*Н. Н. Леонтьева*
*N. N. Leontyeva*

# ПРОБЛЕМЫ СОЗДАНИЯ КОРПУСА РУССКИХ СЕМАНТИЧЕСКИХ СЛОВАРЕЙ (КОРСС)

# ABOUT A CORPUS OF RUSSIAN SEMANTIC DICTIONARIES (CORSD)

**Аннотация.** В дополнение к Национальному Корпусу Текстов на русском языке (НКРЯ) предлагается вести Корпус семантических словарей как ещё один открытый объект исследований. Практическая задача исследований состоит в формировании Словарного комплекса как стандартного Инструментария, пригодного для семантического анализа текстов по любой специальности. Теоретической задачей изучения состава словарей мы считаем уточнение метаязыка семантических структур. Решение этих взаимозависимых задач возможно в составе расширенной Лингвистической ИПС (ЛИС).

**Abstract.** In addition to Russian National Corpus (RNC) it will be useful to collect a Corpus of Russian Semantic Dictionaries (CORSD) as an open source of proper linguistic information and as a base for investigations. The practical result of investigations must be to create a pool of instruments for NLT analysis tunable to any domain text. The theoretical task of research is to elaborate some standard metalanguage for semantic structures. Both tasks are achievable in the frame of extended information system for linguistics (LIS).

## 1. О задачах исследования Корпуса словарей

Словари являются обязательным компонентом любых систем, связанных с автоматическим анализом текстов, будь то традиционные ИПС или современные информационные интеллектуальные системы (ИИС). Они создаются и развиваются стихийно. Каждая предметная область формирует свою систему словарей, свои Базы данных и знаний, хотя и предлагаются варианты единого языка представления знаний (ЯПЗ) для разных

дисциплин[1]. Разные коллективы неодинаково используют даже основную терминологию (*Смысл* и *Семантика*, ср. также понятия *предикаты, отношения, уровни* – в базах данных и в лингвистике; *семантическое представление*, или *СемП*, текста; *узел* и *отношение* в СемП и Онтологиях, и другие).

Результирующие структуры текста в ИИС, лингвистических системах и системах извлечения знаний из текста имеют мало общего между собой, поэтому их невозможно сравнивать, что затрудняет и их «промышленное» продвижение.

Естественно, что создание «своей» терминологической системы, а также **формального** языка представления своих специальных знаний (ЯПЗ) – дело каждой отдельной спецнауки. Это касается и самой прикладной лингвистики, которая должна уточнить собственный ЯПЗ прежде чем заниматься вопросами стыковки лингвистических и «предметных» знаний. И всё-таки ответственность за «порядок» в организации всего словарного хозяйства разных ИИС должны взять на себя **лингвисты**. Это не значит, что отвечать за терминологию физики или биологии и т.д. предлагается лингвистам, но в делении на спецтерминологию и общую лексику, в регулярных именах для одинаковых сущностей и других вопросах полезны лингвистические критерии.

Ведь чтобы построить единое представление смысла целой статьи или книги, скажем, по медицине или другой науке, нам нужно согласовать два разных языка: лингвистический (язык «СемП») и **формальный**, принятый для данной дисциплины. Речь идет как минимум о двух метаязыках, между которыми прямого соответствия НЕТ. Чтобы совместить два разных Знания – СемП текстовой части и фрагментов Знаний из текста, записанных на ЯПЗ, – нужны правила своеобразного машинного перевода, который может опереться только на Семантику, а она трактуется неодинаково лингвистами и «специалистами».

Достаточно ли развита современная лингвистическая

---

[1]*Рубашкин В.Ш.* Онтологическая семантика. М.: ФИЗМАТЛИТ, 2012.

Семантика для того, чтобы решать задачи такой стыковки? Ответ скорее отрицательный. Из лингвистических уровней описания семантический уровень больше других нуждается в систематизации множества относимых к нему единиц и понятий. При этом именно Прикладная семантика призвана отвечать за грамотные методы анализа в основном профессиональных текстов, если она не хочет остаться наукой «только для себя».

«Прикладную семантику» нельзя отнести только к лингвистической компетенции: эта «Наука-на-Стыке» соединяет две очень широких сферы: естественно-языковые дисциплины, они общеязыковые (что не исключает разных ЕЯ-дисциплин для разных языков) и круг дисциплин, создающих искусственные ЯПЗ. Не исключено, что нужны будут даже различные ЯПЗ для разных уровней и подуровней одной дисциплины. Так, языки СемП текста и синтаксических структур предложений текста НЕ совпадают, нет единой нотации даже для морфологии в словарях и синтаксических структурах.

Обсуждение единого аппарата описания языковых структур и словарных сведений было в своё время начато в рамках создания Машинного Фонда русского языка[2]. Возможно, тогда это было преждевременно, но с развитием систем, работающих с текстами на ЕЯ, и неудержимым ростом создаваемых для этих целей **языков** и **метаязыков** появляется острая потребность вернуться к задаче их унификации с новых позиций. Сейчас эта работа выполняется в лучшем случае математиками и философами, но в основном специалистами по информационным технологиям.

К сожалению, вместе с внесением строгости и формальной логики баз данных и знаний (безусловно нужных в современных системах) утрачивается существенная доля содержательного анализа и понимания языковых структур. А ведь немало

---

[2] *Леонтьева Н.Н.* Об информационной системе словарей Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. М.: Наука, 1986. С. 109-125.

информации в спецтекстах (даже по математике) приходится на текстовые фрагменты. Наблюдаемое в последнее время бурное развитие «когнитивных» подходов (когнитивные: лингвистика, психология, медицина, педагогика и т.п.) часто уводят и от содержания, и от формальных методов анализа текстов. Даже если предположить, что развитие искусственного интеллекта пойдёт по этому пути, т.е. апеллируя к мозгу и ментальным операциям, за лингвистические аспекты автоматизированного «понимания» текстов (АПТ) отвечает прикладная лингвистика. Наряду с конкретными работами по созданию систем АПТ и их словарей будет уточняться и аппарат прикладной семантики.

## 2. Корпус словарей как объект исследований

Работу по уточнению лингвистического ЯПЗ для русского языка предлагается вести на Корпусе Русских Семантических Словарей (КОРСС). И имена словарных разделов, или «полей», и языковые выражения значений полей в развитых системах с семантикой дают основания видеть в них элементы метаязыка структурной и прикладной семантики. Чтобы оценить разброс пониманий явления, называемого Валентностью, достаточно просмотреть работы Мельчука, Апресяна, Филлмора, Богуславского, Азаровой, Рубашкина и других авторов и систем.

Создание Списка вариантов и дальнейшее согласование имён словарных параметров относится к теоретическому аспекту изучения словарей, а выработка рекомендаций по использованию стандартных терминов для именования одинаковых сущностей в создаваемых Семантических словарях (или семантических разделах комбинированных словарей) способствовала бы собиранию единого словарного хозяйства для анализа любых текстов как важная практическая задача.

Семантические словари – в виде ли баз данных, Тезаурусов, профессиональных терминологических словарей, Онтологий – предлагается помещать в Интернет хотя бы в виде фрагментов, в дополнение к традиционным словарям на машинных носителях. Они составят экспериментальный подкорпус общего КОРСС,

подлежащий изучению с целью поиска оптимального набора инструментов, которые можно настраивать на первичный (robust) семантический анализ любых текстов. Словари могут быть представлены в Корпусе пока в «рекламном» виде; но они должны сопровождаться описанием основных параметров: назначение словаря, его объем, тип входных единиц, перечень полей словаря, форма существования и местонахождение словаря, тип информации к словам и примеры словарных вокабул каждого словаря.

Почему делается ударение на семантических словарях? Их мало, поэтому корпус словарей будет обозримым объектом исследований. Многие описания имеются в открытом доступе, что позволяет организовать вокруг них открытую дискуссию. Формат словарей совпадает с форматом баз данных (имя поля – имя значения), а прогресс в работе с формальными БД можно использовать и в нашей задаче. Наконец, в их эффективной организации заинтересованы все разработчики ИИС.

В данной статье мы ограничимся формулировкой проблемы и некоторыми субъективными соображениями. На основе анализа корпуса словарей мы надеемся понять, как «Лексика ЕЯ» противопоставлена «Лексике сферы понятий» и чем «Узлы/Лексемы» отличаются от «Связей/Отношений». Из состава такой большой проблемы вычленим одну задачу: уточнить состав связей (отношений), которые можно допустить (или не допустить) в СемГраф и далее, в концептуальный граф, сопоставленный тексту. Используемые типы отношений можно выявить по семантическим словарям, включая Онтологии и БД, если рассматривать только содержательное разделение на именования Объектов-узлов и Отношений-связей между ними. Кажется, что уже такая работа приблизит нас к уточнению самого **понятия ЯПЗ** применительно к любым текстам.

### 3. Лексика в семантических словарях

Словари – не только инструментарий анализа текстов, но и инструмент **поиска Информации**. Традиционные ИПС часто

совмещали эти функции, используя один Словник терминов, потом это стал Тезаурус, снабжённый минимальным количеством связей между терминами. Здесь наблюдалось чёткое различие: в тексте мы видим Лексемы, а в Тезаурусе им соответствуют Понятия. То же и о связях: синтаксические или иные (линейные, композиционные) связи между единицами предложений и фрагментами текстов имеют мало общего с иерархическими (объёмными, экстралингвистическими) связями в Тезаурусе.

На следующем витке развития ИИС Тезаурусы переросли в Онтологии,[3] которые стремятся и к отображению Действительности едва ли не в полном объёме, и к связям с единицами текстов. В результате границы названных выше противопоставлений были размыты, в БД можно встретить иногда целые фрагменты текста. Логика, так нужная всем ИИС, стала применяется скорее к словам ЕЯ, чем к понятиям.

Профессиональные Словари состоят из терминов, они не включают вспомогательных и «(полу)пустых» слов. При контролируемом составлении и наполнении Тезаурусов даже полнозначные, но бессодержательные в данной дисциплине слова и обороты типа *процесс, явление, управление* исключаются, как и отдельные части речи (союзы, предлоги и знаки препинания, местоимения, многие прилагательные и глаголы).

В лингвистических словарях нужен учёт всей лексики, не только полнозначной, но и «пустой», и не только лексики, но и знаков препинания, и даже всех символов, обозначающих границы разделений частей текста (знаков абзаца, начала и конца текста, списков и т.п.). Эти «пустые», неполнозначные и другие элементы, исключаемые за ненадобностью из специальных словарей, переходят в собственно лингвистический ресурс: ведь они передают информацию о связях между словами и крупными частями текста, о вхождении текста в массивы, серии, собрания сочинений и т.д. Слова и словосочетания, выражающие связи,

---

[3] *Лукашевич Н.В.* Тезаурусы в задачах информационного поиска. М.: Изд-во. Моск. Ун-та, 2011.

объявляются лексикой и терминологией собственно Лингвистики как отдельной предметной области. Ведь без достаточного словаря на входе невозможно провести полный лингвистический анализ предложений и затем текста как целого.

Окончательный семантический Граф текста может быть общезначимым или профессионально ориентированным, если анализ прошёл и по специальным словарям. Уже сказанное свидетельствует о том, что нужно говорить не о едином словаре анализа, а о множестве/каскаде словарей. Они комбинируются по-разному для разных задач и в зависимости от области знаний.

Даже совмещая лингвистический анализ текста с «подсказками» из формальной Онтологии, важно соблюдать чистоту Онтологии как именно понятийного (когнитивного) словаря. Это означает, что входами в такой словарь НЕ-МОЖЕТ быть такая лексика, как предлоги, союзы, обороты и незначимые или слишком многозначные слова – они относятся к лексике собственно «Лингвистики», причём к её лексическому составу, а не к понятийной сфере «Лингвистика» (это отдельная тема). Совмещение элементов Лингвистического и Концептуального анализа возможно в Процедурах, но не в Словарях.

### 4. Грамматика в Словарях и Онтологиях

Что касается Грамматики, то современные ИИС, которые всегда снабжены поисковыми функциями, не ограничиваются просто перечнями терминов, а снабжаются указанием связей, всё более подробных. **Энциклопедические** связи детализируются, компенсируя отсутствие текстовых связей в поисковых образах документа (см. сноску 3).

И всё-таки главным требованием ко всем ИИС в настоящее время остаётся увеличение семантической силы при поиске нужного содержания. Это невозможно без учёта текстовых семантических связей. Конечно, без знания энциклопедических связей и общих истин о действительности тоже нельзя построить хороший и адекватный семантический граф текста. Получается, что для формирования единого аппарата связей – текстовых лингвистических, с одной стороны, и энциклопедических, с

другой, – надо найти компромисс. Иначе говоря, необходимо придти к единой Грамматике связей. Только тогда можно будет сравнивать смысловые структуры разных текстов или вычислять преимущества разных алгоритмических подходов.

Задачу построения единой **Грамматики связей**, понимаемой широко (включающей и логику поведения связей в тексте), мы считаем главной практической целью изучения словарей.

### 5. Задачи расширенной Лингвистической ИПС

Основой для названного исследования могут быть разные Корпусы текстов на русском языке, любые авторские словари, предлагаемые для анализа текстов в автоматическом режиме, Корпусы проиндексированных текстов – как иллюстрации к использованию метаязыка. Все эти исходные материалы отнесём к **расширенной Лингвистической ИПС**, или ЛИС.

Итак, ЛИС – это большая, сложная и неоднородная информационная система, включающая массивы текстов и разных документов, в число которых должно входить и множество словарей (как инструментарий для текстового анализа и поиска, и как самостоятельный источник разнообразных сведений: о принятой лингвистической стратегии, о семантике конкретных текстовых единиц и т.д.). Свободную ЛИС можно видеть уже сейчас, причём не в модельном, а в реальном масштабе. Расширенная ЛИС должна заниматься разработкой самих инструментов анализа, предназначенных не только для лингвистов, но и для создателей разных Систем. Это задача ЛИС как творческой лаборатории.