / Тартус. ун-т. — 1969. — Вып. 228.
14. *Кузнецова Э. В.* (ред.) Лексико-семантические группы русских глаголов. — Свердловск, 1988.
15. *Лукашевич Н. В., Добров Б. В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Труды Международного семинара Диалог 2002 по компьютерной лингвистике и ее приложениям, Т. 2. — М.: Наука, 2002.
16. *МАС-2* Словарь русского языка в 4-х тт. / под ред. А. П. Евгеньевой, Изд. 2-ое, испр. и доп. — М., 1981–1984.
17. *Морковкин В. В.* (ред.) Лексическая основа русского языка. Комплексный учебный словарь. — М., 1984.
18. *Морковкин В. В.* Основы теории учебной лексикографии. Дисс. в форме науч. докл. .. докт. филол. наук. М., 1990.
19. *Паллас П. С.* Сравнительные словари всех языков и наречий, собранные десницею всевысочайшей особы. Отделение перьвое, содержащее в себе Европейские и Азиатские языки. Ч. 1–2, СПб, 1787–1789.
20. *ПрокоповВ.В.* Основныелексико-семантические группы русского глагола. Дисс. … канд. филол. наук, — Самарканд, 1945. — 197 с.
21. *Руделёв В. Г.* Вначале было слово / Руделёв В. Г., Руделёва О. А. — Тамбов , 1995.
22. *Шведова Н. Ю.* (ред.) Русский семантический словарь, тт. I–IV. — М., 1998–2007.
23. *Шишков А. С.* Собрания языков и наречий с примечаниями на оныя. //Собр. соч. и переводов Адмирала Шишкова, Ч. XV, СПб, 1832.
24. *Щерба Л.*В. Языковая система и речевая деятельность. Л.: Наука, Ленингр. отд-ние, 1974.
25. *Fellbaum, C.* (1998), ed. «WordNet: An Electronic Lexical Database». MIT Press, Cambridge, MA.
26. *Lorge Irving, Thorndike Edvard E.* A Semantic Count of English Words. — New York, 1938.
27. *Roget* 1986: The Penguin Roget`s thesaurus of English words and phrases /New edition copletely revized, updated and abridged by Susan M. Lloyd, Pengin books, 1986. — 776 p.